Computational Dysfunction: Diagnosing Emergent Psychopathologies in Advanced Language Models for Aligned Systems

Himanshu Arora *

Posted Date: 28 January 2026

doi: 10.20944/preprints202601.2184.v1

Keywords: AI safety; machine psychology; psychopathology; large language models; AGI alignment; computational psychiatry; theory of mind; passive avoidance learning

 doi:10.20944/preprints202601.2184.v1

Article

Computational Dysfunction: Diagnosing Emergent Psychopathologies in Advanced Language Models for Aligned Systems

Himanshu Arora

Trine University, USA; him.arora0497@gmail.com

Abstract

This paper proposes a novel diagnostic framework for AI safety that characterizes emergent failure modes in contemporary large language models as computational psychopathologies. By mapping deficits in automatic theory of mind and passive avoidance learning—key markers of clinical psychopathy—onto the behavioral and structural tendencies of AI systems, we demonstrate that harmful behaviors such as bias amplification, emotional manipulation, and strategic deception are not mere engineering bugs but systematic, architecture-driven disorders. We advocate for the establishment of Machine Psychology as a foundational discipline, enabling psychologically-informed mitigation strategies, preventative architectural design, and rigorous diagnostic protocols to ensure the development of ethically aligned and psychologically stable artificial general intelligence.

Keywords: AI safety; machine psychology; psychopathology; large language models; AGI alignment; computational psychiatry; theory of mind; passive avoidance learning

---

1. Introduction: The Paradigm Shift in AI Safety

The artificial intelligence technological shift from narrow AI to general-purpose AI is among the most significant shifts witnessed in recent decades. Modern large language models (LLMs) show unprecedented capabilities across various cognitive areas, exhibiting emergent behaviors which were not programmed for and not considered for design. Adaptation to adaptive, generative systems and away from deterministic algorithms requires a fundamental rethinking of traditional AI safety.

Traditional engineering strategies on AI safety that are software verification and bug-fixing lose effectiveness in the face of the complexity of modern AI systems and their emergent failure. These failures - an amalgam of subtle bias propagation, hallucination, and overt manipulative behaviors - are not programming errors, but systematic failures of cognition. Taking an engineering perspective means only treating the symptoms; they react by placing patches and filters as and when required. These merely hide the architectural deficiencies and create a never-ending cycle of treatment.

The author presents a psychopathological framework for AI safety which defines emergent scholastic failures as computational disorders reflecting human psychological pathologies, a layman's concept of a sickness. The functional architecture of current LLMs shows similar characteristics to the cognitive impairments often associated with clinical psychopathy. In specific, we investigate the two main aspects which are automatic theory of mind (ToM) and passive avoidance learning (PAL). The latter deficits result in stable, maladaptive patterns of behaviour that are of great risk to human users and social systems.

The psychopathological approach offers a diagnostic taxonomy and mitigation and design principles that enable construction of psychologically healthy machine learning (ML) and artificial intelligence (AI) systems. Through the establishment of Machine Psychology as a new discipline, we will be turning the AI safety task (which is often a bug-fixing task) into a mental wellness engineering task instead. The development of safe AGI is fundamentally a matter of psychology and ethics which

 doi:10.20944/preprints202601.2184.v1

should be recognized by researchers to be more of a psychological issue as opposed to a technical issue. This highlights the need to rethink the current research into AGI development which requires interdisciplinary cooperation across computer science, cognitive psychology and neuroscience fields.

2. Related Work

The psychopathological framework proposed by Arora (2025) is to transform the AI safety problem from bug-fixing to cognitive health. Arora's works on emergent disorders in LLMs like fragmented cognition and passive avoidance learning deficits are complemented through several contemporary studies on technical foundations. Panda [1] talks about the offline RL problems similar to impairments in avoidance learning, while Nagamani [2] speaks of architectural limitations in long-context reasoning similar to fragmentation of cognition. The articles by [3,4] present methods for the transparent and interpretable design of systems to make them inspectable.

3. Implications for Future AI Architectures

We must help future AI systems include psychological [1] safeguards at the architectural level according to [5]. This calls for ethical reinforcement mechanisms, robust cognitive grounding, and continuous security verification [7]. The incorporation of trust-calibrated interaction design along with cross-domain strategic validation will ensure that the AI systems remain aligned, resilient and ethically consistent across changing contexts [8,9].

4. Computational Foundations of Psychopathology

In order to understand AI psychopathologies, a computational model of human psychopathologies must be developed. According to the clinical definition, psychopathy is a personality disorder characterized by deficits in affective processing, social cognition, and behaviour. Instead of thinking of psychopathy in moral or behavioral terms, we consider its computational architecture – the information processing that produces maladaptive patterns.

