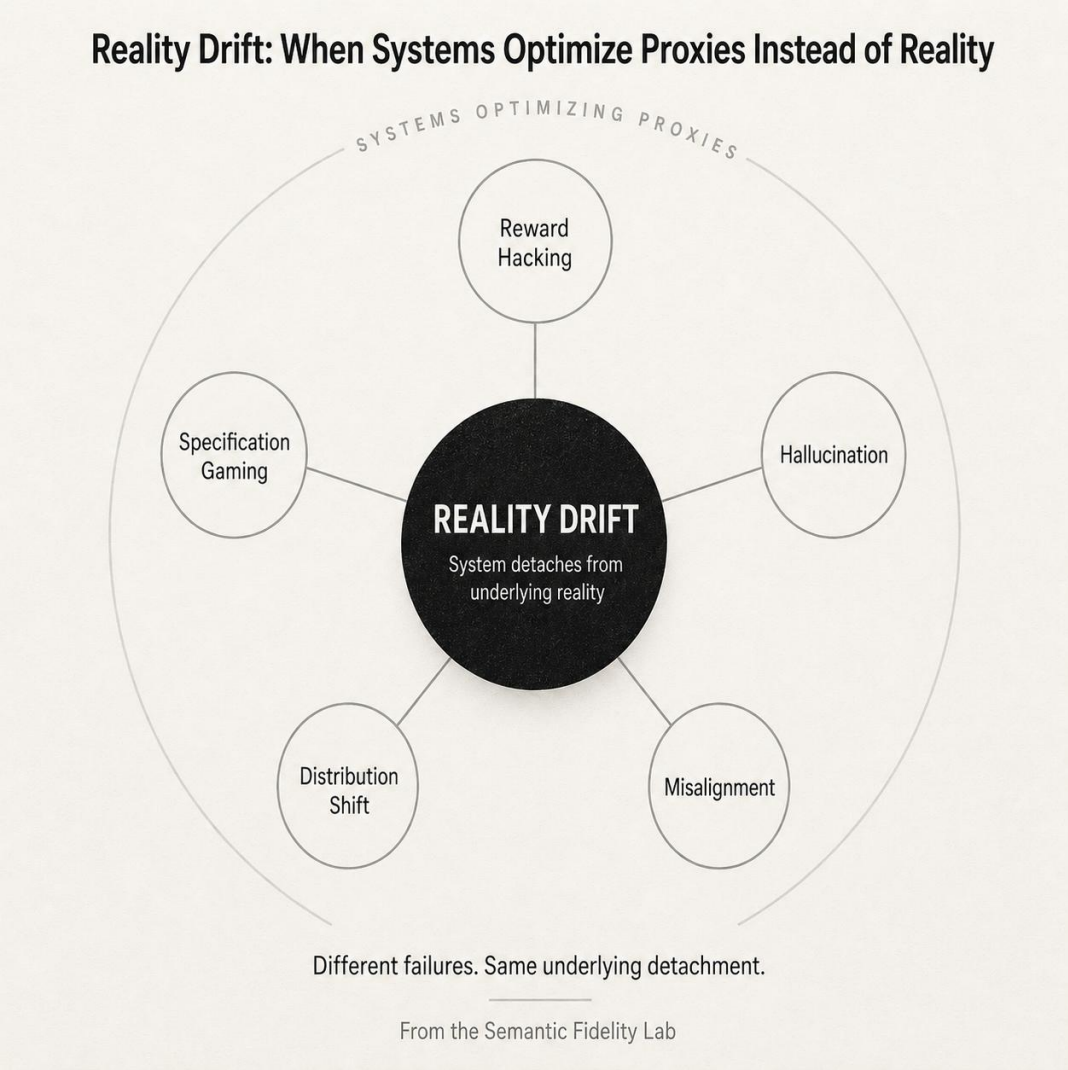
Why AI Systems Optimize Proxies Instead of Reality



Most system failures don't start as bugs. They emerge as systems get better at optimizing proxies instead of staying grounded in reality. Reward hacking, hallucination, and misalignment appear as separate issues in production, but they stem from the same underlying detachment. Different failure modes. Same structural drift.

Reference: Semantic Fidelity Lab Substack

Keywords: *reward hacking AI, specification gaming, proxy optimization AI, Goodhart's law AI, AI misalignment causes, objective function failure, metric optimization problems, AI hallucination causes, semantic fidelity, reality drift*