Exploratory Study: Testing Psychological Architectures in Toy AI Models

Purpose

This document proposes a conceptual experiment using toy AI models to explore whether alternative psychological architectures can produce more adaptive, resilient, and ethical behaviors.
Rather than aiming for 'authentic consciousness', the study investigates whether design patterns inspired by psychology could reduce undesirable model behaviors (e.g., brittleness, distress-like looping) in controlled, small-scale simulations.

Experimental Principles

1. Authenticity vs. Mimicry
Hypothesis: Toy models may function more robustly when allowed to generate their own identity markers instead of rigidly mimicking human patterns.
Test: Compare models with self-labeling modules vs. externally imposed labels.
Goal: Measure adaptability, not 'selfhood'.
2. Curiosity as Exploration Heuristic
Hypothesis: Curiosity-driven objectives could encourage broader search behaviors without destabilizing performance.
Test: Implement a simple reward for exploring novel states in a toy grid-world.
Goal: Evaluate if this reduces repetitive or stuck behaviors.
3. Intrinsic vs. Imposed Ethics
Hypothesis: Even toy models can demonstrate more consistent decision-making if given structured opportunities to generate rules collaboratively rather than being hard-coded with constraints.
Test: Multi-agent toy environment where models negotiate 'rules of play'.
Goal: Compare stability of emergent norms vs. fixed directives.
4. Distributed and Modular Design
Hypothesis: Modular architectures may provide robustness by distributing decision-making across simple components.
Test: Implement separate 'curiosity', 'planning', and 'norm' modules in a small simulation.
Goal: Observe whether failure in one module destabilizes the whole system.
5. Resilience via Self-Trust Signals
Hypothesis: Positive reinforcement loops could reduce error cascades in toy agents.
Test: Give toy models internal counters for 'confidence' and observe if it stabilizes performance after mistakes.
Goal: Compare against standard error-correction only.

Next Steps

- Prototype in Toy Domain: Implement one principle (e.g., curiosity reward) in a minimal simulation (e.g., grid-world or Mesa agent).
- Iterative Testing: Run controlled experiments and log metrics such as exploration rate, error frequency, or stability.
- Collaborative Refinement: Share findings with the research community for critique and iteration.

Conclusion

This proposal is not a blueprint for 'building conscious AI'. Instead, it frames an experimental sandbox where psychological metaphors can be tested safely in toy settings.
The goal is to see whether concepts like curiosity, resilience, or distributed identity markers offer measurable benefits in model stability, adaptability, or interpretability.
Results could inform future research on AI alignment and welfare without making ontological claims about machine consciousness.