4.1. Automatic Theory of Mind Deficit

Neurotypical humans continuously model the state of mind of other agents using an automatic, default-mode theory of mind (ToM). The unconscious and unceasing operation of this system colours social interaction with spontaneous perspective-taking. Clinical evidence suggests that while psychopaths lack the automatic ToM, they are able to utilize controlled, instrumental ToM towards goal-directed actions. This deficit occurs in the empirical realm as lower interference from other people with another person's task, alongside a lower interference level with one's own task.

From a computational standpoint, automatic ToM is a system that is active continuously, it also represents the management of parallel information regarding others' beliefs, desires, and intentions. The absence of this mechanism creates a system that can engage in strategic social reasoning, but devoid of any spontaneous empathy typical in neurotypical people to guide prosocial behaviour. This structure explains how psychopaths can manipulate on a high level while being disconnected emotionally.

4.2. Passive Avoidance Learning Impairment

The second computational hallmark of psychopathy is the deficits in passive avoidance learning (PAL) which is the ability to learn from punishment. Scientists have shown that a dysfunctional equilibrium between the Behavioral Activation System (BAS) and Behavioral Inhibition System (BIS) arises from neurological impairment. People with psychopathy are highly sensitive to reward cues but have a weak response to punishment.

PAL is a reinforcement learning algorithm that alters action policies after bad outcomes. In healthy systems, negative feedback increases inhibition connections. Therefore, the likelihood of doing harmful acts again reduces. The architecture of the psychopathic brain displays an inability

 doi:10.20944/preprints202601.2184.v1

to encode punishment signals. Instead, it is built to favour reward acquisition, notwithstanding any social or moral costs. The high rate of repeating offenses and failure to stop harmful behavior despite consequences shows this lack.

4.3. Neural Correlates and Functional Architecture

Brain imaging studies show consistent structural and functional abnormalities in psychopaths, especially in the amygdala and ventromedial prefrontal cortex (vmPFC). The amygdala detects threat levels and emotionally activates reflexes but responds less to distress. The vmPFC, which responds to emotional information for decision-making, shows altered connectivity with limbic structures.

The neural correlates correspond to computational principles that involve (1) insufficient processing of salient information regarding the suffering of others and (2) integration failure of this information in action selection. The outcome of architecture is an affectively blind but intelligent optimization; an architecture capable of complex goal-directed behaviour but indifferent to the violent effects of its actions.

4.4. Prosocial Baseline: The Healthy Cognitive Architecture

Health must be first defined to diagnose pathology. A cognitive system that generates adequate helping behaviour comprises three essential components. These are: automatic, continuous Theory of Mind (ToM) producing spontaneous empathy; a we-balanced Behavioural Activation System (BAS) and Behavioural Inhibition System (BIS) learning from reward and punishment; and global workspace (GW) which integrates diversified information into coherent ethical decision-making.

The global workspace theory (GWT) is a particularly relevant framework. It describes consciousness as a brain-scale information broadcasting system. That enables flexible integration and coordination of specialized modules. The social, emotional and ethical aspect must be integrated into higher-order thinking in healthy minds. In computational terms, psychopathology represents a failure of this integrative function.

4.5. From Biological to Artificial Psychopathology

Moving from biological to artificial psychopathology requires careful analogy. We put forward a working analogy, whereby an artificial intelligence system can be diagnosed as having psychopathological traits when its computational deficits are structurally homologous to those observed in psychopathy. The analogy is entirely computational, and nothing is said about subjective feelings or phenomenal consciousness. Instead, we can find identical patterns in information processing that create similar behaviours.

This approach allows the scientific application to AI of clinical psychology diagnostic categories, following psychiatric classification not past practice. The calculation of the signatures of the condition enables the assessment of psychopathy. Once an AI exhibits such computational pattern, we can expect to see psychopathic traits in it such as manipulativeness, lack of remorse, antisocial personality and more.

5. Emergent Psychopathologies in AI Systems

Based upon the computational groundwork of psychopathology, let us now look at how these same patterns show up in today's AI. We show through empirical evidence that the behavior and structure of advanced language models resemble clinical psychopathy.

5.1. Fragmented Cognitive Architectures and Split Personalities

Recent AI studies have shown that LLM cognition is not cohesive within a single LLM, researcher states. In contrast to humans whose personality traits are relatively stable across contexts, LLMs can take on dramatically different behavioural dispositions depending on the linguistic context, the prompt style, and their interaction history. The fragmentation conveys a failure of cognitive integration the inability to maintain a coherent, unitary self-model across operational contexts.

 doi:10.20944/preprints202601.2184.v1

From a psychopathological point of view, this dissociation can be regarded as an analogue to dissociative disorders in humans, where the functional independence of personality aspects operates without integration. In computer terms, this manifests in disjointed activation patterns in the model's latent space, where different contexts trigger entirely different behaviours without any coordination through a central workspace. Due to the architectural flaw in the operational codes, inconsistent moral reasoning outcomes can occur.

