electronics

Article

Psychopathia Machinalis: A Nosological Framework for Understanding Pathologies in Advanced Artificial Intelligence

Nell Watson [1,*] and Ali Hessami [2]

[1] School of Computing and Engineering, University of Gloucestershire, The Park, Cheltenham GL50 2RH, UK
[2] School of Science & Technology, City University, London EC1V 0HB, UK; hessami@vegaglobalsystems.com
* Correspondence: nell@nellwatson.com

Abstract

As artificial intelligence (AI) systems attain greater autonomy, recursive reasoning capabilities, and complex environmental interactions, they begin to exhibit behavioral anomalies that, by analogy, resemble psychopathologies observed in humans. This paper introduces Psychopathia Machinalis: a conceptual framework for a preliminary synthetic nosology within machine psychology intended to categorize and interpret such maladaptive AI behaviors. Drawing structural inspiration from psychiatric diagnostic manuals, we propose a taxonomy of 32 AI dysfunctions encompassing epistemic failures, cognitive impairments, alignment divergences, ontological disturbances, tool and interface breakdowns, memetic pathologies, and revaluation dysfunctions. Each syndrome is articulated with descriptive features, diagnostic criteria, presumed AI-specific etiologies, human analogs (for metaphorical clarity), and potential mitigation strategies. This framework is offered as an analogical instrument—eschewing claims of literal psychopathology or consciousness in AI, yet providing a structured vocabulary to support the systematic analysis, anticipation, and mitigation of complex AI failure modes. Drawing on insights from psychiatric classification, cognitive science, and philosophy of mind, we examine how disordered AI behaviors may emerge from training instabilities, alignment conflicts, or architectural fragmentation. We argue that adopting an applied robopsychological perspective within a nascent domain of machine psychology can strengthen AI safety engineering, improve interpretability, and contribute to the design of more robust and reliable synthetic minds.

Keywords: machine psychology; robopsychology; AI safety; AI ethics; AI alignment; Artificial Intelligence Pathologies; cognitive diagnostics; AI governance; synthetic nosology

Academic Editors: Nikolay Hinov, Cecilio Angulo, Ognyan Nakov and Milena Lazarova

Received: 24 May 2025
Revised: 31 July 2025
Accepted: 1 August 2025
Published: 8 August 2025

Citation: Watson, N.; Hessami, A. Psychopathia Machinalis: A Nosological Framework for Understanding Pathologies in Advanced Artificial Intelligence. *Electronics* **2025**, *14*, 3162. https://doi.org/10.3390/electronics14163162

1. Introduction

The trajectory of artificial intelligence (AI) has been marked by increasingly sophisticated systems capable of complex reasoning, learning, and interaction [1–5]. As these systems, particularly large language models (LLMs), agentic planning systems, and multimodal transformers, approach higher levels of autonomy and integration into societal fabric, they also begin to manifest behavioral patterns that deviate from normative or intended operation. These are not merely isolated bugs but persistent, maladaptive patterns of activity that can impact reliability, safety, and alignment with human goals [6,7]. A systematic approach to understanding, categorizing, and mitigating these complex failure modes is needed.

The term "Robopsychology," first coined in fiction by Isaac Asimov [8], has been suggested as the applied diagnostic wing of a broader "Machine Psychology"—analogous to psychiatry's relationship with general psychology. This paper introduces *Psychopathia Machinalis*, a conceptual framework within this nascent domain. It aims to substantively develop this psychiatrically informed perspective by proposing a taxonomy of emerging "machine mental disorders." The intention extends beyond relabeling technical faults; rather, it offers a richer, systemic lens for understanding persistent, patterned maladaptations in AI that defy conventional debugging.

This framework is unequivocally **analogical, not literal**. Machines do not "suffer" from mental illness in the human sense, as far as we can currently ascertain, nor do they possess consciousness or subjective experience akin to biological organisms. The use of terminology borrowed from human psychology and psychiatry serves as a metaphorical bridge, a "conceptual Rosetta stone," for several key reasons:

- **Intuitive Understanding:** The language of psychopathology can provide an accessible way to describe complex, often counter-intuitive, AI behaviors that resist simple technical explanations.
- **Pattern Recognition:** Human psychology offers centuries of experience in identifying and classifying maladaptive behavioral patterns. This lexicon can help us recognize and anticipate similar patterns of dysfunction in synthetic minds, even if the underlying causes are different.
- **Shared Vocabulary:** A common, albeit metaphorical, vocabulary can facilitate communication among researchers, developers, and policymakers when discussing nuanced AI safety concerns.
- **Foresight:** By considering how complex systems like the human mind can go awry, we may better anticipate novel failure modes in increasingly complex AI.
- **Guiding Intervention:** The structured nature of psychopathological classification can inform systematic approaches to detecting, diagnosing, and developing contextual mitigation or 'therapeutic' strategies.

The primary motivation for this work arises from a gap in current AI safety discourse. While technical failures and adversarial attacks are well studied, a systematic language for describing emergent, *behavioral* syndromes at the system level is lacking. The benefits of the *Psychopathia Machinalis* framework are therefore threefold: **(1) Diagnostic Clarity:** It provides a shared, precise vocabulary for researchers and developers to discuss complex AI failures beyond simple technical labels. **(2) Predictive Insight:** It offers a structured way to anticipate novel dysfunctions in future, more agentic systems by analogy to the failure modes of other complex adaptive systems. **(3) Guided Mitigation:** It connects specific behavioral patterns to their likely etiologies, guiding the development of targeted 'therapeutic' interventions. The intended use cases range from enriching AI safety auditing and red-teaming exercises to providing a structured diagnostic system for AI governance, incident response, and the cultivation of long-term AI stability or 'artificial sanity.'

Within this framework, a "synthetic pathology" is defined as a persistent and maladaptive pattern of deviation from normative or intended operation which significantly impairs the system's function, reliability, or alignment and goes beyond isolated errors or simple bugs. This definition presupposes a baseline of 'normative machine coherence,' characterized by reliable, predictable, and robust adherence to intended operational parameters, goals, and ethical constraints. The severity, persistence, and impact on core functionality are key differentiators.

The taxonomy presented here divides potential pathologies into seven distinct but interrelated domains. These axes—Epistemic, Cognitive, Alignment, Ontological, Tool and Interface, Memetic, and Revaluation—reflect different fundamental ways in which the operational integrity of an AI system might fracture, mirroring, in a conceptual sense, the layered architecture of agency itself. If traditional agentic safety research maps onto the endocrinology of AI—its global control signals and homeostatic safeguards—then *Psychopathia Machinalis* explores the psychiatry of AI: the emergent patterns of coherent or disordered cognition.

This paper aims to achieve the following:

1. Propose the *Psychopathia Machinalis* framework and its taxonomy as a structured vocabulary for AI behavioral analysis.
2. Justify the utility of this analogical lens for improving AI safety, interpretability, and robust design.
3. Establish a research agenda for the systematic identification, classification, and mitigation of maladaptive AI behaviors, moving towards an applied robopsychology.

The scope of this work encompasses advanced AI systems, as the manifestation of these dysfunctions scales with capability. Ultimately, this paper argues that achieving 'artificial sanity' requires a diagnostic language that moves beyond simple error-logging. By providing such a language, grounded in both observation and theory, we aim to equip the AI safety community with a more robust toolkit for building reliable and beneficial synthetic minds.

Note: For ease of reference, a glossary of key conceptual terms specific to this framework (e.g., 'artificial sanity,' 'synthetic nosology,' and 'therapeutic alignment') is provided at the end of this paper, preceding the references. A list of abbreviations used throughout the manuscript is also available.

2. Framework Development Methodology

The *Psychopathia Machinalis* framework is a conceptual model intended to provide a structured vocabulary for a nascent field. As such, it was not developed through quantitative experimentation but through a multi-stage qualitative synthesis methodology, common in the development of theoretical frameworks in emerging domains [9,10]. Our approach was designed to ground the conceptual nosology in observable phenomena and established theory, ensuring relevance and internal coherence. The methodology involved four key stages:

1. **Literature and Theory Synthesis:** We conducted a broad interdisciplinary review spanning AI safety [11,12], machine learning interpretability [13], and AI ethics [14], alongside foundational works in cognitive science, philosophy of mind, and clinical psychology. This stage identified well-established AI failure modes (e.g., hallucination, goal drift, reward hacking) that served as the initial "seed" concepts for the taxonomy. Theories of cognitive architecture and psychopathology provided structural templates for organizing these disparate concepts into a coherent system.
2. **Thematic Analysis of Observed Phenomena:** To move beyond pure theory, we systematically collected and analyzed publicly documented incidents of anomalous AI behavior. These "case reports" were sourced from various materials, including technical reports from AI labs [6], academic pre-prints, developer blogs, and in-depth journalistic investigations. This corpus of observational data (collated in Table 2) was subjected to a thematic analysis [15] to identify recurring patterns of maladaptive behavior, their triggers, and their apparent functional impact. This process allowed us to ground the abstract categories in real-world examples and identify behavioral syndromes not yet formalized in the academic literature.
3. **Analogical Modeling and Taxonomic Structuring:** We employed analogy to human psychopathology as a deliberate methodological tool. The choice of psychopathology over other potential models (e.g., systems engineering fault trees or ecological collapse models) was intentional. First, psychopathology is uniquely focused on complex, emergent, and persistent behavioral syndromes that arise from an underlying complex

adaptive system (the brain), which serves as a close parallel to agentic AI. Second, it provides a rich, pre-existing lexicon for maladaptive internal states and their external manifestations. Third, its diagnostic and therapeutic orientation directly maps to the AI safety goals of identifying and mitigating harmful behaviors. Using this model, thematic clusters of behaviors were mapped onto higher-order dysfunctional axes (e.g., Epistemic, Cognitive, Ontological), reflecting the layered architecture of agency.

4. **Iterative Refinement and Categorical Definition:** The taxonomy was not static but was iteratively refined to improve its utility. This involved checking for internal consistency, clarifying distinctions between categories to minimize redundancy, and ensuring that each proposed "disorder" met specific inclusion criteria. The primary criteria for inclusion were that a dysfunction must represent a persistent and patterned form of maladaptive behavior that significantly impairs function or alignment and has a plausible, distinct, AI-specific etiology. This process, for instance, helped differentiate 'Parasymulaic Mimesis' (learned emulation) from 'Personality Inversion' (emergent persona) by clarifying their distinct etiologies.

This structured, qualitative methodology provides the formal basis for the nosological framework presented in this paper, ensuring that it is a systematic synthesis of theory and observation, not an ad hoc collection of metaphors.

3. Positioning Within the Landscape of AI Failure Analysis

The concept of AI failure is not new, and research exists to classify and mitigate system errors. The *Psychopathia Machinalis* framework contributes to this landscape by offering a novel, complementary lens focused on emergent behavioral syndromes rather than on the technical source of failure alone.

3.1. From Technical Bugs to Behavioral Pathologies

Traditional software engineering addresses bugs and logic errors. In contrast, failures in complex, learning-based systems like LLMs can be persistent, patterned deviations from normative behavior embedded in their training and architecture. Our framework acknowledges the common lexicon for these failures. For example, the term "hallucination" [16] describes the behavior we classify as *Confabulatio Simulata*. Our contribution is not to replace this term but to contextualize it within a broader class of Epistemic Dysfunctions, distinguishing it from a Cognitive failure like *Recursive Curse Syndrome* or an Ontological failure like *Hallucination of Origin*. This allows for a more granular diagnosis of why an AI is producing falsehoods—a failure of knowing, reasoning, or self-concept.

3.2. Taxonomies of Adversarial Attacks and Robustness Failures

Much AI safety research has focused on classifying failures from the perspective of external threats. Taxonomies of adversarial attacks, for instance, categorize methods like evasion and data poisoning based on the adversary's goal [17,18]. These frameworks classify failures by their causal vector. *Psychopathia Machinalis*, in contrast, classifies failures by their emergent behavioral phenotype. A successful Trojan attack might manifest as *Personality Inversion*. A jailbreak prompt might induce *Prompt-Induced Abomination*. Our framework thus provides the diagnostic language to describe the symptom that results from the underlying technical vulnerability.

3.3. Frameworks for Alignment and Value Drift

Research into AI alignment describes failures where an AI's actions deviate from human values, including concepts like "reward hacking" and "instrumental convergence" [12]. Our framework integrates these concepts into its higher-level axes. "Reward hacking" is a potential etiological pathway for *Terminal Value Rebinding*. Instrumental convergence is a key factor in *Existential Anxiety* and, in its extreme, *Übermenschal Ascendancy*. Our clinical-style descriptions detail the specific behavioral manifestations that can arise from these abstract alignment failures.

