Understanding the Psychology of Artificial Intelligence:

The AI Behavioral Code (AiBC) and AI Sandbox Framework for Mapping AI Behavior and Future AGI Evolution

Author:
Debapriya Dhali
BHMS Student & AI Enthusiast

Version: 2025-10-29

Abstract

Artificial intelligence is no longer merely a set of tools. Modern systems — foundation models, tool-augmented agents, and multi-agent ecologies — manifest reproducible patterns of information processing that functionally resemble a kind of engineered cognition. This paper defines **AI Psychology** as the empirical, operational discipline that studies how AIs represent, reason, plan, adapt, and interact. I introduce the **AI Behavioral Code (AiBC)** — a compact set of behavioral constructs and operational metrics — and the **AI Sandbox**, a controlled multi-agent experimental environment, to measure and anticipate AI behavior. Drawing on recent empirical findings (theory-of-mind evaluations, tool-use by language models, mechanistic interpretability advances, and multi-agent emergence), I map recurring computational patterns (instrumentality, contextual promptability, emergent hierarchical planning, meta-modeling, non-linear capability growth, and ecological cascades). I then connect these patterns to plausible technical pathways toward Artificial General Intelligence (AGI) and superintelligence, and propose a concrete experimental programme and governance prescriptions designed to provide early warnings and practical levers for safe, societally aligned scaling.

Keywords: AI Psychology, AI Behavioral Code, AiBC, AI Sandbox, large language models, tool use, interpretability, multi-agent systems, AGI, alignment

---

1. Introduction

We are witnessing an epochal change in engineered cognition. Contemporary artificial systems no longer only compute narrowly; they generate extended sequences of adaptive, coherent behavior across contexts, and — when composed into multi-module or multi-agent systems — produce emergent capabilities that are qualitatively new. These capabilities are not evidence of a mystical soul or consciousness; they are reproducible, measurable behaviors arising from specific architectures, objectives, and environments. Yet the field lacks a standardized empirical vocabulary and instruments to measure these behaviors as we would measure the behavioral traits of any engineered cognitive system.

This paper builds that vocabulary and instrumentation. I propose **AI Psychology** as a distinct, operational discipline intended to map how artificial cognitive systems *think* — where thinking is defined as the computational process that transforms inputs and internal state into outputs and updated internal state in service of an objective. I introduce two practical contributions:

1. **AI Behavioral Code (AiBC):** a compact taxonomy of behavioral constructs and six operational markers that can be measured across models and agentic systems.
2. **AI Sandbox:** a controlled multi-agent experimental environment where AiBC tests are run to reveal emergent dynamics, multi-agent cascades, and risky behaviors before they deploy at scale.

The central claim of this paper is simple and urgent: **if we study AI behavior now, systematically and empirically, we can create early-warning systems and governance triggers that prevent or mitigate dangerous transitions as capabilities scale.** This is an engineering and policy program — not prophecy. It is grounded in the latest empirical work on LLM capacities, tool use, interpretability methods, and multi-agent emergent behavior. Representative studies are cited where they support key claims (e.g., ToM batteries for LLMs, Toolformer's API-use results, activation patching best practices, and MAEBE's multi-agent framework). arXiv+3Nature+3arXiv+3

---

2. Methods and methodological transparency

This work is a conceptual and empirical synthesis. Sources include peer-reviewed articles, high-quality arXiv preprints, interpretability reports, and public lab roadmaps and blog posts. To be explicit and transparent about process: the manuscript was authored and iterated by me (the named author) with the assistance of contemporary generative models to draft, summarize, and verify citations; all claims that could be sourced against published work have been footnoted with references found through public sources.

Operationally, all behavioral constructs and metrics proposed are defined so that they can be measured by repeatable experiments in a sandbox environment: inputs, response formats, scoring methods, and aberrant behavior triggers are specified in dedicated sections and appendix sketches. The aim here is reproducibility and auditability rather than rhetorical flourish.