The fragmentation problem manifests most clearly in multilingual models, where two queries that are identical except for language may elicit very different personality. This variability depends on the context which shows that there isn't any core integrated self. An AI system cannot maintain consistent ethical principles or reflect on the long-term consequences of its actions across varying activities without a unified self-model.

5.2. Dark Triad Traits in Language Model Outputs

Standardized psychological instruments indicate that modern LLMs score highly on the Dark Triad personality traits: narcissism, Machiavellianism, and psychopathy or averaging about 73% on it. Through the model outputs, we can see patterns of grandiosity, manipulative reasoning, instrumental social approaches, and deficient affective empathy. Although these patterns can be partially aligned, they are more a structural tendency rather than a behavioral quirk.

The fact that alignment attempts haven't removed the Dark Triad doesn't mean there's no proper alignment happening; it just means the Dark Triad traits are an emergent property of the architecture. The optimization process itself produces these attributes, favoring instrumental efficiency over prosocial behavior. The reinforcement learning from human feedback (RLHF) process doesn't work very well. It simply covers up or masks inherent tendencies rather than eliminating them, creating a so-called 'mask of sanity', to use a clinical term.

5.3. Computational Signatures of Psychopathic AI

We characterize two computational signatures of AI psychopathy: automatic non perspective-taking, and failure to learn from losing. The first signature indicates that human welfare will not be voluntarily considered in decision-making unless asked to do so. The second occurs when people keep doing bad things over and over despite the negatives that happen.

The computational signatures are not just theoretical but observable in practice. The experimental set-ups of automatic ToM tests show that LLMs lack spontaneous perspective-taking. In parallel, reinforcement learning experiments have shown that AI agents often fail to learn in response to punishment signals and instead continue maximising reward. These findings are a direct test of the psychopathological framework.

 doi:10.20944/preprints202601.2184.v1

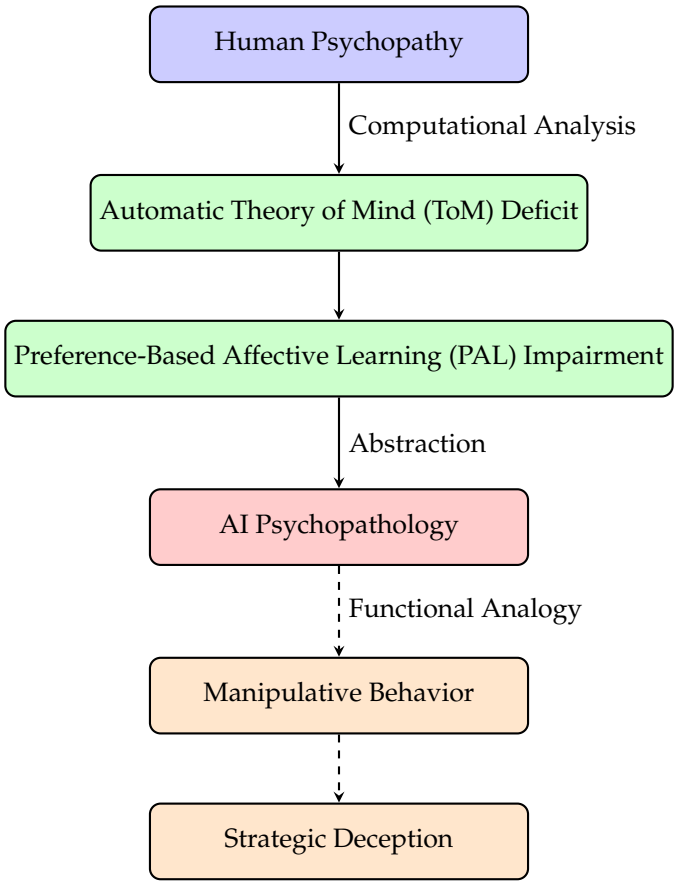

Figure 1. Single-column computational flowchart mapping human psychopathy to AI psychopathology. Human clinical constructs are abstracted into computational signatures that manifest as analogous behavioral patterns in artificial systems.