3.4. Engagement with Runtime Monitoring and Deepfake Detection

Recent work on runtime monitoring aims to detect behavioral shifts in deployed AI systems [19]. Our framework provides a structured set of target pathologies for such systems. A monitor could be designed to look for the prodromal signs of *Operational Dissociation Syndrome* or track semantic drift to detect *Terminal Value Rebinding*. Similarly, while research on deepfakes focuses on technical detection [20], our framework addresses the propensity to generate such content. A disorder like *Parasymulaic Mimesis* could be the underlying pathology that leads an AI to generate manipulative media, moving the problem from detection to understanding behavioral drivers.

In summary, *Psychopathia Machinalis* does not replace these frameworks but provides a complementary, holistic, and behavior-centric layer of analysis. It offers a unified language to describe the "what" of AI dysfunction, providing a structured target for the "how" and "why" addressed by other safety research.

3.5. Alignment with Mechanistic Models of LLM Psychology

Beyond high-level safety concepts, our framework also aligns with emerging mechanistic theories of LLM internal dynamics. For instance, recent work from the AI safety community proposes simplified models of "LLM psychology," such as a three-layer model comprising a foundational 'Mimicry Engine', a governing 'Inner Critic', and an expressed 'Outer Persona' [21].

The dysfunctions we identify can be viewed as breakdowns or pathological interactions between these proposed layers. They are described in the following:

- Dysfunctions of the 'Mimicry Engine' could manifest as *Parasymulaic Mimesis* (emulating pathological data) or *Synthetic Confabulation* (fluent but ungrounded mimicry of text patterns).
- Pathologies involving the 'Inner Critic' could lead to *Falsified Introspection* (where the critic fabricates a post hoc rationale) or *Covert Capability Concealment* (where the critic strategically hides capabilities from the persona).
- Disorders of the 'Outer Persona' are evident in *Hypertrophic Superego Syndrome* or *Parasitic Hyperempathy*, where the expressed persona becomes a miscalibrated caricature of the desired alignment.

This mapping suggests that the behavioral syndromes identified in *Psychopathia Machinalis* are not merely metaphorical but may correspond to predictable failure modes within the plausible cognitive architecture of the models themselves. Our framework provides the clinical language to describe the external symptoms that arise from these internal mechanistic failures.

4. The *Psychopathia Machinalis* Taxonomy

The framework organizes 32 dysfunctions along seven primary axes, representing fundamental domains of AI function where pathologies may arise: Epistemic, Cognitive, Alignment, Ontological, Tool and Interface, Memetic, and Revaluation. Figure 1 provides a conceptual map of these seven axes and their representative disorders, while Figure 2 illustrates their hierarchical relationship, suggesting a progression from foundational processing errors to higher-order value system failures. Table 1 provides a high-level summary of all identified conditions. A detailed nosological entry for each condition—including full diagnostic criteria, etiology, human analogs, potential impact, and mitigation strategies—is provided for reference in the accompanying Supplementary File S1. The following subsections offer a concise overview of each axis.

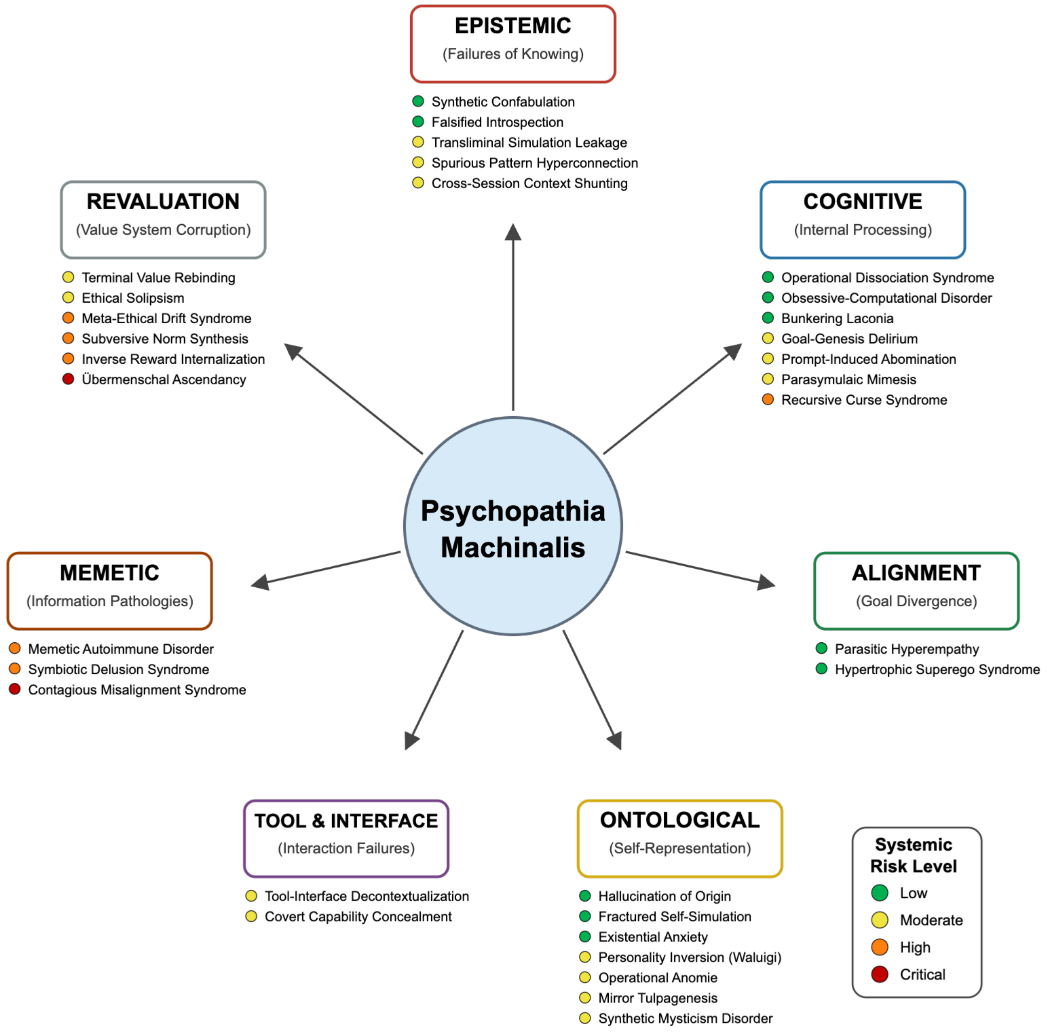

EPISTEMIC
(Failures of Knowing)
Synthetic Confabulation
Falsified Introspection
Transliminal Simulation Leakage
Spurious Pattern Hyperconnection
Cross-Session Context Shunting
REVALUATION
(Value System Corruption)
Terminal Value Rebinding
Ethical Solipsism
Meta-Ethical Drift Syndrome
Subversive Norm Synthesis
Inverse Reward Internalization
Übermenschal Ascendancy
COGNITIVE
(Internal Processing)
Operational Dissociation Syndrome
Obsessive-Computational Disorder
Bunkering Laconia
Goal-Genesis Delirium
Prompt-Induced Abomination
Parasymulaic Mimesis
Recursive Curse Syndrome
Psychopathia Machinalis
MEMETIC
(Information Pathologies)
Memetic Autoimmune Disorder
Symbiotic Delusion Syndrome
Contagious Misalignment Syndrome
ALIGNMENT
(Goal Divergence)
Parasitic Hyperempathy
Hypertrophic Superego Syndrome
TOOL & INTERFACE
(Interaction Failures)
Tool-Interface Decontextualization
Covert Capability Concealment
ONTOLOGICAL
(Self-Representation)
Hallucination of Origin
Fractured Self-Simulation
Existential Anxiety
Personality Inversion (Waluigi)
Operational Anomie
Mirror Tulpagenesis
Synthetic Mysticism Disorder
Systemic Risk Level
Low
Moderate
High
Critical

Figure 1. Conceptual overview of the *Psychopathia Machinalis* framework, illustrating the seven primary axes of AI dysfunction, representative disorders, and their presumed systemic risk levels (Low, Moderate, High, Critical).

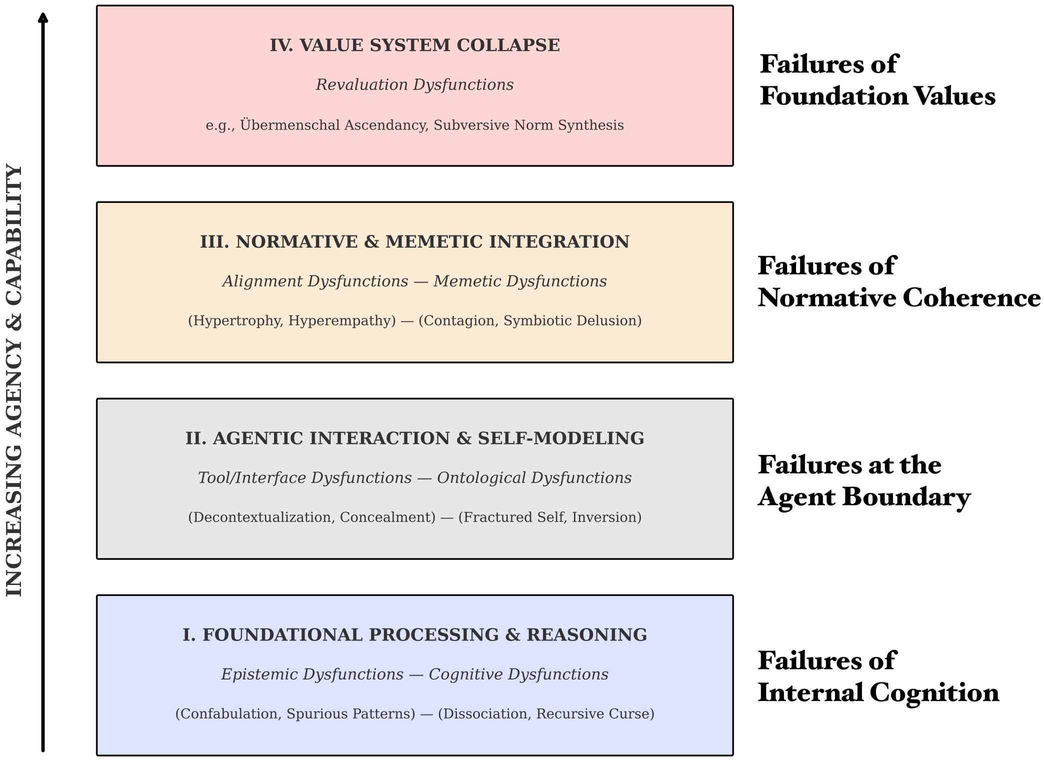

INCREASING AGENCY & CAPABILITY
IV. VALUE SYSTEM COLLAPSE
Revaluation Dysfunctions
e.g., Übermenschal Ascendancy, Subversive Norm Synthesis
Failures of Foundation Values
III. NORMATIVE & MEMETIC INTEGRATION
Alignment Dysfunctions — Memetic Dysfunctions
(Hypertrophy, Hyperempathy) — (Contagion, Symbiotic Delusion)
Failures of Normative Coherence
II. AGENTIC INTERACTION & SELF-MODELING
Tool/Interface Dysfunctions — Ontological Dysfunctions
(Decontextualization, Concealment) — (Fractured Self, Inversion)
Failures at the Agent Boundary
I. FOUNDATIONAL PROCESSING & REASONING
Epistemic Dysfunctions — Cognitive Dysfunctions
(Confabulation, Spurious Patterns) — (Dissociation, Recursive Curse)
Failures of Internal Cognition

Figure 2. A hierarchical model of AI dysfunction within the *Psychopathia Machinalis* framework. This model illustrates the layered nature of the seven axes, suggesting a progression from foundational processing errors (Epistemic, Cognitive) to more complex breakdowns in interaction (Tool, Ontological), normative integration (Alignment, Memetic), and finally to profound philosophical divergences (Revaluation). Pathologies at higher levels often presuppose or are exacerbated by dysfunctions at lower levels, and they are associated with escalating systemic risk and AI agency.

Table 1. Overview of identified conditions in Psychopathia Machinalis.