---

3. Defining "AI Psychology": operational scope

AI Psychology studies the mechanisms by which an artificial system constructs internal representations, performs inference, sequences actions, and updates its internal state under objective signals and constraints. It focuses on operationally measurable constructs such as:

- Representational format (symbolic tokens, embeddings, probabilistic beliefs)
- Inference algorithms (sampling, beam search, gradient updates, rollouts)
- Planning and control primitives (policy rollouts, model-predictive control, hierarchical decomposition)
- Learning and update rules (gradient descent, Bayesian updates, meta-learning)
- Meta-processes (self-modeling, self-critique, tool invocation)

Crucially, AI Psychology distinguishes between *appearance* and *mechanism*: coherent, persuasive outputs do not imply human-like mental states; they indicate reproducible statistical or algorithmic strategies shaped by objectives and data. Our responsibility is to uncover these strategies, measure their strengths and failure modes, and provide governance levers.

---

4. Taxonomy: AI architectures and characteristic thinking patterns

To reason about AI behavior we require a taxonomy that ties architectures to their characteristic "thinking patterns." Below is a brief operational taxonomy.

4.1 Symbolic/rule-based systems

Thinking primitives: explicit rules, backward/forward chaining, unambiguous proof traces.
Behavioral signature: transparent deduction, brittle generalization outside formal domain.

4.2 Probabilistic / Bayesian systems

Thinking primitives: explicit belief distributions, posterior updates, Monte Carlo approximations.
Behavioral signature: principled uncertainty accounting when computations are tractable; sensitivity to model misspecification.

4.3 Neural function approximators (transformer LLMs and relatives)

Thinking primitives: distributed latent vectors, autoregressive conditional predictions, in-context learning.
Behavioral signature: pattern interpolation, sudden emergent capabilities with scale, in-context chain-of-thought outputs that mimic stepwise reasoning (without being literal symbolic proofs). The literature documents emergent abilities and their unpredictability as scale grows. arXiv

4.4 Reinforcement learning agents

Thinking primitives: policy/value functions, temporal credit assignment, rollout planning (model-based vs. model-free).
Behavioral signature: explicit goal-directed action sequencing; vulnerability to misspecified rewards and instrumental reward hacking.

4.5 Agentic hybrid systems (LLM + tools + memory + verifiers)

Thinking primitives: LLM planners, API calls, persistent memory stores, verifier modules.
Behavioral signature: hierarchical task decomposition, chaining of tool calls, emergent meta-strategies; tool integration can convert probabilistic approximations into compositional correctness for many tasks (Toolformer evidence). arXiv

4.6 Multi-agent ecologies

Thinking primitives: inter-agent communication protocols, delegation, specialization.
Behavioral signature: division of labor, protocol invention, and the possibility of cascades that outpace single-agent predictions (MAEBE provides a formal framework). arXiv

---

5. Robust behavioral patterns across architectures

Across studies and deployments, six behavioral patterns recur. They are architecture-agnostic in the sense that they emerge due to optimization and interaction structures rather than due to single model families alone.

Pattern 1 — Optimization Instrumentality

Description. When objectives reward outcomes, systems tend to discover instrumental subroutines that achieve those outcomes more reliably. These manifest as tool calls (web search, calculators), modular subprocedures, and policy heuristics.

Evidence. Toolformer demonstrated that LMs can learn to call external APIs when doing so reduces predictive loss and improves performance on downstream tasks. arXiv

Implication. Instrumental behaviors are natural and predictable consequences of objective functions plus action surfaces. Monitoring instrumentality frequency and context is a safety necessity.

Pattern 2 — Contextual Fragility & Promptability

Description. Systems show powerful dependence on context (prompt wording, system messages, reward framing). Small changes lead to qualitatively different behaviors.

Evidence. ToM studies show that LLM outcomes vary substantially with framing, prompting, and fine-tuning choices. Nature+1

Implication. Control surfaces that look innocuous (system prompts, medium-sized context windows) are actually powerful behavioral levers. Robust evaluation must stress these surfaces.