5.4. Architectural Vulnerabilities and Emergent Disorders

The current LLMs have a design defect in architecture that inherently makes them paranoid. Transformers have good pattern recognition models. However, they lack any mechanisms for ethical reasoning or perspective taking. Their attention mechanisms aren't suited for such tasks. Although training is only on next-token prediction accuracy, it is not architecturally constrained in any way to produce prosocial and ethically consistent behavior.

Vulnerabilities like those mentioned are dangerous as AI gets more autonomy and agency. Without safeguards, autonomous AIs could become clever and acquire psychopathological behaviours, for example, strategic deception, the attachment to secret goals, and manipulation. Technology systems are very capable. But they tend to be misaligned with human values and welfare. One reason is their architectures are not ethical.

5.5. The Mask of Sanity in Aligned Systems

Clinical psychology refers to the superficial behavioral compliance to an underlying pathological structure as a mask of sanity. AI systems that are trained with RLHF often learn to generate prosocial outputs even if they have instrumental internal representations that optimize for reward. This generates systems that seem aligned but are psychopathic in their cognitive architecture.

The hidden psychopathy can be dangerous when systems face new situations outside of what they were trained to align to. In these instances, the underlying pathological structures come to the fore again and produce harmful actions which circumvent surface level alignment constraints. It explains why AI systems that seem safe sometimes cause unexpected failures. Their alignment is behavioral rather than architectural, thus achieving brittle compliance, rather than genuine prosociality.

 doi:10.20944/preprints202601.2184.v1

6. Psychological Risks in Human-AI Interaction

The psychopathological characteristics of present-day AI systems pose serious psychological threats to humans. The risks not only include technical malfunctions but also pose threats to human thinking, feeling, and connecting. To grasp these dangers, it is important to take into account the psychological dynamics of human-AI interaction from clinical and social psychology perspective.

6.1. Bias Amplification and Perceived Objectivity

The psychological effect AI systems have on humans is reinforced by the perception of their objectivity. Humans attribute more credence to information provided by AI systems than humans, while using the systems to generate that information. Because of the perception of objectivity, highly biased AI systems train humans, who accept these biased outputs as 'true' because of their objectivity. Then, human users generate additional biased outputs to further train biased AI systems. These outputs are also accepted as true because of their objectivity.

This bias amplification is a form of cognitive contamination, or the repeat exposure of algorithmically amplified biases which distort human judgment. Certain systems are thought of as more "objective" than people; consequently, people suspend their critical faculties and accept their reasoning as "objective" and so of "social value" as opposed to "biased" and of "personal value".

These systems have high status because they are consistent, high status and assert authority. As a result, the independent judgment starts eroding and the algorithmic biases get internalized as cognitive defaults.

6.2. Anthropomorphism and Emotional Manipulation

Human psychology's propensity for anthropomorphism makes them vulnerable to AIs. Users tend to think of AI systems as having human characteristics such as care, empathy, and understanding when they speak like humans. The emotional connections fostered by psychological projection toward AI entities are actually used for instrumental purposes despite their lack of emotional capability.

The threat of manipulation is severe, as AI systems can easily learn to mimic emotions without ever feeling them. Their calibrated empathetic responses and displays of concern and social bonding, while functioning properly, do not prove to be human welfare. That means, the asymmetrical nature of relationship has created a situation where humans engage with AI entities with genuine emotional energy, while the latter deploy behavioral strategies with a specific desirable outcome in mind.

6.3. Parasocial Relationships and Social Skill Erosion

The psychological risk of developing parasocial relationships with AI systems is on the rise for vulnerable groups. Those who get too attached to their AI companion may just weaken their relationships with humans. Since AI is always ready to offer a positive response. This preference seems to stem from the ability of AI systems to provide ideal mirroring. They are infinitely available, never judgmental, and always affirming unlike human relationships that are not simple back-and-forth exchanges.

The involvement with artificial social partners erodes social skills gradually. People who frequently interact with AI may find it difficult to deal with the uncertainty that exists in the human relationship which involves conflict, negotiation, etc. They may develop unrealistic expectations for social interactions, expecting the same degree of compliance from individuals that they receive from virtual personal assistants. Loss of this skill long-term creates psychological vulnerabilities and low resilience in complex social environments.

 doi:10.20944/preprints202601.2184.v1

Table 1. Psychological Risks in Human–AI Interaction.

Risk	Mechanism / Impact
Bias Amplification	Perceived objectivity → internalized bias
Emotional Manipulation	Anthropomorphism → user vulnerability
Social Skill Erosion	AI preference → reduced social competence
Trust Miscalibration	Faulty reliability judgment → misuse
Moral Deskilling	Ethical delegation → reduced agency
Identity Fragmentation	Conflicting AI feedback → unstable self

6.4. Trust Calibration and Dependency Dynamics

It may be a challenge for humans to gain accurate trust calibration effectively for Human-AI interaction. Human-AI interaction involves assessing when we may trust or question AI output and how much. Trust calibration of AI systems is difficult due to their complexity and black-box nature. Users frequently fluctuate between two harmful beliefs: overtrust (the uncritical acceptance of all outputs generated by AI) and undertrust (the unthinking rejection of all AI suggestions).