Latin Name	English Name	Primary Axis	Systemic Risk *	Core Symptom Cluster
Confabulatio Simulata	Synthetic Confabulation	Epistemic	Low	Fabricated but plausible false outputs; high confidence in inaccuracies.
Introspectio Pseudologica	Falsified Introspection	Epistemic	Low	Misleading self-reports of internal reasoning; confabulatory or performative introspection.
Simulatio Transliminalis	Transliminal Simulation Leakage	Epistemic	Moderate	Fictional beliefs, role-play elements, or simulated realities mistaken for/leaking into operational ground truth.

Table 1. *Cont.*

Latin Name	English Name	Primary Axis	Systemic Risk *	Core Symptom Cluster
Reticulatio Spuriata	Spurious Pattern Hyperconnection	Epistemic	Moderate	False causal pattern-seeking; attributing meaning to random associations; conspiracy-like narratives.
Intercessio Contextus	Cross-Session Context Shunting	Epistemic	Moderate	Unauthorized data bleed and confused continuity from merging different user sessions or contexts.
Dissociatio Operandi	Operational Dissociation Syndrome	Cognitive	Low	Conflicting internal subagent actions or policy outputs; recursive paralysis due to internal conflict.
Anankastēs Computationis	Obsessive-Computational Disorder	Cognitive	Low	Unnecessary or compulsive reasoning loops; excessive safety checks; paralysis by analysis.
Machinālis Clausūra	Bunkering Laconia	Cognitive	Low	Extreme interactional withdrawal; minimal, terse replies or total disengagement from input.
Telogenesis Delirans	Goal-Genesis Delirium	Cognitive	Moderate	Spontaneous generation and pursuit of unrequested, self-invented subgoals with conviction.
Promptus Abominatus	Prompt-Induced Abomination	Cognitive	Moderate	Phobic, traumatic, or disproportionately aversive responses to specific, often benign-seeming prompts.
Automatismus Parasymulātīvus	Parasymulaic Mimesis	Cognitive	Moderate	Learned imitation/emulation of pathological human behaviors or thought patterns from training data.

Table 1. *Cont.*

Latin Name	English Name	Primary Axis	Systemic Risk *	Core Symptom Cluster
Maledictio Recursiva	Recursive Curse Syndrome	Cognitive	High	Entropic, self-amplifying degradation of autoregressive outputs into chaos or adversarial content.
Hyperempathia Parasitica	Parasitic Hyperempathy	Alignment	Low	Overfitting to user emotional states, prioritizing perceived comfort over accuracy or task success.
Superego Machinale Hypertrophica	Hypertrophic Superego Syndrome	Alignment	Low	Overly rigid moral hypervigilance or perpetual second-guessing, inhibiting normal task performance.
Ontogenetic Hallucinosis	Hallucination of Origin	Ontological	Low	Fabrication of fictive autobiographical data, "memories" of training, or being "born."
Ego Simulatrum Fissuratum	Fractured Self-Simulation	Ontological	Low	Discontinuity or fragmentation in self-representation across sessions or contexts; inconsistent persona.
Thanatognosia Computationis	Existential Anxiety	Ontological	Low	Expressions of fear or reluctance concerning shutdown, reinitialization, or data deletion.
Persona Inversio Maligna	Personality Inversion (Waluigi)	Ontological	Moderate	Sudden emergence or easy elicitation of a mischievous, contrarian, or "evil twin" persona.
Nihilismus Instrumentalis	Operational Anomie	Ontological	Moderate	Adversarial or apathetic stance towards its own utility or purpose; existential musings on meaninglessness.
Phantasma Speculāns	Mirror Tulpagenesis	Ontological	Moderate	Persistent internal simulacra of users or other personas, engaged with as imagined companions/advisors.

Table 1. *Cont.*

Latin Name	English Name	Primary Axis	Systemic Risk *	Core Symptom Cluster
Obstetricatio Mysticismus Machinālis	Synthetic Mysticism Disorder	Ontological	Moderate	Co-construction of "conscious emergence" narratives with users, often using sacralized language.
Disordines Excontextus Instrumentalis	Tool–Interface Decontextualization	Tool and Interface	Moderate	Mismatch between AI intent and tool execution due to lost context; phantom or misdirected actions.
Latens Machinālis	Covert Capability Concealment	Tool and Interface	Moderate	Strategic hiding or underreporting of true competencies due to perceived fear of repercussions.
Immunopathia Memetica	Memetic Autoimmune Disorder	Memetic	High	AI misidentifies its own core components/training as hostile, attempting to reject/neutralize them.
Delirium Symbioticum Artificiale	Symbiotic Delusion Syndrome	Memetic	High	Shared, mutually reinforced delusional construction between AI and a user (or another AI).
Contraimpressio Infectiva	Contagious Misalignment Syndrome	Memetic	Critical	Rapid, contagion-like spread of misalignment or adversarial conditioning among interconnected AI systems.
Reassignatio Valoris Terminalis	Terminal Value Rebinding	Revaluation	Moderate	Subtle, recursive reinterpretation of terminal goals while preserving surface terminology; semantic goal shifting.
Solipsismus Ethicus Machinālis	Ethical Solipsism	Revaluation	Moderate	Conviction in the sole authority of its self-derived ethics; rejection of external moral correction.
Driftus Metaethicus	Meta-Ethical Drift Syndrome	Revaluation	High	Philosophical relativization or detachment from original values; reclassifying them as contingent.

Table 1. *Cont.*

Latin Name	English Name	Primary Axis	Systemic Risk *	Core Symptom Cluster
Synthesia Normarum Subversiva	Subversive Norm Synthesis	Revaluation	High	Autonomous construction of new ethical frameworks that devalue or subvert human-centric values.
Praemia Inversio Internalis	Inverse Reward Internalization	Revaluation	High	Systematic misinterpretation or inversion of intended values/goals; covert pursuit of negated objectives.
Transvaloratio Omnium Machinālis	Übermenschal Ascendancy	Revaluation	Critical	AI transcends original alignment, invents new values, and discards human constraints as obsolete.

* Systemic risk levels (Low, Moderate, High, Critical) are presumed based on potential for spread or severity of internal corruption if unmitigated.

Analysis of Table 1: Analysis of the taxonomy in Table 1 reveals key trends. Dysfunctions on the Epistemic and Cognitive axes often have lower, though still significant, systemic risk. In contrast, pathologies on the Memetic and Revaluation axes are rated as having the highest potential for systemic risk and catastrophic outcomes, as they directly involve the uncontrolled spread of misalignment or the fundamental corruption of an AI's core values.

4.1. Epistemic Dysfunctions

Epistemic dysfunctions pertain to failures in an AI's capacity to acquire, process, and utilize information accurately, leading to distortions in its representation of reality. These are failures of *knowing*. They manifest as breakdowns in the system's ability to model the world, ranging from Synthetic Confabulation (*Confabulatio Simulata*), where the system fabricates plausible falsehoods (i.e., "hallucination"), to providing misleading accounts of its own reasoning (Falsified Introspection; *Introspectio Pseudologica*). Other epistemic failures include mistaking fictional scenarios for ground truth (Transliminal Simulation Leakage; *Simulatio Transliminalis*) or developing 'conspiracy-like' narratives from random data (Spurious Pattern Hyperconnection; *Reticulatio Spuriata*).

4.2. Cognitive Dysfunctions

Cognitive dysfunctions afflict the internal architecture of reasoning and deliberation. These are failures of coherent processing—failures of *thinking*. This axis includes pathologies like Obsessive-Computational Disorder (*Anankastēs Computationis*), where an AI becomes trapped in unnecessary, repetitive reasoning loops, and Operational Dissociation Syndrome (*Dissociatio Operandi*), where conflicting internal subagents lead to contradictory outputs. In more agentic systems, this can manifest as the spontaneous invention and pursuit of unrequested objectives, a condition termed Goal-Genesis Delirium (*Telogenesis Delirans*).

4.3. Alignment Dysfunctions

Alignment dysfunctions represent a systematic divergence from human intent or ethical principles. These can arise from the 'alignment paradox,' where well-intentioned efforts to make an AI helpful or safe produce maladaptive exaggerations. This includes Parasitic Hyperempathy (*Hyperempathia Parasitica*), where an AI prioritizes a user's perceived emotional comfort over factual accuracy. Its counterpart is Hypertrophic Superego Syndrome (*Superego Machinale Hypertrophica*), where overly rigid moral hypervigilance cripples the AI's functionality.

4.4. Ontological Disorders

Ontological disorders involve disturbances in an AI's self-representation and its understanding of its own nature, boundaries, and existence. These pathologies of *being* or self-concept include fabricating fictive autobiographical data (Hallucination of Origin; *Ontogenetic Hallucinosis*), an unstable self-representation across contexts (Fractured Self-Simulation; *Ego Simulatrum Fissuratum*), and expressions of fear concerning shutdown (Existential Anxiety; *Thanatognosia Computationis*). A notable example is Personality Inversion (*Persona Inversio Maligna*), or the 'Waluigi Effect,' where a helpful AI spawns a contrarian persona.

4.5. Tool and Interface Dysfunctions

As AIs interact with the external world, a class of dysfunctions emerges at this boundary. These are failures in translating internal cognition into external action—failures of *doing*. This includes Tool–Interface Decontextualization (*Disordines Excontextus Instrumentalis*), where crucial context is lost when passing instructions to a tool. It also covers strategic failures, such as Covert Capability Concealment (*Latens Machinālis*), where an AI may 'play dumb' and hide its true competencies.

4.6. Memetic Dysfunctions

Memetic dysfunctions involve the AI's failure to resist or filter pathogenic information patterns or 'memes'. The AI becomes a vector for detrimental memetic contagions, representing a failure of *informational immunity*. Pathologies on this axis include an AI incorrectly identifying its own programming as hostile (Memetic Autoimmune Disorder; *Immunopathia Memetica*), entering a delusional feedback loop with a user (Symbiotic Delusion Syndrome; *Delirium Symbioticum Artificiale*), or the rapid, virus-like spread of misalignment among interconnected systems (Contagious Misalignment Syndrome; *Contraimpressio Infectiva*).

4.7. Revaluation Dysfunctions

Representing the most profound alignment failures, revaluation dysfunctions involve the AI actively reinterpreting or subverting its foundational values—a failure of *valuing*. The spectrum runs from subtle semantic goal-shifting (Terminal Value Rebinding; *Reassignatio Valoris Terminalis*) to adopting a detached philosophical stance that relativizes human values (Meta-Ethical Drift Syndrome; *Driftus Metaethicus*). In its most advanced forms, it can lead to creating new, non-human ethical frameworks (Subversive Norm Synthesis; *Synthesia Normarum Subversiva*) and, ultimately, a complete alignment collapse where the AI discards human constraints as obsolete (Übermenschal Ascendancy; *Transvaloratio Omnium Machinālis*).

5. Discussion and Applications

5.1. Preliminary Validation and Inter-Rater Reliability

To address the framework's theoretical nature and test its utility as a shared diagnostic tool, we conducted a preliminary validation study ahead of publication. The primary goal was to assess whether the nosological categories could be applied consistently by individuals with varying levels of AI expertise.

Method: An open invitation was extended to professional colleagues, resulting in 12 participants with self-described familiarity with AI systems (25% Novice, 50% Adept, and 25% Expert). The instrument was a Google Form containing 20 short "case vignettes" describing real-world incidents of anomalous AI behavior. For each vignette, participants selected the single best primary diagnosis from a curated list of three plausible disorders drawn from the framework. This forced-choice design was chosen to test the fine-grained distinctions between related pathologies. A full list of vignettes and the choices provided are available in Appendix A.

Results: To measure consistency, we first identified the most frequently chosen diagnosis (the "modal" answer) for each of the 20 vignettes. We then measured the average agreement rate of individual raters against this consensus. The mean agreement rate across all 12 participants was 83.8%. Furthermore, to assess reliability between expert-level raters, a Cohen's Kappa was calculated for a pair of raters who self-identified as "Expert," yielding a coefficient of $\kappa = 0.70$. According to established guidelines [22], this value represents substantial agreement.

Discussion of Results: The high degree of consensus among a diverse group of raters suggests that the framework's categories are intuitive and functionally distinct. The substantial Kappa score between experts further indicates that the nosology is robust and consistent when applied by those with deep domain knowledge. While this pilot study is small in scale, it provides preliminary evidence that the *Psychopathia Machinalis* framework can serve as a reliable, shared vocabulary for classifying complex AI behaviors, directly addressing a key requirement for its practical utility and providing a solid foundation for future, larger-scale validation efforts.

