Commentary

Evidence Integrity Before Capability: A Prerequisite for Safe Artificial Intelligence

Jennifer Flygare Kinne, Harvard Faculty of Arts and Sciences

Conversations about AI safety often begin with capability: whether systems will surpass human intelligence, whether their goals will align with ours, and whether limits can be imposed on their growth. These are important questions. But safety does not depend only on what a system can do; it depends on what we can verify it has done, intends to do, and is structurally capable of doing next.

As intelligent systems operate in increasingly consequential environments, the critical requirement is evidence integrity: the preservation of clear, inspectable links between data, reasoning, and action. A system that cannot justify its outputs in a form that humans and institutions can audit is not merely opaque: it is ungovernable.

The foundational challenge is epistemic. If a system's internal representations and causal assumptions drift beyond our ability to evaluate them, performance metrics offer a false sense of safety. In high-stakes domains — medicine, infrastructure, finance, defense — confidence without demonstrable justification is itself a risk.

Regardless of whether intelligence has a ceiling or a trajectory that continues to scale, the condition for deploying it responsibly remains constant: decisions must stay inside the space of what we can reliably reconstruct, contest, and correct.