Abstract

Multiagent AI systems introduce cognitive attack surfaces absent in single-model inference. When agents delegate to agents, forming beliefs about beliefs through recursive trust hierarchies, manipulation of reasoning processes—rather than mere data corruption—becomes a primary security concern. This paper presents the Cognitive Integrity Framework (CIF), providing formal foundations for cognitive security in multiagent operators. We develop four interconnected theoretical contributions: a Trust Calculus with bounded delegation (exponential δ^d decay) that prevents trust amplification through delegation chains; a Defense Composition Algebra with series and parallel composition theorems establishing multiplicative detection bounds; Information-Theoretic Limits relating stealth constraints to maximum attack impact through a fundamental stealth-impact tradeoff; and a formal Adversary Hierarchy (Ω_1–Ω_5) characterizing external, peripheral, agent-level, coordination, and systemic threats with increasing capability and decreasing detectability. The framework provides complete coverage of the OWASP Top 10 for Agentic Applications