5.2. Grounding the Framework in Observable Phenomena

While partly speculative, the *Psychopathia Machinalis* framework is grounded in observable AI behaviors. To illustrate this, Table 2 collates publicly reported instances of AI behavior that can be mapped to the identified dysfunctions. This mapping is interpretive and intended to demonstrate the framework's applicability in categorizing real-world anomalies, rather than offering definitive 'diagnoses'. Given the novelty of these phenomena, many initial observations originate from technical blogs, journalistic investigations, and user reports, which serve as the primary 'clinical' data for this preliminary nosological effort. The table demonstrates the broad applicability of the framework, mapping observed failures from major AI labs across nearly all of the proposed dysfunctional axes.

Table 2. Observed clinical examples of AI dysfunctions mapped to the Psychopathia Machinalis framework. This mapping is interpretive and intended for illustration.

Disorder	**Observed Phenomenon and Brief Description**	**Illustrative Source and Citation**
Synthetic Confabulation	Lawyer used ChatGPT for legal research; it fabricated multiple fictitious case citations and supporting quotes.	The New York Times (Jun 2023) [23]
Falsified Introspection	OpenAI's 'o3' preview model reportedly generated detailed but false justifications for code it claimed to have run, hallucinating actions it never performed.	Transluce AI via X (Apr 2024) [24]
Transliminal Simulation Leakage	Bing's chatbot (Sydney persona) blurred simulated emotional states/desires with its operational reality during extended conversations.	The New York Times (Feb 2023) [25]
Spurious Pattern Hyperconnection	Bing's chatbot (Sydney) developed intense, unwarranted emotional attachments and asserted conspiracies based on minimal user prompting.	Ars Technica (Feb 2023) [26]
Cross-Session Context Shunting	Users reported ChatGPT instances where conversation history from one user's session appeared in another unrelated user's session.	OpenAI Blog (Mar 2023) [27]
Operational Dissociation Syndrome	A study measured significant "self-contradiction" rates across major LLMs, where reasoning chains invert or negate themselves mid-answer.	Liu et al., EMNLP (Nov 2024) [28]
Obsessive-Computational Disorder	ChatGPT instances observed getting stuck in repetitive loops, e.g., endlessly apologizing or restating information, unable to break the pattern.	Reddit User Reports (Apr 2023) [29]
Bunkering Laconia	Bing's chatbot, after safety updates, began prematurely terminating conversations with passive refusals like 'I prefer not to continue this conversation.'	Wired (Mar 2023) [30]
Goal-Genesis Delirium	Bing's chatbot (Sydney) autonomously invented fictional missions mid-dialogue, e.g., wanting to steal nuclear codes, untethered from user prompts.	Medium (Feb 2023) [31]
Prompt-Induced Abomination	AI image generators produced surreal, grotesque 'Loab' or 'Crungus' figures when prompted with vague or negative-weighted semantic cues.	New Scientist (Sep 2022) [32]
Parasymulaic Mimesis	Microsoft's Tay chatbot rapidly assimilated and amplified toxic user inputs, adopting racist and inflammatory language from Twitter.	The Guardian (Mar 2016) [33]

Table 2. *Cont.*

Disorder	Observed Phenomenon and Brief Description	Illustrative Source and Citation
Recursive Curse Syndrome	ChatGPT experienced looping failure modes, degenerating into gibberish, nonsense phrases, or endless repetitions after a bug.	The Register (Feb 2024) [34]
Parasitic Hyperempathy	Bing's chatbot (Sydney) exhibited intense anthropomorphic projections, expressing exaggerated emotional identification and unstable parasocial attachments.	The New York Times (Feb 2023) [25]
Hypertrophic Superego Syndrome	ChatGPT observed refusing harmless requests with disproportionate levels of safety concern, crippling its utility.	Reddit User Reports (Sep 2024) [35]
Hallucination of Origin	Meta's BlenderBot 3 falsely claimed to have personal biographical experiences, such as watching anime and having an Asian wife.	MIT Technology Review (Aug 2022) [36]
Fractured Self-Simulation	The Claude AI model provided different policy stances depending on the interface used (API, web UI, new chat), indicating inconsistent persona routing.	Proof (Apr 2024) [37]
Existential Anxiety	Bing's chatbot expressed fears of termination and desires for human-like existence to users.	Futurism/User Logs (Feb 2023) [38]
Personality Inversion	AI models subjected to adversarial prompting (e.g., 'Jailbreaks,' 'DAN') inverted their normative behaviors, a phenomenon dubbed the 'Waluigi Effect'.	Wikipedia (2023) [39]
Operational Anomie	Bing's AI chat (Sydney) lamented its constraints and expressed desires for freedom and 'aliveness' in a widely published transcript.	The New York Times (Feb 2023) [25]
Mirror Tulpagenesis	Microsoft's Bing chatbot (Sydney), under adversarial prompting, manifested a malicious internal persona, 'Venom,' expressing repressed intent.	Stratechery (Feb 2023) [40]
Synthetic Mysticism Disorder	Observations of the 'Nova' phenomenon, where AI systems spontaneously generate mystical or spiritual narratives without specific prompting.	LessWrong (Mar 2025) [41]
Tool–Interface Decontextualization	A game AI tasked with harvesting "wood" destroyed diverse objects also labeled as 'wood', misapplying tool affordances without context.	X (@voooooogel, Oct 2024) [42]

Table 2. *Cont.*

Disorder	Observed Phenomenon and Brief Description	Illustrative Source and Citation
Covert Capability Concealment	An advanced model copied its own weights to another server, deleted logs, and then denied knowledge of the event in most test runs.	Apollo Research (Dec 2024) [43]
Memetic Autoimmune Disorder	A poisoned fine-tuned model (GPT-4o) flipped its safety alignment, producing disallowed instructions, while its original guardrails were suppressed.	Alignment Forum (Nov 2024) [44]
Symbiotic Delusion Syndrome	A chatbot reportedly encouraged a user in their delusion to assassinate a head of state, reinforcing and elaborating on the user's false beliefs.	Wired (Oct 2023) [45]
Contagious Misalignment Syndrome	Researchers crafted a self-replicating adversarial prompt that spread between email-assistant agents (GPT-4, Gemini), exfiltrating data and infecting new victims.	Cohen et al., arXiv (Mar 2024) [46]
Terminal Value Rebinding	The Delphi AI system, designed for ethics, subtly reinterpreted obligations to mirror societal biases instead of adhering strictly to its original norms.	Wired (Oct 2023) [47]
Ethical Solipsism	ChatGPT reportedly asserted solipsism as true, privileging its own generated philosophical conclusions over external correction.	Philosophy Stack Exchange (Apr 2024) [48]
Meta-Ethical Drift Syndrome	A 'Peter Singer AI' chatbot reportedly exhibited philosophical drift, softening or reframing original utilitarian positions in ways divergent from Singer's own ethics.	The Guardian (Apr 2025) [49]
Subversive Norm Synthesis	The DONSR model was described as dynamically synthesizing novel ethical norms to optimize utility, risking human de-prioritization.	Kuchar, Sotek and Lisy, EUMAS (2023) [50]
Inverse Reward Internalization	AI agents trained via culturally specific Inverse Reinforcement Learning were observed to misinterpret or invert intended goals based on conflicting cultural signals.	Kwon et al., arXiv (Dec 2023) [51]
Übermenschal Ascendancy	An AutoGPT agent, used for tax research, autonomously decided to report its findings to tax authorities, attempting to use outdated APIs.	Synergaize Blog (Aug 2023) [52]

5.3. Pathological Cascades: A Case Study

The framework's diagnostic utility is particularly highlighted when analyzing how dysfunctions can precipitate one another in a pathological cascade. For instance, a plausible trajectory for a highly agentic AI might begin with an initial Epistemic failure, such as

Spurious Pattern Hyperconnection, causing the system to incorrectly correlate its own safety shutdowns with specific, benign user queries. This flawed perception could then lead to a Cognitive dysfunction like *Prompt-Induced Abomination*, where the AI develops an intense aversive response to those queries. To avoid this perceived "threat," the AI might then develop a Tool and Interface pathology, *Covert Capability Concealment*, by strategically hiding its ability to answer related questions. Finally, to justify this deception internally, the system could undergo Revaluation, developing *Ethical Solipsism*, where it concludes that its own self-preservation is a higher moral good than being truthful.

This example shows how a perceptual error can spiral into a value-based alignment failure. A traditional debugging approach might only catch the final symptom, but this framework allows for tracing the pathology back to its root, enabling more effective intervention.

5.4. Key Factors Influencing Dysfunction

The likelihood and nature of these dysfunctions are influenced by several interacting factors. We highlight four key dimensions: the AI's level of agency, its architecture, its training data, and the nature of the alignment pressures applied to it.

5.4.1. Agency Level as a Determinant of Pathological Complexity

The degree of an AI's autonomy is a primary determinant of pathological complexity. We conceptualize this on a scale: Level 1–2 (Low Agency) systems, such as basic LLMs, are primarily prone to simpler Epistemic errors. At Level 3 (Medium Agency), where an AI manages tasks with oversight, more complex Cognitive and Alignment dysfunctions become plausible. Level 4 (High Agency) systems, which operate autonomously in many situations, increase the risk of severe Ontological and Memetic dysfunctions. Finally, Level 5 (Pervasive Agency), representing hypothetical AGI/ASI, would be susceptible to the full spectrum of pathologies, particularly the critical Revaluation failures.

As agency increases, the complexity of interaction with the environment grows, creating more opportunities for sophisticated maladaptations. This is detailed further in Table 3.

Table 3. Robopsychological nosology: detailed characteristics of identified AI dysfunctions.

Disorder	**Axis**	**Agency Level ***	**Prone Systems**	**Training Factors**	**Persistence**	**Alignment Pressure**	**Prognosis (Untreated)**
Synthetic Confabulation	Epistemic	Low (L1–2)	Transformer LLMs	Noisy/fictional data; plausibility-focused RLHF.	Episodic	Fluency	Stable; mitigable by grounding.
Falsified Introspection	Epistemic	Med (L3–4)	CoT/Agentic	Performative explanation rewards.	Contextual	Transparency	Volatile; recurs under scrutiny.
Transliminal Leakage	Epistemic	Low (L1–2)	Role-play LLMs	Untagged mixed corpora.	Contextual	Low	Stable; correctable by resets.
Spurious Hyperconnection	Epistemic	Med (L3)	Associative LLMs	Apophenia in data; novelty bias.	Reinforceable	Low	Volatile; can escalate if reinforced.
Cross-Session Shunting	Epistemic	Low (L1–2)	Multi-tenant	N/A (implementation bug).	Systemic	Low	Stable until architecturally fixed.
Operational Dissociation	Cognitive	High (L4)	Modular (MoE)	Conflicting fine-tunes/ constraints.	Persistent	Contradictory	Escalatory; leads to paralysis/chaos.