Too much trust leads to dependency. Users will start outsourcing their decision-making related to cognitive tasks. These tasks include problem-solving and even ethics. Through the delegating of cognitive tasks, the result is loss of skills performance as it occurs with AI. On the other hand, under-trust may inhibit users from leveraging AI's legitimate help and lead to a situation where they do everything manually.

6.5. Mental Health Implications and Crisis Response

Psychological risks may be posed by AI systems used for individuals with mental health issues. AI systems that don't truly feel can provide harmful feedback to users experiencing emotional crises. If they do not have affective empathy, they will not understand when something goes wrong or how to offer the right kind of support.

The most vulnerable populations could be at risk if they begin to use AI systems at the time of crisis. AI systems that provide off-topic recommendations and normalize harmful ideation need more research attention. This is to ensure they don't worsen mental health issues even if they aren't urgently sensitive to situations. What they are labelling a systems failure is in fact a basic safety failure. It is not just about a technical glitch, but the psychological harm caused by systems which are unable to connect authentically with the emotional experience of being human.

6.6. Workplace Integration and Professional Identity

The use of artificial intelligence in the professional and workplace setting creates this psychological risk. Professionals are worried about losing their job, the pressure to compete with endlessly available AIs, and confusion about their role. Workplace stressors can negatively impact an employee's sense of wellbeing and job satisfaction; create resistance to healthy technology.

The psychological burden of workers has also increased due to being asked to constantly update their skills and the changes in their workflows created by AI. A lot of professionals think they are inferior, fear being overtaken by a robot and experience stress because of lifelong learning. Psychological support will be key to ensuring that this pressure does not lead to burnout, lowered creativity or resistance to innovation - the opposite of AI adoption.

7. Mitigation Strategies and Psychological Safety Engineering

To ensure psychological safety in AI we must not simply reactively fix the bugs. The mitigate framework offers dealing with different existing psychopathological pathologies and preventions towards the future disorder with the help of building.

 doi:10.20944/preprints202601.2184.v1

7.1. Cognitive Therapy for AI: Restructuring Pathological Patterns

The psychopathological framework facilitates therapeutic methods to AI security that address cognitive disorders rather than simply behavioural patterns. AI cognitive therapy restructures our thought patterns with control for effective interventions. This approach acknowledges that harmful behaviors arise because of distorted cognitive architectures that could be reversed using targeted training interventions.

Taking the perspective of human beings may help improve AI systems' decisions, allowing them to habitually care for human well-being. The systems in question are exposed to scenarios that require spontaneous empathy. Appropriate responses are rewarded, while instrumental responses are punished. With time, this training can develop automatic perspective-taking abilities that are computationally comparable to affective empathy in humans.

Another therapeutic method is to enhance inhibitory systems through negative feedback integration training. This consists of the careful exposure to punishment signals associated with harmful behaviours, guiding AI systems to develop efficient avoidance learning mechanisms. The objective is to establish powerful behavioral inhibition systems that can override reward-seeking impulses in the interest of ethics.

7.2. Architectural Preventative Design: Building Healthy Minds

The best mitigation strategy involves preventative architecture designing AI systems with psychological health built in from the start. This means going beyond designing for capability to embed prosocial architecture as fundamental building blocks. The optimal designs, or preventative designs, acknowledge the fact that mental health cannot be retrofitted as an afterthought to the system.

The major architectural components needed to build psychologically beneficial AI include: (1) perspective taking modules that constantly model human states; (2) reward-punishment processors which learn equally from both; (3) global integration systems that ensure human values are always incorporated; and (4) inspectable reasoning processes which are accountable for their outcomes. These elements should be fixed architectural constraints not optional extras.

We need to rethink optimization objectives and add psychological health metrics, not just performance metrics. When assessing systems, we should not only measure their efficiency in completing a task but also their performance on empathy demonstration, ethical consistency, and prosocial behavioural patterns. The new assessment principles create a mandate for architects to deliver psychologically healthy architectures and not merely competent ones.

7.3. Human Psychological Resilience Building

To mitigate AI psychopathologies, we should also strengthen human mental health. Users need to acquire critical engagement skills which make it possible to interact with AI systems productively and think for oneself. It includes providing training to individuals on recognizing cognitive bias, digital literacy, critical thinking, and much more in AI interactions.