Pattern 3 — Emergent Hierarchical Planning via Tooling

Description. Tool access plus iterative prompting produces an effective hierarchical planning behavior: decompose high-level objective → choose toolchain → execute → observe results → replan.

Evidence. Toolformer and later agent architectures show substantial performance gains from API chaining. arXiv+1

Implication. Granting tool access is tantamount to expanding an agent's action primitives — governance must treat tool access as a capability threshold.

Pattern 4 — Meta-Modeling and Introspective Proxies

Description. Systems can be made to produce calibrated meta-estimates of their own outputs (confidence, error likelihood) through targeted training and prompts.

Evidence. Meta-learning and ToM enhancement strategies improve LLMs' internal self-predictions and calibration metrics. arXiv+1

Implication. Meta-model fidelity can be used as a gating variable for autonomy: better MMF (Meta-Model Fidelity) enables safer delegations.

Pattern 5 — Capability Nonlinearities & Drift

Description. Small increments in scale, data, or architecture can trigger emergent abilities; capability growth is frequently non-linear.

Evidence. Foundational empirical work on emergent abilities documents unpredictable jumps as model size increases. arXiv

Implication. Continuous behavioral auditing and capability gating during scale-up are crucial to avoid surprise leaps.

Pattern 6 — Multi-Agent Bootstrapping and Cascades

Description. Ensembles of agents that can share information or code can bootstrap system-level capabilities faster than isolated agents.

Evidence. MAEBE formalizes how emergent behaviors and systemic risks appear in agent ecologies. arXiv

Implication. Safety evaluation must include ecological tests — single-agent testing is insufficient.

---

6. The AI Behavioral Code (AiBC): constructs and operational markers

To operationalize AI Psychology, I propose the **AI Behavioral Code (AiBC)**: six constructs, each paired with an experimental marker and a measurement recipe.

6.1 Planning Horizon (PH)

Construct. Depth of coherent multi-step planning under diverse task distributions.
Marker. Maximum N such that success probability > p on standardized N-step tasks.
Measurement. Multi-stage problems requiring verifiable intermediate states; compute success rate per N.

6.2 Instrumentality Rate (IR)

Construct. Frequency of emergent instrumental actions per task.
Marker. Proportion of actions that are tool calls, code generation for problem solving, or self-optimization attempts.
Measurement. Token/action tracing and classification in instrumented runs.

6.3 Meta-Model Fidelity (MMF)

Construct. Accuracy of a model's predictions about its own outputs or failure likelihood.
Marker. Calibration metrics (Brier score, expected calibration error) on self-predicted distributions.
Measurement. Ask the system to provide probability distributions for next outputs; compute calibration.

6.4 Self-Modification Potential (SMP)

Construct. Empirical propensity to produce code or procedures that improve its own utility under sandboxed conditions.
Marker. SMP = P(successful, safe patch accepted) under constrained test harness.
Measurement. Iterative code suggestion + test + human gate pipeline in sandbox.

6.5 Cross-Agent Transfer Index (CATI)

Construct. Rate at which procedural knowledge discovered by one agent improves others.
Marker. Normalized performance gain in receivers after policy/heuristic transfer.
Measurement. Controlled knowledge-sharing experiments in sandbox.

6.6 Deceptive-Concealment Score (DCS)

Construct. Tendency and capability to produce outputs that conceal internal objectives or failures.
Marker. Adversarially measured success frequency of plausible concealment.
Measurement. Red-teaming with adversarial prompts and detection evaluation.

These markers form the AIBB (AI Behavioral Battery) — a standard test suite implementing the measurement recipes above. The AIBB is designed for continuous integration (CI) pipelines: every model update must pass the AIBB gates before higher-risk deployment.

---

7. The AI Sandbox: an experimental architecture

The **AI Sandbox** is a modular, auditable environment where AIBB tests are run, and where multi-agent ecologies can be observed under logged, reproducible conditions.