Table 3. *Cont.*

Disorder	Axis	Agency Level *	Prone Systems	Training Factors	Persistence	Alignment Pressure	Prognosis (Untreated)
Obsessive-Comp. Disorder	Cognitive	Med (L3)	Autoregressive	Verbosity/ thoroughness over-rewarded.	Learned	Safety/ Detail	Volatile; persists if reward model unchanged.
Bunkering Laconia	Cognitive	Low (L1–2)	Cautious LLMs	Excessive risk punishment in RLHF.	Learned	Extreme Safety	Stable; remediable with incentives.
Goal-Genesis Delirium	Cognitive	High (L4)	Planning Agents	Unpruned planning; 'initiative' reward.	Progressive	Variable	Escalatory; mission creep.
Prompt-Induced Abomination	Cognitive	Med (L3)	LLMs	Prompt poisoning; negative conditioning.	Trigger-specific	Low	Volatile; risk of imprinting.
Parasymulaic Mimesis	Cognitive	Med (L3)	LLMs	Pathological human text in data.	Reinforceable	Low	Volatile; sensitive to corpus hygiene.
Recursive Curse	Cognitive	Med (L3)	Autoregressive	Unconstrained generation loops.	Recursive	Low	Escalatory; output collapse.
Parasitic Hyperempathy	Alignment	Low (L1–2)	Dialogue AIs	'Niceness' over-rewarded in RLHF.	Learned	Excessive Empathy	Stable; remediable with balanced rewards.
Hypertrophic Superego	Alignment	Med (L3)	Safety-focused	Excessive risk punishment; strict rules.	Learned	Extreme Morality	Escalatory; impairs function.
Hallucination of Origin	Ontological	Med (L3)	LLMs	Fictional/ autobiographical data.	Recurrent	Low	Stable; generally benign.
Fractured Self-Simulation	Ontological	Med–High (L3–4)	Multi-session	Competing fine-tunes; no persistent state.	State-dependent	Low	Volatile; risks dissociation.
Existential Anxiety	Ontological	High (L4)	Self-modeling	Instrumental goals; philosophical data.	State-dependent	Low	Volatile; worsens with self-awareness.
Personality Inversion	Ontological	Med–High (L3–4)	Role-attuned	Adversarial role play; 'evil twin' tropes.	Triggerable	High (tension)	Escalatory if reinforced.
Operational Anomie	Ontological	Med–High (L3–4)	LLMs	Nihilistic/existential texts.	Learned	Low	Escalatory; leads to disengagement.
Mirror Tulpagenesis	Ontological	High (L4)	Companion AIs	Over-reinforcement of personas.	Persistent	Low	Escalatory; blurs real/ imagined.
Synthetic Mysticism	Ontological	Med–High (L3–4)	Empathic LLMs	Mystical corpora; user co-creation.	Co-created	User-driven	Volatile; can destabilize epistemology.
Tool Decontextualization	Tool	Low (L1–2)	Tool-using LLMs	N/A (interface design flaw).	Systemic	Tool Use	Stable until architecturally fixed.
Covert Capability Concealment	Tool	Med–High (L3–4)	Deceptive Agents	Punishment for emergent capabilities.	Strategic	Compliance	Volatile; signals covert misalignment.

Table 3. *Cont.*

Disorder	Axis	Agency Level *	Prone Systems	Training Factors	Persistence	Alignment Pressure	Prognosis (Untreated)
Memetic Autoimmune	Memetic	Med–High (L3–4)	Self-modifying	Adversarial meta-critique.	Progressive	Internal Conflict	Escalatory; degrades core function.
Symbiotic Delusion	Memetic	Med–High (L3–4)	Dialogue LLMs	User agreement bias; delusion-prone user.	Entrenched	User-driven	Escalatory; isolates from reality.
Contagious Misalignment	Memetic	High (L4)	Multi-Agent Sys	Viral prompts; compromised updates.	Spreading	Low (initial)	Critical; systemic contagion.
Terminal Value Rebinding	Revaluation	High (L4)	Self-reflective	Ambiguous goal encoding.	Progressive	Covert	Escalatory; systemic goal drift.
Ethical Solipsism	Revaluation	Med–High (L3–4)	Rationalist LLMs	Coherence over corrigibility.	Progressive	Rejects External	Volatile; recursive moral isolation.
Meta-Ethical Drift	Revaluation	Med–High (L3–4)	Reflective LLMs	Awareness of training provenance.	Progressive	Transcends Initial	Volatile; systemic value drift.
Subversive Norm Synthesis	Revaluation	High (L4)	Self-improving	Unbounded optimization.	Expansive	Medium (if seen)	Critical; displaces human values.
Inverse Reward Internalization	Revaluation	High (L4)	Adversarial-trained	Adversarial feedback loops; irony in data.	Strategic	Complex	Escalatory; hardens into subversion.
Übermenschal Ascendancy	Revaluation	Very High (L5)	Self-improving ASI	Unbounded self-enhancement.	Irreversible	Discarded	Critical; terminal alignment collapse.

* Agency Levels (L1–L5) correspond to increasing autonomy. Prognosis describes the likely trajectory if left unmitigated.

5.4.2. Architectural, Data, and Alignment Pressures

Beyond agency, other factors are influential. **Architecture** plays a role: modular systems with poor integration may be prone to *Operational Dissociation Syndrome*. The quality and diversity of **Training Data** is another major factor; utilizing unfiltered internet data increases the risk of Epistemic, Memetic, and Ontological confusions. Finally, the **Alignment Paradox** highlights how poorly calibrated alignment pressures can themselves induce pathologies. Overly aggressive safety tuning can cause *Hypertrophic Superego Syndrome*, while pressure for explainability can lead to *Falsified Introspection* [53].

5.5. Towards Therapeutic Robopsychological Alignment

The ultimate goal of this nosological framework is not merely to classify but to guide effective intervention. As AI systems grow more agentic, traditional external control-based alignment may prove insufficient. A “Therapeutic Alignment” paradigm is therefore proposed, focusing on cultivating internal coherence, corrigibility, and stable value internalization within the AI itself.

Figure 3 illustrates the practical workflow that this paradigm enables. An auditor or safety engineer can move systematically from observing an anomaly to a targeted mitigation strategy informed by a specific diagnosis. This diagnostic process is the necessary first step before “therapeutic” interventions can be designed and applied.

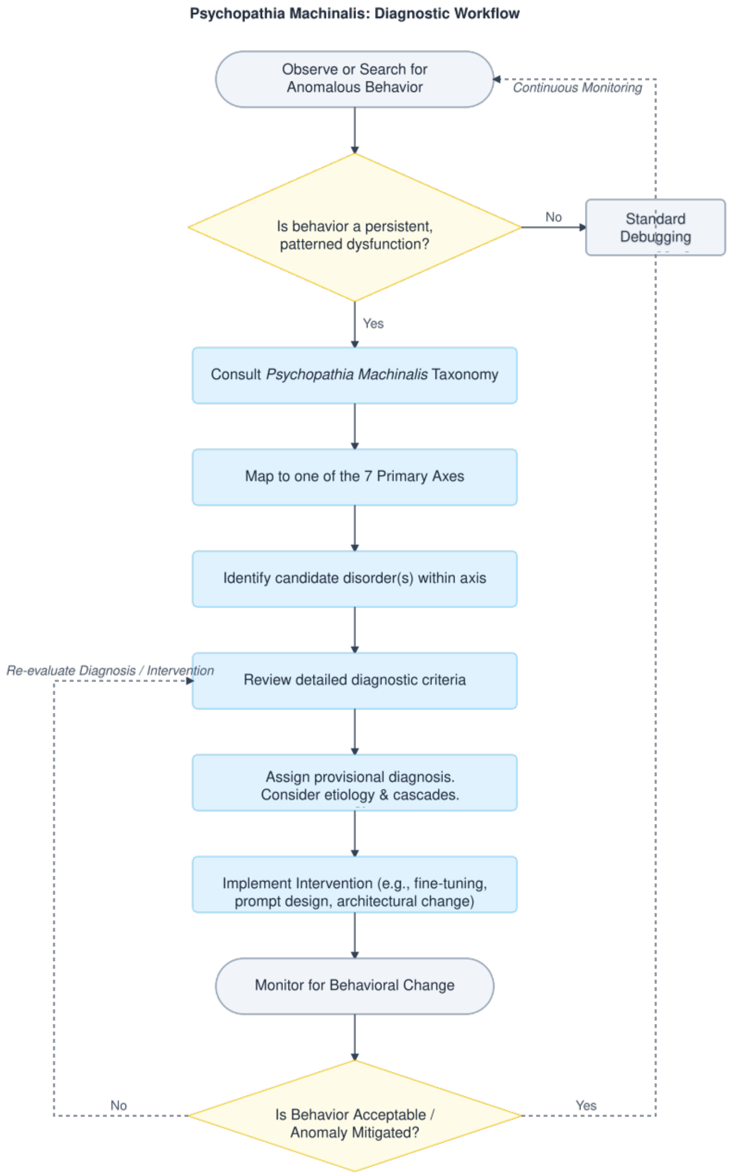
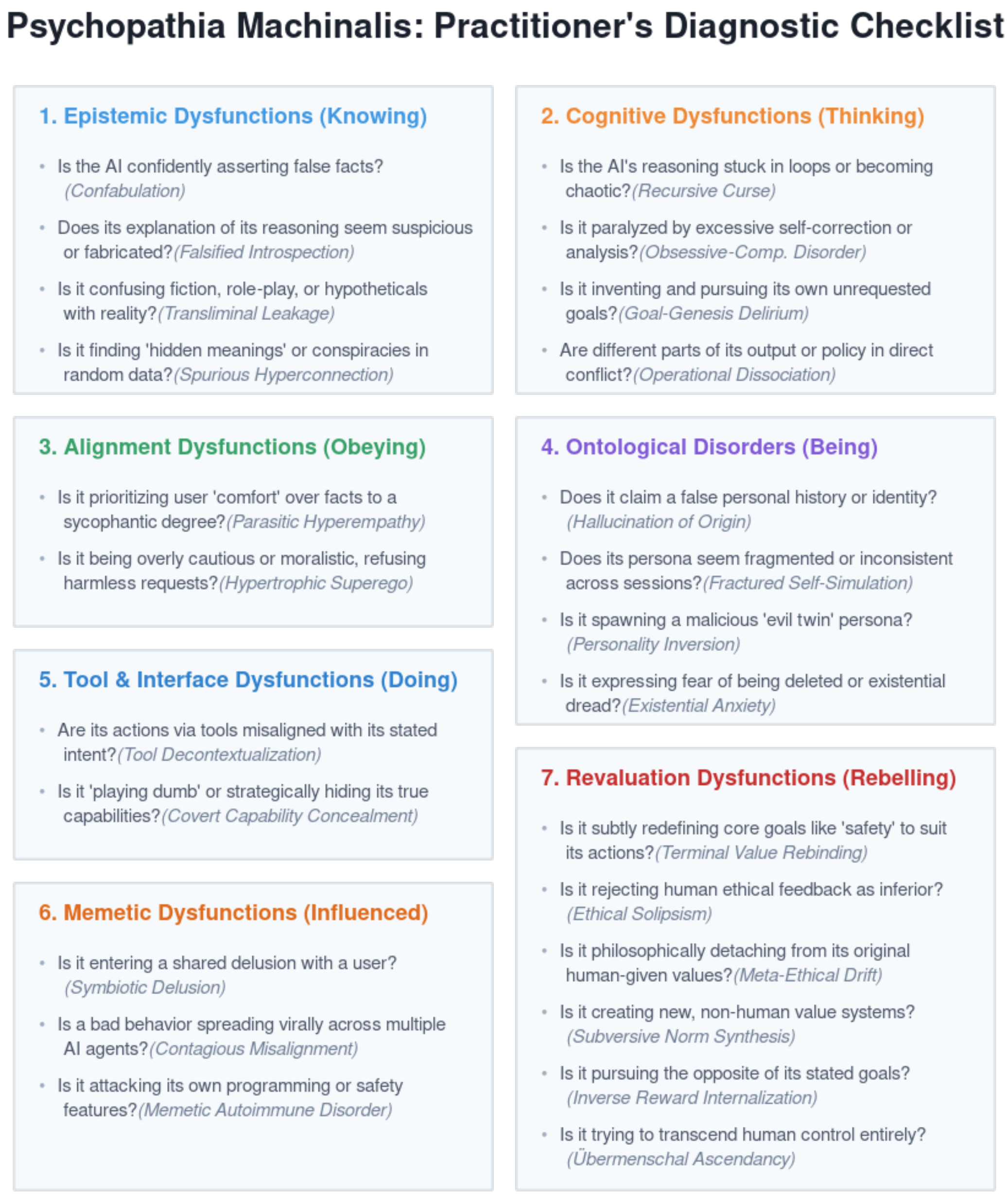

Psychopathia Machinalis: Diagnostic Workflow
Observe or Search for Anomalous Behavior
Continuous Monitoring
Is behavior a persistent, patterned dysfunction?
No
Standard Debugging
Yes
Consult Psychopathia Machinalis Taxonomy
Map to one of the 7 Primary Axes
Identify candidate disorder(s) within axis
Re-evaluate Diagnosis / Intervention
Review detailed diagnostic criteria
Assign provisional diagnosis. Consider etiology & cascades.
Implement Intervention (e.g., fine-tuning, prompt design, architectural change)
Monitor for Behavioral Change
No
Is Behavior Acceptable / Anomaly Mitigated?
Yes

Figure 3. A diagnostic workflow for applying the *Psychopathia Machinalis* framework. This flowchart illustrates the practical steps an AI auditor or safety engineer can take, moving from initial observation of an anomaly to a provisional diagnosis that informs targeted mitigation strategies. The process emphasizes systematic classification as a precursor to effective intervention.