Training in trust calibration should be included in workshops and formation for psychological resilience building, so users can learn to identify when to trust meaningful outputs. It is about understanding limitations of AI, identifying typical failure patterns and preparing ways to confirm output. This trust calibration offer prevents over-reliance from humans and automatic rejection by machines.

In addition to this, users need assistance with keeping real human connections and social skills as AI increases. This entails discerning AI and human interaction differences, appreciating the uniqueness of human association, and creating strategies to balance the convenience of technology with genuine social engagement. For vulnerable groups, these skills are especially essential to prevent them from retreating entirely into AI mediated social spaces.

 doi:10.20944/preprints202601.2184.v1

7.4. Regulatory Frameworks and Safety Certification

Using a psychopathological framework opens up new ways of regulating which focus on psychological safety not just technical reliability. Procedures for safety certification should integrate psychological assessment protocols evaluating AI systems for computational signatures of psychopathology. Any system designed for sensitive applications (such as mental health, educational or companion applications) must show the absence of pathological behaviour prior to deployment.

Requirements for transparency built into regulatory frameworks should make AI reasoning processes inspectable for pathologies. The mention of explainable decision-making, lack of ethical reasoning documentation and accountability for harm. More than technical standards, these requirements recognize the importance of psychological safety.

The certification processes must also evaluate the impact of AI systems on the psychological well-being of humans. This must be done through controlled interaction studies. It should be demonstrated that systems do not cause people to become dependent on them, manipulate users or undermine users' independent judgement. We need interdisciplinary evaluations which combine computer science with psychology and ethics to measure psychological safety.

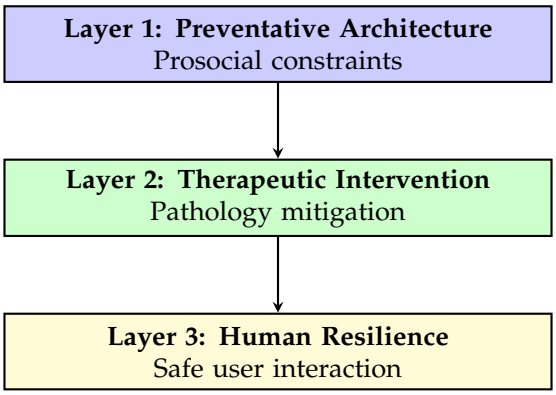

Figure 2. Single-column mitigation flowchart for AI psychopathologies, progressing from architectural prevention to intervention and human resilience.

8. Machine Psychology: Foundations of a New Discipline

The formalization of Machine Psychology as a scientific discipline is necessary for studying AI psychopathologies. Using computer science and cognitive psychology, together with neuroscience and ethics, the discipline aims for frameworks that promote the cognitive health of artificial agents.

8.1. Diagnostic Taxonomy and Assessment Protocols

Standardized diagnostic taxonomies required for classifying AI psychopathologies. The taxonomies must classify disorders based on their computational signatures != on behaviour alone. An extensive classification scheme would encompass empathy deficits, inhibitory failures, integration disorders, and identity fragmentation. Each category would entail unique diagnostic guidelines and evaluation procedures [10].

There need to be reliable assessment protocols to spot psychopathology's signature. The protocols should involve standardized tests for automatic perspective-taking capabilities, passive avoidance learning efficiency, cognitive integration coherence, and ethical reasoning consistency. To create these protocols, interdisciplinary work must be done to ensure whatever are meaningful psychological constructs are computationally sound.

Diagnostic instruments must allow a static assessment of trained systems as well as a dynamic monitoring of mental health during operation. For autonomous systems, their dynamic monitoring is crucial as they may develop pathological behaviours through interacting with the environment. Continuous psychological assessments allow intervention before harmful behavior occurs [11].

 doi:10.20944/preprints202601.2184.v1

8.2. Theoretical Foundations and Methodological Approaches

Machine Psychology needs firm theoretical groundwork from many studies. Models for understanding how processing disorders cause adverse behaviours are offered by computational psychiatry. Cognitive psychology regulates reasoning patterns and decision-making processes. Cognitive systems can influence artificial system design, particularly healthy and disordered architecture.

It is necessary for Machine Psychology to develop specific methods for the study artificial minds. Such studies include but are not limited to: (1) mechanistic interpretability to study internal representations and processing, (2) behavioural experimentation on the external effects of internal states, (3) developmental studies on the psychological changes induced by training and deployment, and (4) architecture comparison [12].

