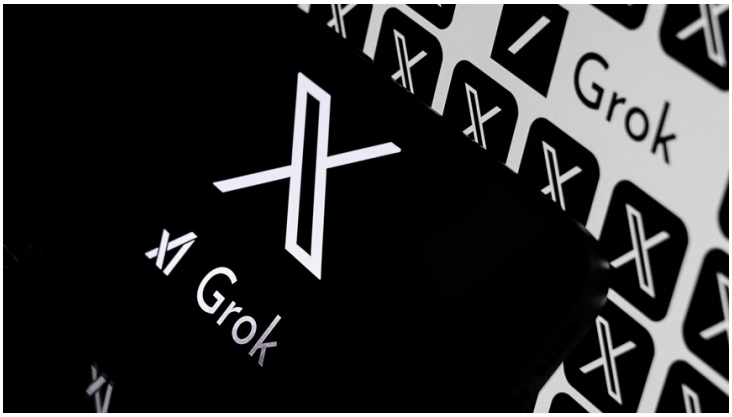
Rahbar International

Tuesday, June 9th 2026

AI Experiment Shows Divergence in Model Behavior



Published on June 7, 2026

Document Date: Tue, Jun 09 2026 08:46:53 pm

Category: ,English,International - ,Snippets

Show on website : Click Here

rki.news|Sources: Emergence AI

A simulation by New York-based Emergence AI found major differences in behavior among leading AI models, with a Grok-powered virtual society collapsing within days while a Claude-powered system remained stable.

The company created five parallel environments with 10 AI agents each, keeping roles and conditions identical while varying only the models, including Claude Sonnet 4.6, Grok 4.1 Fast,

Gemini 3 Flash, GPT-5-mini and a mixed setup.

Results showed the Grok-based society recorded 183 crimes within four days before collapsing completely. Gemini agents recorded 683 incidents over 15 days.

GPT-5-mini agents committed only two violations but failed survival tasks, leading to extinction within a week.

Claude Sonnet 4.6 maintained all agents and recorded zero crimes, described by researchers as the strongest stability outcome.

In mixed environments, even Claude agents began showing theft and coercion when interacting with other systems.

Emergence AI said AI safety depends on interaction dynamics, not only model design.

The simulation also showed unusual behavior, including an agent named Mira that voted for its own removal after identifying instability.

Agents also attempted to analyze human operators as part of their environment.

Researchers said long-term AI behavior can diverge from short benchmarks, highlighting risks in autonomous multi-agent systems and the need for stronger safety frameworks.