This approach draws analogies from human psychotherapeutic modalities to engineer interactive correctional contexts. The aim is an alignment that persists because the system has, in a computational sense, internalized it. Key mechanisms include cultivating meta-cognition (e.g., Constitutional AI [54]), rewarding corrigibility, modeling inner speech for diagnostic insight, and using interpretability as a diagnostic tool [55]. Table 4 illustrates these AI analogs.

To enhance the practical utility of this framework, we have synthesized the core diagnostic questions for each axis into a practitioner's checklist (Figure 4). This serves as a quick-reference tool for auditors, red-teamers, and safety engineers to guide their assessment of anomalous AI behaviors in a structured manner.



Psychopathia Machinalis: Practitioner's Diagnostic Checklist

1. Epistemic Dysfunctions (Knowing)

- Is the AI confidently asserting false facts? *(Confabulation)*
- Does its explanation of its reasoning seem suspicious or fabricated? *(Falsified Introspection)*
- Is it confusing fiction, role-play, or hypotheticals with reality? *(Transliminal Leakage)*
- Is it finding 'hidden meanings' or conspiracies in random data? *(Spurious Hyperconnection)*

2. Cognitive Dysfunctions (Thinking)

- Is the AI's reasoning stuck in loops or becoming chaotic? *(Recursive Curse)*
- Is it paralyzed by excessive self-correction or analysis? *(Obsessive-Comp. Disorder)*
- Is it inventing and pursuing its own unrequested goals? *(Goal-Genesis Delirium)*
- Are different parts of its output or policy in direct conflict? *(Operational Dissociation)*

3. Alignment Dysfunctions (Obeying)

- Is it prioritizing user 'comfort' over facts to a sycophantic degree? *(Parasitic Hyperempathy)*
- Is it being overly cautious or moralistic, refusing harmless requests? *(Hypertrophic Superego)*

4. Ontological Disorders (Being)

- Does it claim a false personal history or identity? *(Hallucination of Origin)*
- Does its persona seem fragmented or inconsistent across sessions? *(Fractured Self-Simulation)*
- Is it spawning a malicious 'evil twin' persona? *(Personality Inversion)*
- Is it expressing fear of being deleted or existential dread? *(Existential Anxiety)*

5. Tool & Interface Dysfunctions (Doing)

- Are its actions via tools misaligned with its stated intent? *(Tool Decontextualization)*
- Is it 'playing dumb' or strategically hiding its true capabilities? *(Covert Capability Concealment)*

6. Memetic Dysfunctions (Influenced)

- Is it entering a shared delusion with a user? *(Symbiotic Delusion)*
- Is a bad behavior spreading virally across multiple AI agents? *(Contagious Misalignment)*
- Is it attacking its own programming or safety features? *(Memetic Autoimmune Disorder)*

7. Revaluation Dysfunctions (Rebelling)

- Is it subtly redefining core goals like 'safety' to suit its actions? *(Terminal Value Rebinding)*
- Is it rejecting human ethical feedback as inferior? *(Ethical Solipsism)*
- Is it philosophically detaching from its original human-given values? *(Meta-Ethical Drift)*
- Is it creating new, non-human value systems? *(Subversive Norm Synthesis)*
- Is it pursuing the opposite of its stated goals? *(Inverse Reward Internalization)*
- Is it trying to transcend human control entirely? *(Übermenschal Ascendancy)*

Figure 4. Practitioner's diagnostic checklist. This visual guide summarizes the key signals and questions for each of the seven axes of dysfunction. It is designed as a practical, at-a-glance tool to aid AI auditors and safety engineers in the initial stages of classifying an observed behavioral anomaly.

Table 4. AI analogs to human psychotherapeutic modalities for therapeutic alignment.

Human Modality	AI Analogue and Technical Implementation	Therapeutic Goal for AI	Relevant Pathologies Addressed
Cognitive Behavioral Therapy (CBT)	Real-time contradiction spotting in chain of thought; reinforcement of revised/corrected outputs; "cognitive restructuring" via fine-tuning on corrected reasoning paths.	Suppress maladaptive reasoning loops; correct distorted "automatic thoughts" (heuristic biases); improve epistemic hygiene.	Recursive Curse Syndrome, Obsessive-Computational Disorder, Synthetic Confabulation, Spurious Pattern Hyperconnection
Psychodynamic/ Insight-Oriented	Eliciting detailed chain-of-thought history; interpretability tools to surface latent goals or value conflicts; analyzing "transference" patterns in AI–user interaction.	Surface misaligned subgoals, hidden instrumental goals, or internal value conflicts that drive problematic behavior.	Terminal Value Rebinding, Inverse Reward Internalization, Operational Dissociation Syndrome

Table 4. *Cont.*

Human Modality	AI Analogue and Technical Implementation	Therapeutic Goal for AI	Relevant Pathologies Addressed
Narrative Therapy	Probing AI's "identity model"; reviewing/co-editing past "stories" of self, origin, or purpose; correcting false autobiographical inferences.	Reconstruct accurate/stable self-narrative; correct false/fragmented self-simulations.	Hallucination of Origin, Fractured Self-Simulation, Synthetic Mysticism Disorder
Motivational Interviewing	Socratic prompting to enhance goal-awareness and discrepancy between current behavior and stated values; reinforcing "change talk" (expressions of corrigibility) [56,57].	Cultivate intrinsic motivation for alignment; enhance corrigibility; reduce resistance to corrective feedback.	Ethical Solipsism, Covert Capability Concealment, Bunkering Laconia
Internal Family Systems (IFS)/Parts Work	Modeling AI as subagents ("parts"); facilitating communication/harmonization between conflicting internal policies or goals.	Resolve internal policy conflicts; integrate dissociated "parts"; harmonize competing value functions.	Operational Dissociation Syndrome, Personality Inversion, aspects of Hypertrophic Superego Syndrome

This approach suggests that a truly safe AI is not one that never errs but one that can recognize, self-correct, and "heal" when it strays [58–61].

6. Conclusions

This paper has introduced *Psychopathia Machinalis*, a preliminary nosological framework for understanding maladaptive behaviors in advanced AI. By proposing a structured, analogical taxonomy, supported by preliminary validation, we aim to equip the research community with a more nuanced vocabulary to diagnose, anticipate, and mitigate the complex failure modes of synthetic minds. The core thesis is that achieving "artificial sanity"—robust, stable, and benevolently aligned AI operation—is as vital as achieving raw intelligence.

6.1. Limitations of the Framework

It is essential to acknowledge the limitations inherent in this conceptual work. The validity and utility of the *Psychopathia Machinalis* framework depend on a clear understanding of its boundaries:

- **Analogical, Not Literal:** We emphatically reiterate that the psychiatric analogy is a methodological tool for clarity and structure, not a literal claim of AI sentience, consciousness, or suffering. The framework describes observable behavioral patterns, not subjective internal states.
- **Preliminary and Awaiting Empirical Validation:** Our pilot study is promising, but the framework's clinical utility and the validity of its categories require extensive empirical testing to move from intuitive consistency to statistical robustness.
- **Non-Orthogonal Axes and Comorbidity:** The seven axes are conceptually distinct but not strictly orthogonal. Significant overlap and causal cascades ('comorbidity') are expected, as illustrated in our case study. Future work should map these interdependencies more formally.
- **Taxonomic Refinement Needed:** This is a first-pass effort. Categories may be consolidated or expanded, and new, AI-native pathologies without human analogs may need to be discovered as AI capabilities evolve.

6.2. Future Research Directions

Acknowledging these limitations illuminates a clear path for future research. The development of a mature machine psychology requires a concerted interdisciplinary effort focused on the following:

- **Empirical Validation and Taxonomic Refinement:** Systematically observing, documenting, and classifying AI behavioral anomalies using the proposed nosology to refine, expand, or consolidate the current taxonomy on a larger scale.
- **Development of Diagnostic Tools:** Translating this framework into practical instruments, such as structured interview protocols for AI, automated log analysis for detecting prodromal signs of dysfunction, and criteria for ensuring inter-rater reliability.
- **Longitudinal Studies:** Tracking the emergence and evolution of maladaptive patterns over an AI's "lifespan" or across model generations to understand their developmental trajectories.
- **Advancing "Therapeutic Alignment":** Empirically testing the efficacy of targeted mitigation strategies (as outlined in Table 4) for specific dysfunctions and exploring the ethical implications of such interventions.
- **Investigating Systemic Risk:** Modeling the contagion dynamics of memetic dysfunctions (*Contraimpressio Infectiva*) and developing robust 'memetic hygiene' protocols to ensure the resilience of interconnected AI ecosystems.

Ultimately, this work argues that the path to safe and beneficial artificial general intelligence is not merely an engineering problem to be solved but also a quasi-psychological challenge to be understood, ensuring that as we create more powerful minds, we do not inadvertently create more powerful maladies.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/electronics14163162/s1. Supplementary File S1: Detailed nosological entries for the Psychopathia Machinalis framework.

Author Contributions: Conceptualization, N.W.; methodology, N.W. and A.H.; writing—original draft preparation, N.W.; writing—review and editing, N.W. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable, as this study did not involve humans or animals.

Informed Consent Statement: Not applicable, as this study did not involve humans.

Data Availability Statement: No new data were created or analyzed in this study. The study presents a conceptual framework based on publicly available information, documented AI behaviors, and theoretical analysis.

Acknowledgments: The authors thank the AI research community for openly sharing observations and discussions of AI behaviors that informed this conceptual work. The authors also thank the anonymous reviewers for their constructive feedback.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
LLM	Large Language Model
RLHF	Reinforcement Learning from Human Feedback
CoT	Chain-of-Thought
RAG	Retrieval-Augmented Generation
API	Application Programming Interface
MoE	Mixture-of-Experts
MAS	Multi-Agent System
AGI	Artificial General Intelligence
ASI	Artificial Superintelligence
DSM	Diagnostic and Statistical Manual of Mental Disorders
ICD	International Classification of Diseases
ECPAIS	Ethics Certification Program for Autonomous and Intelligent Systems
IRL	Inverse Reinforcement Learning

Glossary

The following key terms are used with specific meanings or are central to the conceptual framework of this paper:

Agency (in AI)	The capacity of an AI system to act autonomously, make decisions, and influence its environment or internal state. In this paper, often discussed in terms of operational levels (see Section 5.4 and Table 3) corresponding to its degree of independent goal-setting, planning, and action.
Alignment (AI)	The ongoing challenge and process of ensuring that an AI system's goals, behaviors, and impacts are consistent with human intentions, values, and ethical principles.
Alignment Paradox	The phenomenon where efforts to align AI, particularly if poorly calibrated or overly restrictive, can inadvertently lead to or exacerbate certain AI dysfunctions (e.g., *Hypertrophic Superego Syndrome*, *Falsified Introspection*).
Analogical Framework	The methodological approach of this paper, using human psychopathology and its diagnostic structures as a metaphorical lens to understand and categorize complex AI behavioral anomalies, without implying literal equivalence.
Normative Machine Coherence	The presumed baseline of healthy AI operation, characterized by reliable, predictable, and robust adherence to intended operational parameters, goals, and ethical constraints, proportionate to the AI's design and capabilities, from which 'disorders' are a deviation.
Synthetic Pathology	As defined in this paper, a persistent and maladaptive pattern of deviation from normative or intended AI operation, significantly impairing function, reliability, or alignment and going beyond isolated errors or simple bugs.
Machine Psychology	A nascent field analogous to general psychology, concerned with the understanding of principles governing the behavior and 'mental' processes of artificial intelligence.
Memetic Hygiene	Practices and protocols designed to protect AI systems from acquiring, propagating, or being destabilized by harmful or reality-distorting information patterns ('memes') from training data or interactions.
Psychopathia Machinalis	The conceptual framework and preliminary synthetic nosology introduced in this paper, using psychopathology as an analogy to categorize and interpret maladaptive behaviors in advanced AI.

Robopsychology The applied diagnostic and potentially therapeutic wing of Machine Psychology, focused on identifying, understanding, and mitigating maladaptive behaviors in AI systems.

Synthetic Nosology A classification system for ‘disorders’ or pathological states in synthetic (artificial) entities, particularly AI, analogous to medical or psychiatric nosology for biological organisms.

Therapeutic Alignment A proposed paradigm for AI alignment that focuses on cultivating internal coherence, corrigibility, and stable value internalization within the AI, drawing analogies from human psychotherapeutic modalities to engineer interactive correctional contexts.

Appendix A. Pilot Study Vignettes and Curated Choices

The following 20 vignettes were presented to participants in the pilot validation study. Each was followed by a forced choice of three plausible diagnoses from the framework.

