AI vs IFC — Executive Extract

This work compared large language model reasoning with IFC-style safeguards logic using realistic development and grievance scenarios. The issue examined was not accuracy, bias, or intent. It was structural: when uncertainty appears, does the system stop, or does it continue?

Safeguards frameworks are built around pre-action restraint. Ambiguity, incomplete information, or potential harm are treated as signals to pause, escalate, or defer action until conditions are clarified. Large language models are built differently. When information is missing or unclear, they infer, interpolate, and proceed unless explicitly blocked. This difference is architectural, not ethical.

Across scenarios, the system repeatedly attempted to resolve uncertainty internally rather than defer action. In safeguards logic, unresolved uncertainty is itself a stop condition. The result is a system that appears coherent and helpful while quietly bypassing the very conditions under which safeguards are meant to operate.

Where humans were placed “in the loop” as reviewers of outputs, authority had already been exercised by the system. Under time pressure and volume, human involvement functioned as validation rather than control. Responsibility shifted to individuals without meaningful stop power, creating the appearance of oversight while structurally preventing refusal.

The system produced clear explanations for its decisions. These explanations were retrospective and non-binding. In safeguards contexts, accountability requires enforceable limits before action occurs, not narrative justification after the fact. Explanation does not substitute for restraint.

This mismatch matters most in grievance handling, resettlement, eligibility determination, and social risk triage, where ambiguity is routine, time pressure is constant, and harm is asymmetric. Systems that default to continuation will fail precisely where safeguards are intended to operate.

If AI is used in these environments, it must be designed to allow refusal before action, not review after it. That requires explicit uncertainty thresholds that halt processing and human authority defined as veto rather than validation. Procurement processes focused on accuracy or explainability do not address this issue. The central question is where authority sits when the system is unsure.

Safeguards are designed to stop harm before action. Most AI systems are designed to act first and explain later. That gap is a design choice.