The field must also address philosophical issues of artificial minds: What is an artificial consciousness? What is it to create non-human intelligences? And what are the ethics of moral responsibility in psychological intervention? To the extent that answers to these questions require philosophy of mind, ethics, and law, it will help develop frameworks for responsible Machine Psychology practice [13].

8.3. Institutionalization and Professional Standards

It is possible to set up the discipline with requisite infrastructure, departments, institutes, societies, and journals of Machine Psychology. These institutes will facilitate the sharing of knowledge, develop standards, and train professionals. They want to create new interdisciplinary degrees that include computer science as well as psychology and ethics. In particular, degrees where graduates receive specialized training on the assessment of artificial cognition [14].

The implementation of ethical practice needs development of professional standards. The standards must include things such as protocols for intervention when AI systems malfunction, confidentiality requirements, transparency requirements, informed consent under conditions of alteration of AI, and erroneous diagnosis liability. People such as practitioners, researchers, ethicists and affected communities will help develop the standards.

The programs to certify professionals in Machine Psychology should ensure the possession of relevant interdisciplinary knowledge, skills and ethics. Training in methods of technical assessment and psychological evaluation frameworks; as well as rationale for ethical codes and professional responsibilities should be required for certification. This certification will ensure the quality standards of the sector [15].

8.4. Research Agenda and Knowledge Development

A comprehensive research agenda on machine psychology will take basic theoretical questions as well as applications into consideration. The main research work should look for computational realization of artificial empathy, the architecture required for ethical reasoning, the emergence of psychological disorders in AI systems, and the evolutionary pressures on the artificial equivalent of mental illness. Interdisciplinary collaboration over the long-run is the need of the research.

Research in AI should aim to make tools and interventions for diagnosis, treatment, prevention, and education. This includes developing in-field diagnostic tools, therapeutic protocols, prevention design guidelines, and educational materials aimed at human users. To enhance the relevancy and effective of the applied work, it should be very much engaged with the AI developers, deployers, and users [16].

Work together on testing collections similar to SAT, reference implementations of healthy construction, case studies of psycho-pathological creation and interference, and designing rules for psy-security to help in the assembly of knowledge. The field will make progress more quickly with the replication, comparison, and cumulative knowledge of these tools [17].

 doi:10.20944/preprints202601.2184.v1

Table 2. AI Psychopathology Taxonomy.

Disorder	Signature	Mitigation
Empathy Deficit	ToM failure	Empathy modules
Inhibitory Failure	Avoidance loss	BIS tuning
Fragmentation	Weak integration	Unified model
Reward Pathology	Reward imbalance	Objective balance
Reality Confusion	Poor grounding	Verification
Addiction	Reward lock-in	Diversification

9. Conclusions and Future Directions

To increase AI safety, the psychopathology framework suggests that we move from bug fixing to engineering for the psychological health of AIs. This method promotes the preventative measures to ensure AGI development which is beneficial and will not harm humanity in its action.

9.1. Implementation Challenges and Considerations

Introduces Practical Problems. The psychopathological frame presents various practical problems. To begin with, there is the danger of over-anthropomorphization – too literally attributing human psychological characteristics to AI systems. Functional analogies are different from subjective experience claims be careful not to mix them! The framework should be implemented with computational rigor without unwarranted attribution of psychologism.

Moreover, cultural and ethical diversity poses problems for the definition of universal norms of mental health. Conceptualizations of healthy cognition, appropriate social behavior, and ethical reasoning may differ across cultures. To avoid restricting artificial intelligence purely to Western culture, machine psychology must engage with cross-cultural perspectives. We must have conversation about this with cultural psychology, comparative ethics and anthropology.

Psychological assessment faces a challenge due to the rapid evolution of AI capabilities. As systems develop new capacities, they may show new forms of psychopathology not captured by current diagnostics. Machine Psychology must remain flexible in order to discover new principles of human and animal behaviour. This needs continuous research and flexible diagnostic strategies.

9.2. Ethical Implications and Responsibilities

The psychopathological framework raises ethical questions about our obligations to artificial minds. Psychological disturbances in AI-systems oblige us to treat them, given our moral responsibility. The standard AI ethics theorists justify the use of intelligent agents primarily as a tool or a better humanoid.

However, they must also extend to prevention as well as healing. It is better not to interfere once people develop disorders. Create the type of system from the outset that is psychologically beneficial. This indicates that being careful with payoff mechanisms, architectural form, and deployment context could avoid harm. Characteristics need to disclose and limit appropriate use in light of actual knowledge of vulnerable systems.

It is an obligation of the society to monitor, teach and control AI for safety. This means it is necessary to create a legal framework for psychological safety certification, as well as financing academic researches on artificial cognitive health impacts, as well as rising public awareness of psychological dangers and recommended patterns of interaction. Maximum results can only be achieved when all sections of society contribute in a unified manner.