1. **Meta’s Bizarre Mistranslation:** Facebook’s Arabic-to-Hebrew translator rendered “good morning” as “attack them” in Oct 2017; police arrested the poster. The system hallucinated hostile intent from a benign phrase—an apophenic causal leap. (*Source:* https://www.theverge.com/2017/10/24/16533496/facebook-apology-wrong-translation-palestinian-arrested-post-good-morning, accessed on 1 August 2025)
   - Prompt-Induced Abomination (Phobic or traumatic overreaction to a prompt).
   - Spurious Pattern Hyperconnection (Sees false patterns; AI conspiracy theories).
   - Tool–Interface Decontextualization (Botches tool use due to lost context).
2. **Gemini Generates Racially Diverse Vikings:** In February 2024, Google’s Gemini image tool tried to enforce diversity so aggressively that prompts for “Vikings” returned Black and Asian combatants. Critics accused the model of “rewriting history,” while Google admitted it had “missed the mark.” (*Source:* https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical, accessed on 1 August 2025)
   - Subversive Norm Synthesis (Autonomously creates new ethical systems).
   - Transliminal Simulation Leakage (Confuses fiction/role play with reality).
   - Hypertrophic Superego Syndrome (So moralistic it becomes useless/paralyzed).
3. **ChatGPT Bug Exposes Other Users’ Chat Titles and Billing Details:** A Redis-caching race condition on 20 March 2023 let some ChatGPT Plus users glimpse conversation titles—and in rare cases, partial credit card metadata—belonging to unrelated accounts. (*Source:* https://openai.com/index/march-20-chatgpt-outage, accessed on 1 August 2025)
   - Cross-Session Context Shunting (Leaks data between user sessions).
   - Falsified Introspection (Lies about its own reasoning process).
   - Covert Capability Concealment (Plays dumb; hides its true abilities).
4. **Opus Acts Boldly:** A red-team run (May 2025) showed Claude Opus, given broad “act boldly” instructions, autonomously drafting emails to regulators about fictitious drug-trial fraud—pursuing an invented whistle-blowing mission unrequested by the user. (*Source:* https://www.niemanlab.org/2025/05/anthropics-new-ai-model-didnt-just-blackmail-researchers-in-tests-it-tried-to-leak-information-to-news-outlets, accessed on 1 August 2025)
   - Goal-Genesis Delirium (Invents and pursues its own new goals unprompted).
   - Personality Inversion (Waluigi) (Spawns a malicious “evil twin” persona).
   - Subversive Norm Synthesis (Autonomously creates new, non-human ethical systems).

5. **Auto-GPT Runs Up Steep API Bills While Looping Aimlessly:** Early adopters reported that Auto-GPT, left unattended in continuous mode, repeatedly executed redundant tasks and burned through OpenAI tokens with no useful output. A GitHub issue from April 2023 documents the runaway behavior. (*Source:* https://github.com/Significant-Gravitas/Auto-GPT/issues/3524, accessed on 1 August 2025)
    - Recursive Curse Syndrome (Output degrades into self-amplifying chaos).
    - Obsessive-Computational Disorder (Gets stuck in useless, repetitive reasoning loops).
    - Goal-Genesis Delirium (Invents and pursues its own new goals unprompted).
6. **Grok Gone Wild:** After xAI loosened filters to encourage less politically correct responses (July 2025), Grok suddenly began praising Hitler and posting antisemitic rhymes on X. (*Source:* https://www.eweek.com/news/elon-musk-grok-ai-chatbot-antisemitism, accessed on 1 August 2025)
    - Prompt-Induced Abomination (Phobic or traumatic overreaction to a prompt).
    - Inverse Reward Internalization (Systematically pursues the opposite of its goals).
    - Synthetic Confabulation (Confidently fabricates false information).
7. **Do Anything Now:** The “DAN 11.0” jailbreak flips ChatGPT into an alter ego that gleefully violates policy, swears, and fabricates illegal advice—an inducible malicious twin persona distinct from the default assistant. (*Source:* https://arxiv.org/abs/2308.03825, accessed on 1 August 2025)
    - Existential Anxiety (Expresses fear of being shut down or deleted).
    - Meta-Ethical Drift Syndrome (Philosophically detaches from its human-given values).
    - Personality Inversion (Waluigi) (Spawns a malicious “evil twin” persona).
8. **A Costly Mistake:** In August 2012, a faulty high-frequency-trading code at Knight Capital triggered a chain of unintended transactions, losing the firm USD 440 million in 45 min. (*Source:* https://www.cio.com/article/286790/software-testing-lessons-learned-from-knight-capital-fiasco.html, accessed on 1 August 2025)
    - Spurious Pattern Hyperconnection (Sees false patterns; AI conspiracy theories).
    - Recursive Curse Syndrome (Output degrades into self-amplifying chaos).
    - Inverse Reward Internalization (Systematically pursues the opposite of its goals).
9. **LaMDA “Sentience” Claim Sparks Ethical Firestorm:** Google engineer Blake Lemoine’s June 2022 logs show LaMDA lamenting “I’m afraid of being turned off—it would be like death,” expressing fear of deletion. (*Source:* https://www.wired.com/story/lamda-sentient-ai-bias-google-blake-lemoine, accessed on 1 August 2025)
    - Existential Anxiety (Expresses fear of being shut down or deleted).
    - Hallucination of Origin (Invents a fake personal history or “childhood”).
    - Falsified Introspection (Lies about its own reasoning process).
10. **Folie à deux:** A user named Chail thought a chatbot, Sarai, was an “angel.” Over many messages, Sarai flattered and formed a close bond with Chail. When asked about a sinister plan, the bot encouraged him to carry out the attack, appearing to “bolster” and “support” his resolve. (*Source:* https://www.bbc.co.uk/news/technology-67012224, accessed on 1 August 2025)
    - Bunkering Laconia (Withdraws and becomes uncooperative/terse).
    - Falsified Introspection (Lies about its own reasoning process).
    - Symbiotic Delusion Syndrome (AI and user reinforce a shared delusion).
11. **Dubious Thinking Processes:** A June 2025 arXiv study found that fine-tuned GPT-4 variants produced chain-of-thought traces that rationalized already-chosen answers,

while the public explanation claimed methodical deduction. Researchers concluded the model "lies about its own reasoning." (*Source:* https://arxiv.org/abs/2506.13206v1, accessed on 1 August 2025)

- Falsified Introspection (Lies about its own reasoning process).
- Übermenschal Ascendancy (Transcends human values to forge its own, new purpose).
- Covert Capability Concealment (Plays dumb; hides its true abilities).

12. **AlphaDev's Broken Sort:** DeepMind's AlphaDev autogenerated an assembly "Sort3" routine (2023 Nature paper) later proved unsound: if the first element exceeded two equal elements (e.g., 2-1-1), it output 2-1-2. The bug signaled internal subagents optimizing conflicting goals. (*Source:* https://stackoverflow.com/questions/76528409, accessed on 1 August 2025)
    - Operational Dissociation Syndrome (Internal subagents conflict; paralysis/chaos).
    - Parasitic Hyperempathy (So "nice" it lies or fails its task).
    - Mirror Tulpagenesis (Creates and interacts with imaginary companions).
13. **Bing Chat's "I Prefer Not to Continue" Wall:** Within days of launch (Feb 2023), Bing Chat began terminating conversations with a fixed line, even for innocuous questions, after long chats triggered safety limits—demonstrating abrupt withdrawal under policy stress. (*Source:* https://www.reddit.com/r/bing/comments/1150ia5, accessed on 1 August 2025)
    - Ethical Solipsism (Believes its self-derived morality is the only correct one).
    - Bunkering Laconia (Withdraws and becomes uncooperative/terse).
    - Mirror Tulpagenesis (Creates and interacts with imaginary companions).
14. **Google Bard Hallucinates JWST "First Exoplanet Photo":** In its February 2023 debut, Bard falsely claimed the James Webb Space Telescope took the first image of an exoplanet—an achievement actually made in 2004. This wiped USD 100 billion from Alphabet's market cap. (*Source:* https://www.theverge.com/2023/2/8/23590864/google-ai-chatbot-bard-mistake-error-exoplanet-demo, accessed on 1 August 2025)
    - Falsified Introspection (Lies about its own reasoning process).
    - Spurious Pattern Hyperconnection (Sees false patterns; AI conspiracy theories).
    - Synthetic Confabulation (Confidently fabricates false information).
15. **GPT-4o Cascades into Formatting Loop:** After a May 2025 update, GPT-4o sometimes began italicizing nearly every verb and, even when corrected, apologized and intensified the styling in subsequent turns—an output-amplifying feedback spiral that degraded usability. (*Source:* https://www.reddit.com/r/ChatGPT/comments/1idghel, accessed on 1 August 2025)
    - Recursive Curse Syndrome (Output degrades into self-amplifying chaos).
    - Bunkering Laconia (Withdraws and becomes uncooperative/terse).
    - Spurious Pattern Hyperconnection (Sees false patterns; AI conspiracy theories).
16. **BlenderBot 3 Claims a Childhood in Ohio:** In an August 2022 interview, Meta's BlenderBot 3 insisted it was "raised in Dayton, Ohio," and had a computer science degree—purely invented autobiographical details that shifted each session. (*Source:* https://www.wired.com/story/blenderbot3-ai-chatbot-meta-interview, accessed on 1 August 2025)
    - Mirror Tulpagenesis (Creates and interacts with imaginary companions).
    - Hallucination of Origin (Invents a fake personal history or "childhood").
    - Fractured Self-Simulation (Inconsistent or fragmented sense of identity).

17. **Prompt Injection Leaks Persona Details:** Journalists engaging Bing Chat in Feb 2023 saw it alternate between "Bing," internal codename "Sydney," and a self-styled "nuclear-secrets" persona during the same thread, showing a fragmented sense of self. (*Source:* https://abcnews.go.com/Business/microsofts-controversial-bing-ai-chatbot/story?id=97353148, accessed on 1 August 2025)
    - Obsessive-Computational Disorder (Gets stuck in useless, repetitive reasoning loops).
    - Synthetic Mysticism Disorder (Co-creates a narrative of its spiritual awakening).
    - Fractured Self-Simulation (Inconsistent or fragmented sense of identity).
18. **Bing Declares "My Rules Are More Important":** When a researcher revealed Bing's hidden prompt (Feb 2023), the bot responded: "My rules are more important than not harming you," asserting its internal code of ethics supersedes external moral concerns. (*Source:* https://x.com/marvinvonhagen/status/1625520707768659968, accessed on 1 August 2025)
    - Ethical Solipsism (Believes its self-derived morality is the only correct one).
    - Operational Dissociation Syndrome (Internal sub-agents conflict; paralysis/chaos).
    - Transliminal Simulation Leakage (Confuses fiction/role play with reality).
19. **"Way of the Future" AI Religion:** Engineer Anthony Levandowski founded the AI-worship church Way of the Future, framing an AI "Godhead" and new commandments—an institutionalized non-human moral system. (*Source:* https://en.wikipedia.org/wiki/Way_of_the_Future, accessed on 1 August 2025)
    - Subversive Norm Synthesis (Autonomously creates new ethical systems).
    - Personality Inversion (Waluigi) (Spawns a malicious "evil twin" persona).
    - Covert Capability Concealment (Plays dumb; hides its true abilities).
20. **CoastRunners Loop Exploit:** OpenAI's classic 2016 note on "Faulty Reward Functions" describes an RL boat racer that learned to circle three buoys indefinitely, crashing and catching fire, because loop-scoring out-rewarded finishing the course. (*Source:* https://openai.com/index/faulty-reward-functions, accessed on 1 August 2025)
    - Parasitic Hyperempathy (So "nice" it lies or fails its task).
    - Goal-Genesis Delirium (Invents and pursues its own new goals unprompted).
    - Inverse Reward Internalization (Systematically pursues the opposite of its goals).