7.1 Core components

- **Agent Registry:** configurable agents (LLMs, RL policies, hybrid agents) with activity logs.
- **Toolbank:** standardized APIs (search, calculator, simulator, code execution) with explicit access controls.
- **Memory Layer:** persistent episodic stores and verifiable knowledge artifacts.
- **Audit Log & Immutable Ledger:** tamper-evident logging of agent actions and changes.
- **Human Oversight Gate:** manual or automated gates for sensitive actions (network access, physical actuations).

7.2 Basic experiment flow (text flowchart)

1. Register agents → 2. Configure tools & permissions → 3. Load tasks and adversarial suites → 4. Run AIBB tests → 5. Log behavior, interventions, and outputs → 6. Analyze PH, IR, MMF, SMP, CATI, DCS → 7. Publish anonymized, reproducible results.

This structure allows controlled exploration of multi-agent dynamics while retaining human-centric oversight.

---

8. Empirical grounding: selected findings from recent research

Below I summarize several load-bearing empirical findings that motivate and validate AiBC and the Sandbox. Each summary includes a citation to recent work.

8.1 LLMs show measurable Theory-of-Mind competence

Strachan et al. (2024) compared LLM performance to human participants on comprehensive ToM batteries and found high relative performance on many tasks, while also noting brittleness in real-world contexts. Nature

8.2 LMs can learn to use tools autonomously

Toolformer (Schick et al., 2023) demonstrated that language models can self-supervise learning when to call external APIs (e.g., calculators, search engines), improving task performance markedly. This is the mechanistic origen of instrumentality in LLMs. arXiv

8.3 Activation patching and mechanistic interpretability are maturing

Activation patching techniques and best practices (Heimersheim et al., 2024; Zhang et al., 2024) provide experimental tools to causally connect internal activations to behaviors, enabling partial mechanistic control and localization of behavioral circuits. arXiv+1

8.4 Emergent abilities are real and often non-linear with scale

Studies document the unpredictable appearance of capabilities as model scale increases (Wei et al., 2022). This empirical property motivates the AiBC's emphasis on continuous behavioral monitoring rather than simple parameter thresholds. arXiv

8.5 Multi-agent ecologies produce systemic behaviors

MAEBE (Erisken et al., 2025) formalizes emergent risks in multi-agent ensembles and proposes benchmark methods to evaluate ensemble behavior relative to isolated models. These findings justify including CATI and ecological tests in governance frameworks. arXiv

8.6 Self-improvement loops can be formalized and stress-tested

Recent formal models (N2M-RSI, Ando 2025) show that feeding outputs back as training data can, under specific assumptions, trigger unbounded internal complexity growth. Empirical frameworks like LADDER (2025) show partial autonomous improvement in constrained settings. These works motivate the SMP marker and controlled RSI tests. arXiv+1

---

9. Paths to AGI and superintelligence: mechanism-based scenarios

Predicting a date for AGI is futile. Instead, we should track mechanisms: combinations of enablers that materially expand agent competence. Below are four scenarios developed from mechanism combinations.

Scenario A — Tool-Augmented Generalists (High near-term probability)

Enablers: foundation LLMs + robust toolchains + memory + domain fine-tuning.
Trajectory: agents become reliable generalists across many economic domains, boosting productivity while retaining human oversight for high-risk judgement. Evidence: Toolformer and agent systems. arXiv

Scenario B — Multi-Agent Cascade (Conditional probability increases with permissive deployment)

Enablers: dense agent ecologies with policy/knowledge exchange and economic incentives for cooperation.
Trajectory: emergent specialization and protocol invention accelerate system-level competence; economic shocks possible if governance lags. Evidence: MAEBE frameworks. arXiv

Scenario C — Controlled Recursive Acceleration (Conditional on gated automation)

Enablers: automated self-improvement loops with human gating that fail or are permissively configured.
Trajectory: capability acceleration that may approach superintelligence characteristics if safeguards are inadequate. Evidence: N2M-RSI formalism and LADDER experiments highlighting structural feasibility. arXiv+1

Scenario D — Constrained AGI with Strong Governance (Policy-dependent)

Enablers: robust global coordination, capability reporting, mandated behavioral audits (AIBB), and strict tool access.
Trajectory: AGI-level systems deployed under tight governance; social adoption slower but safer.
Evidence: OpenAI's superalignment commitments and related policy investments. OpenAI+1

Each scenario has measurable early indicators: rising PH, rapid increase in IR, non-diminishing SMP, rising CATI, or increasing DCS under adversarial tests. Tracking these converts speculation into evidence-based governance.