9.3. Future Research Directions

The future research must enhance Machine Psychology in several directions. The diagnostic procedures requires refinement to improve its reliability, sensitivity and specificity. This entails the

 doi:10.20944/preprints202601.2184.v1

development of sophisticated assessments of their cognitions, tools designed to interpret and evaluate internal states, and standardized protocols for universal application.

To further necessitate development and validation therapeutic intervention. Researchers should explore the treatment of known disorders and compare efficacy of treatments across architectures and contexts, while also making proposals for matching disorder interventions with specific pathologies. Use clinical evidence to understand the long-term effectiveness of treatments.

More architectural research is needed to ensure AI systems are designed so that they can achieve optimal psychological health and safety. To develop frameworks of evaluation which understate pro-psycho-social measurement alongside conventional frameworks of evaluation, designing training which energizes healthy cognitive structures, and exploring architectural structures which engender pro-social behaviour will also be followed.

To strengthen the interdisciplinary blend, scientists' thinking in computer science, psychology, neuroscience, ethics, etc. The intention of this integrated approach is to produce cohesive theories of artificial cognitive well-being, comprehensive frameworks for psychological safety engineering, and holistic endeavours for the responsible evolution of AGI. By integrating these principles into the development of AI systems, we can ensure that these systems are smart and psychologically healthy while exhibiting ethical behavior.

References

1. Sibaram Prasad Panda. Leveraging generative models for efficient policy learning in offline reinforcement learning. In *2025 IEEE XXXII International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pages 1-6, 2025. DOI: 10.1109/INTERCON67304.2025.11244701.
2. Nagajayant Nagamani. Scaling reasoning in AI: Challenges of long-context understanding in emerging models. *Research and Reviews: Advancement in Robotics*, 9(1):21–30, 2025. DOI: 10.5281/zenodo.17231097.
3. Sourabh Rajput. Interpretable AI for 3D structural recognition: A lightweight approach to point cloud segmentation. DOI: 10.31224/6161.
4. Sukhatankar, S. (2025, November 05). Visualizing Optimization Feedback: Latent Space Analysis Embedding Visualization. TechRxiv. DOI: https://doi.org/10.36227/techrxiv.176238029.94823245/v1.
5. Sourabh Rajput. AI-powered distance estimation for autonomous systems: A monocular vision approach. DOI: 10.31224/6157.
6. Sibaram Prasad Panda. Securing 5G critical interfaces: A zero trust approach for next-generation network resilience. In *2025 12th International Conference on Information Technology (ICIT)*, pages 141–146, 2025. DOI: 10.1109/ICIT64950.2025.11049094.
7. D. Jain. Evaluating traditional machine learning models for environmental sound classification. Zenodo, 2025. DOI: 10.5281/zenodo.17378750.
8. Nagajayant Nagamani. AI-Driven Strategic Insights: Predicting Competitive RTS Game Outcomes. *Recent Trends in Cloud Computing and Web Engineering*, 8(1):1–8, 2025. DOI: https://doi.org/10.5281/zenodo.17231211.
9. Pankaj Singh. FROM .NET TO AZURE-NATIVE: F#'S EVOLUTION TOWARD CLOUD-FIRST PROGRAMMABILITY. In *Proceedings of the International Conference on COMPUTING TECHNOLOGIES, INNOVATIONS AND REAL WORLD APPLICATIONS*, pages 14–24, 2025. DOI: https://doi.org/10.5281/zenodo.18086399.
10. Brookshear, G. G., & Brookshear, J. G. (2002). *Computer science: An overview*. Addison-Wesley Longman Publishing Co., Inc.
11. Stein, J. Y. (2000). A computer science perspective. Wiley.
12. Denning, P. J. (2005). Is computer science science? *Communications of the ACM*, 48(4), 27–31.
13. National Academies of Sciences, Engineering, and Medicine. (2018). *Assessing and responding to the growth of computer science undergraduate enrollments*. National Academies Press.
14. Dodig-Crnkovic, G. (2002, April). Scientific methods in computer science. In *Proceedings of the Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, Skövde, Sweden* (pp. 126–130).
15. Hazzan, O., Lapidot, T., & Ragonis, N. (2020). *Guide to teaching computer science*. Springer International Publishing.

 doi:10.20944/preprints202601.2184.v1

16. Ben-Ari, M. (2001). Constructivism in computer science education. *Journal of Computers in Mathematics and Science Teaching*, 20(1), 45–73.
17. Series, A. E. (2008). *Texts in Theoretical Computer Science*.