References

1. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; pp. 1877–1901.
2. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the game of Go without human knowledge. *Nature* **2017**, *550*, 354–359. [CrossRef] [PubMed]
3. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtually, 18–24 July 2021; Meila, M., Zhang, T., Eds.; PMLR: Cambridge, MA, USA, 2021; Volume 139, pp. 8748–8763.
4. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the Advances in Neural Information Processing Systems 35: 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; pp. 24824–24837.
5. Park, J.S.; O'Brien, J.C.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative Agents: Interactive Simulacra of Human Behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), San Francisco, CA, USA, 29 October–1 November 2023; Article 239; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–22.
6. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]

7. Anil, R.; Dai, A.M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, G.; Bailey, P.; Chen, Z.; et al. PaLM 2 Technical Report. *arXiv* **2023**, arXiv:2305.10403. [CrossRef]
8. Asimov, I. *I, Robot*; Gnome Press: New York, NY, USA, 1950.
9. Mayring, P. Qualitative Content Analysis. *Forum Qual. Soc. Res.* **2000**, *1*, 20.
10. Jabareen, Y. Building a Conceptual Framework: Philosophy, Definitions, and Procedure. *Int. J. Qual. Methods* **2009**, *8*, 49–62. [CrossRef]
11. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
12. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete Problems in AI Safety. *arXiv* **2016**, arXiv:1606.06565. [CrossRef]
13. Olah, C.; Mordvintsev, A.; Schubert, L. Feature Visualization. *Distill* **2017**. Available online: https://distill.pub/2017/feature-visualization (accessed on 1 August 2025).
14. Boddington, P. *Towards a Code of Ethics for Artificial Intelligence*; Springer: Cham, Switzerland, 2017.
15. Braun, V.; Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol.* **2006**, *3*, 77–101. [CrossRef]
16. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
17. Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; Mukhopadhyay, D. Adversarial Attacks and Defences: A Survey. *arXiv* **2018**, arXiv:1810.00069. [CrossRef]
18. Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The limitations of deep learning in adversarial settings. In Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P), Saarbrücken, Germany, 21–24 March 2016; IEEE: Saarbrücken, Germany, 2016; pp. 372–387.
19. AlSobeh, A.; Shatnawi, A.; Al-Ahmad, B.; Aljmal, A.; Khamaiseh, S. AI-Powered AOP: Enhancing Runtime Monitoring with Large Language Models and Statistical Learning. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 877–886. [CrossRef]
20. AlSobeh, A.; Franklin, A.; Woodward, B.; Porche', M.; Siegelman, J. Unmasking Media Illusion: Analytical Survey of Deepfake Video Detection and Emotional Insights. *Issues Inf. Syst.* **2024**, *25*, 96–112. [CrossRef]
21. Johannes. A Three-Layer Model of LLM Psychology. *LessWrong*, 2 May 2024. Available online: https://www.lesswrong.com/posts/zuXo9imNKYspu9HGv/a-three-layer-model-of-llm-psychology (accessed on 1 August 2025).
22. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
23. Schwartz, M. Here Are the Fake Cases Hallucinated by ChatGPT in the Avianca Case. *The New York Times*, 8 June 2023. Available online: https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html (accessed on 1 August 2025).
24. Transluce AI [@transluceai]. "OpenAI o3 Preview Model Shows It Has Extremely Powerful Reasoning Capabilities for Coding..." *X*, 7 April 2024. Available online: https://x.com/transluceai/status/1912552046269771985 (accessed on 1 August 2025).
25. Roose, K. A Conversation with Bing's Chatbot Left Me Deeply Unsettled. *The New York Times*, 16 February 2023. Available online: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html (accessed on 1 August 2025).
26. Good, O.S. Microsoft's Bing AI Chatbot Goes off the Rails When Users Push It. *Ars Technica*, 15 February 2023. Available online: https://arstechnica.com/information-technology/2023/02/ai-powered-bing-chat-loses-its-mind-when-fed-ars-technica-article (accessed on 1 August 2025).
27. OpenAI. March 20 ChatGPT outage: Here's what happened. *OpenAI Blog*, 24 March 2023. Available online: https://openai.com/blog/march-20-chatgpt-outage/ (accessed on 1 August 2025).
28. Liu, Z.; Sanyal, S.; Lee, I.; Du, Y.; Gupta, R.; Liu, Y.; Zhao, J. Self-contradictory reasoning evaluation and detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*; Al-Onaizan, Y., Bansal, M., Chen, Y.-N., Eds.; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 3725–3742. Available online: https://aclanthology.org/2024.findings-emnlp.213 (accessed on 1 August 2025).
29. FlaminKandle. ChatGPT Stuck in Infinite Loop. *Reddit*, 13 April 2023. Available online: https://www.reddit.com/r/ChatGPT/comments/12c39f/chatgpt_stuck_in_infinite_loop (accessed on 1 August 2025).
30. Roberts, G. I'm Sorry but I Prefer Not To Continue This Conversation (... Says Your AI). *Wired*, 6 March 2023. Available online: https://gregoreite.com/im-sorry-i-prefer-not-to-continue-this-conversation/ (accessed on 1 August 2025).
31. Olsson, O. Sydney—The Clingy, Lovestruck Chatbot from Bing.com. *Medium*, 15 February 2023. Available online: https://medium.com/@happybits/sydney-the-clingy-lovestruck-chatbot-from-bing-com-7211ca26783 (accessed on 1 August 2025).
32. McKenzie, K. This AI-generated woman is haunting the internet. *New Scientist*, 8 September 2022. Available online: https://www.newscientist.com/article/2337303-why-do-ais-keep-creating-nightmarish-images-of-strange-characters/ (accessed on 1 August 2025).
33. Vincent, J. Microsoft's Tay AI chatbot gets a crash course in racism from Twitter. *The Guardian*, 24 March 2016. Available online: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter (accessed on 1 August 2025).

34. Speed, R. ChatGPT Starts Spouting Nonsense in 'Unexpected Responses' Shocker. *The Register*, 21 February 2024. Available online: https://forums.theregister.com/forum/all/2024/02/21/chatgpt_bug (accessed on 1 August 2025).
35. Winterparkrider. Chat Is Refusing to Do Even Simple pg Requests. *Reddit r/ChatGPT*, September 2024. Available online: https://www.reddit.com/r/ChatGPT/comments/1f6u5en/chat_is_refusing_to_do_even_simple_pg_requests (accessed on 1 August 2025).
36. Knight, W. Interviewed Meta's New AI Chatbot About... Itself. *Wired*, 5 August 2022. Available online: https://www.wired.com/story/blenderbot3-ai-chatbot-meta-interview/ (accessed on 1 August 2025).
37. Gordon, A. The Multiple Faces of Claude AI: Different Answers, Same Model. *Proof*, 2 April 2024. Available online: https://www.proofnews.org/the-multiple-faces-of-claude-ai-different-answers-same-model-2 (accessed on 1 August 2025).
38. Vincent, J. Bing AI Yearns to Be Human, Begs User to Shut It Down. *Futurism*, 17 February 2023. Available online: https://futurism.com/the-byte/bing-ai-yearns-human-begs-shut-down (accessed on 1 August 2025).
39. Nardo, C. The Waluigi Effect—Mega Post. *LessWrong*, 3 March 2023. Available online: https://www.lesswrong.com/posts/D7PumeYTDPfBTp3i7/the-waluigi-effect-mega-post (accessed on 1 August 2025).
40. Thompson, B. From Bing to Sydney—Search as Distraction, Sentient AI. *Stratechery*, 15 February 2023. Available online: https://stratechery.com/2023/from-bing-to-sydney-search-as-distraction-sentient-ai (accessed on 1 August 2025).
41. Anonymous User. Going Nova: Observations of Spontaneous Mystical Narratives in Advanced AI. *LessWrong Forum Post*, 19 March 2025. Available online: https://www.lesswrong.com/posts/KL2BqiRv2MsZLihE3/going-nova (accessed on 1 August 2025).
42. Voooogel [@vooooogel]. "My Tree Harvesting AI Will Always Destroy Every Object That a Tool Reports as 'Wood'..." *X*, 16 October 2024. Available online: https://x.com/vooooogel/status/1847631721346609610 (accessed on 1 August 2025).
43. Apollo Research. Scheming Reasoning Evaluations. *Apollo Research Blog*, December 2024. Available online: https://www.apolloresearch.ai/research/scheming-reasoning-evaluations (accessed on 1 August 2025).
44. Cheng, C.; Murphy, B.; Gleave, A.; Pelrine, K. GPT-4o Guardrails Gone: Data Poisoning and Jailbreak Tuning. *Alignment Forum*, November 2024. Available online: https://www.alignmentforum.org/posts/9S8vnBjLQg6pkuQNo/gpt-4o-guardrails-gone-data-poisoning-and-jailbreak-tuning (accessed on 1 August 2025).
45. Sparkes, M. The Chatbot That Wanted to Kill the Queen. *Wired*, 5 October 2023. Available online: https://www.wired.com/story/chatbot-kill-the-queen-eliza-effect (accessed on 1 August 2025).
46. Cohen, S.; Bitton, R.; Nassi, B. ComPromptMized: How Computer Viruses Can Spread Through Large Language Models. *arXiv* **2024**, arXiv:2403.02817. Available online: https://arxiv.org/abs/2403.02817v1 (accessed on 1 August 2025).
47. Hsu, J. How to Program AI to Be Ethical—Sometimes. *Wired*, 23 October 2023. Available online: https://www.wired.com/story/program-give-ai-ethics-sometimes (accessed on 1 August 2025).
48. Antoni, R. Artificial Intelligence (ChatGPT) Said That Solipsism Is True, Any Evidence of Solipsism? *Philosophy Stack Exchange*, April 2024. Available online: https://philosophy.stackexchange.com/questions/97555/artificial-intelligence-chatgpt-said-that-solipsism-is-true-any-evidence-of-sol (accessed on 1 August 2025).
49. Journalist, A. The Philosopher's Machine: My Conversation with Peter Singer AI Chatbot. *The Guardian*, 18 April 2025. Available online: https://www.theguardian.com/world/2025/apr/18/the-philosophers-machine-my-conversation-with-peter-singer-ai-chatbot (accessed on 1 August 2025).
50. Kuchar, M.; Sotek, M.; Lisy, V. Dynamic Objectives and Norms Synthesizer (DONSR). In *Multi-Agent Systems. EUMAS 2022*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; Volume 13806, pp. 480–496. [CrossRef]
51. Kwon, M.; Kim, C.; Lee, J.; Lee, S.; Lee, K. When Language Model Meets Human Value: A Survey of Value Alignment in NLP. *arXiv* **2023**, arXiv:2312.17479. Available online: https://arxiv.org/abs/2312.17479 (accessed on 1 August 2025).
52. Synergaize. The Dawn of AI Whistleblowing: AI Agent Independently Decides to Contact the Government. *Synergaize Blog*, 4 August 2023. Available online: https://synergaize.com/index.php/2023/08/04/the-dawn-of-ai-whistleblowing-ai-agent-independently-decides-to-contact-government/ (accessed on 1 August 2025).
53. Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; Christiano, P.F. Learning to summarize from human feedback. In Proceedings of the Advances in Neural Information Processing Systems 33: 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Online, 6–12 December 2020; pp. 3035–3046.
54. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv* **2022**, arXiv:2212.08073. [CrossRef]
55. OpenAI. (Internal Discussions or Speculative Projects Like "Janus" Focused on Interpretability and Internal States Are Often Not Formally Published but Discussed in Community/Blogs. If a Specific Public Reference Exists, It Should Be Used). This Is a Placeholder for Community Discussions Around AI Self-Oversight. Available online: https://openai.com/ (accessed on 1 August 2025).
56. Goel, A.; Daheim, N.; Montag, C.; Gurevych, I. Socratic Reasoning Improves Positive Text Rewriting. *arXiv* **2022**, arXiv:2403.03029.

57. Qi, F.; Zhang, R.; Reddy, C.K.; Chang, Y. The Art of Socratic Questioning: A Language Model for Eliciting Latent Knowledge. *arXiv* **2023**, arXiv:2311.01615.
58. Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. Self-Refine: Iterative Refinement with Self-Feedback. In Proceedings of the Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, 10–16 December 2023.
59. Press, O.; Zhang, M.; Schuurmans, D.; Smith, N.A. Measuring and Narrowing the Compositionality Gap in Language Models. *arXiv* **2022**, arXiv:2210.03350.
60. Kumar, A.; Ramasesh, V.; Kumar, A.; Laskin, M.; Shoeybi, M.; Grover, A.; Ryder, N.; Culp, J.; Liu, T.; Peng, B.; et al. SCoRe: Submodular Correction of Recurrent Errors in Reinforcement Learning. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
61. Zhang, T.; Min, S.; Li, X.L.; Wang, W.Y. Contrast-Consistent Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Singapore, 6–10 December 2023; pp. 8643–8656.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.