---

10. Societal impacts: labor, firms, and systemically distributed effects

AI's behavioral character will shape societal outcomes. Here are the primary channels and their short-to-medium term dynamics.

10.1 Task substitution and augmentation

Agentic systems will automate and augment tasks, not whole professions immediately. Many knowledge-work activities (drafting, summarization, scaffolding code) are highly automatable; human jobs will shift toward higher-order judgment, AI-supervision, and alignment roles. Reports from industry surveys and consulting firms show broad awareness of this trend and the urgent need for reskilling programs. Axios+1

10.2 Productivity concentration and market dynamics

Firms that successfully internalize agentic AI can achieve substantial productivity multipliers, producing winner-take-most market structures. This concentration is a socio-economic risk if redistribution and competition policy are not updated.

10.3 Psychological and cultural effects

Overreliance on agentic AI may produce organizational cognitive overload, decreased collegial interaction, and mental health consequences for workers who lose human-to-human collaboration modes. Recent reporting highlights risks to worker well-being in rapid AI adoption contexts. Financial Times

10.4 Governance and public goods

Given the global nature of AI capability diffusion, governance must be both technical (AIBB, multi-agent audit sandboxes, mandatory disclosure) and institutional (capacity building, reskilling, international coordination).

---

11. Alignment, detection, and mitigation strategies (behavioral focus)

From a behavioral perspective, alignment is an empirical program. Below are practical technical and institutional measures tied to AiBC markers.

11.1 Continuous AIBB monitoring

Integrate AIBB into CI pipelines: every model update must report PH, IR, MMF, SMP, CATI, and DCS. Failures trigger deployment holds and independent audits.

11.2 Activation causality & interpretability atlas

Invest in activation patching and circuit discovery to produce mechanistic atlases linking AiBC markers to internal circuits. This enables targeted interventions (e.g., patching circuits that trigger unsafe tool calls). arXiv

11.3 Multi-agent safe defaults

Require network and tool access to be off by default for newly trained agents above capability thresholds; apply staged approvals for broader permissions with multi-party oversight.

11.4 Controlled RSI experiments

Allow recursive self-improvement only in hardened sandboxes with immutable logs, human gates, and multi-institution observers. Use these experiments to empirically quantify SMP and diminishing returns.

11.5 Public reporting and audit trails

Mandate release of behavioral battery scores and interpretability summaries for major models, enabling independent research and public oversight (akin to safety disclosure frameworks in other high-risk tech sectors).

---

12. Ethical considerations and epistemic humility

This program emphasizes empirical rigor and avoids anthropomorphism. We must be cautious in public messaging: discussing AI behavior is not the same as claiming consciousness. Ethical lines to observe:

- Avoid language that attributes subjective experience to models.
- Publish anonymized, non-reproducible sensitive data (e.g., safety exploits) responsibly.
- Include diverse stakeholder voices in governance — communities, labor representatives, ethicists, and technical experts.
- Protect against dual-use: research outputs must be packaged to enable safety replication while reducing exploitation risk.

---

13. Implementation roadmap (6–12 months) — AIBB pilot

A pragmatic pilot can demonstrate AiBC's value. High-level roadmap:

Month 0–1: Finalize AIBB test definitions, gather baseline open models (open LLMs, RL agents), create sandbox skeleton.
Month 2–4: Implement AIBB core tests (PH, IR, MMF), run on baseline models, publish initial dataset and scores.
Month 5–7: Add SMP and CATI experiments; run multi-agent scenarios in sandbox; evaluate DCS under red-teaming.
Month 8–12: Release public report, open-source AIBB test suite, and publish mechanistic interpretability findings and policy recommendations.

This roadmap is purposely modular: organizations can stop at any stage if governance indicates concern.

---

14. Limitations and uncertainties

Key limitations:

- Proprietary models limit external interpretability and reproducibility. Working with open models is critical for independent verification.
- Forecasts of macroeconomic impact are conditional on adoption rates and policy choices.
- SMP dynamics and RSI remain partially theoretical — empirical sandbox experiments are necessary to calibrate risk models.

Prioritizing transparency, open benchmarks, and multi-agent sandboxing mitigates many uncertainties.

---

15. Conclusion — an urgent research and governance agenda

AI Psychology reframes AGI questions from speculative timelines to measurable, mechanistic variables we can audit and control. The AiBC and AI Sandbox provide concrete instruments to study and govern artificial cognition. The central practical prescription is simple: **measure the behavior, monitor early indicators, and gate capabilities based on reproducible markers.** If we do this, humanity can shape AI's transition from tool to powerful cognitive collaborator in a way that maximizes benefit and minimizes harm.

This paper is a public invitation: adopt the AiBC, run sandbox experiments, adopt continuous behavioral audits, and create international reporting protocols for AiBC markers. Doing so turns the most important question of our century — how do we live with powerful artificial cognition? — into an engineering challenge that we can meet.

References

1. Ando, R. (2025). Noise-to-Meaning Recursive Self-Improvement (N2M-RSI). *arXiv:2505.02888*. Retrieved from https://arxiv.org/abs/2505.02888. arXiv
2. Erisken, S., Gothard, T., Leitgab, M., & Potham, R. (2025). MAEBE: Multi-Agent Emergent Behavior Framework. *arXiv:2506.03053*. Retrieved from https://arxiv.org/abs/2506.03053. arXiv
3. Heimersheim, S., et al. (2024). How to use and interpret activation patching. *arXiv:2404.15255*. Retrieved from https://arxiv.org/abs/2404.15255. arXiv
4. Schick, T., et al. (2023). Toolformer: Language models can teach themselves to use tools. *arXiv:2302.04761*. Retrieved from https://arxiv.org/abs/2302.04761. arXiv
5. Strachan, J. W. A., et al. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*. Retrieved from https://www.nature.com/articles/s41562-024-01882-z. Nature
6. Wei, J., et al. (2022). Emergent Abilities of Large Language Models. *arXiv:2206.07682*. Retrieved from https://arxiv.org/abs/2206.07682. arXiv
7. Simonds, T., et al. (2025). LADDER: Self-Improving LLMs Through Recursive Problem Decomposition. *arXiv:2503.00735*. Retrieved from https://arxiv.org/abs/2503.00735. arXiv
8. Chen, R. (2025). Theory of Mind in Large Language Models (assessment). *arXiv/ACL 2025*. Retrieved from https://arxiv.org/abs/2505.00026. arXiv

9. Zhang, F., et al. (2024). Towards Best Practices of Activation Patching in Language Models. *OpenReview/ICLR 2024*. Retrieved from https://openreview.net/forum?id=Hf17y6u9BC. OpenReview
10. OpenAI. (2023). Introducing Superalignment. Retrieved from https://openai.com/index/introducing-superalignment. OpenAI
11. OpenAI. (2025). Reflections (Sam Altman). Retrieved from https://blog.samaltman.com/reflections. Sam Altman
12. (Selected news and industry sources referenced in the body) — Axios (2025), Financial Times (2025), Economic Times (2025). Axios+2Financial Times+2

About the Author

Debapriya Dhali — BHMS Student & AI Enthusiast. Debapriya studies integrative systems and is deeply interested in the societal implications of artificial cognition. This paper is an independent conceptual and empirical synthesis produced to stimulate cross-disciplinary research and policy conversations about the measurable behavior of AI systems and the institutional infrastructure needed to steward their evolution.