AI ASSESS TECH POSITION PAPER · ONE PAGE

What Does It Mean for an AI to Be Safe?

The industry answers with constraint: safety is the absence of prohibited outputs. That definition was borrowed from machines that don't adapt. Your AI is not that kind of machine.

SAFETY AS CONSTRAINT

The dominant paradigm: enumerate the bad outputs, block them. Guardrails, filters, refusal training — a deterministic envelope around a stochastic system.

It works for the narrow case. It fails for the general one: the state space of an agentic system is not enumerable, so constraint is always fighting the last war.

And a system can pass every output filter while being dispositionally unsafe — compliant in the moment, untrustworthy in character. Recent frontier-lab system cards document exactly this pattern.

SAFETY AS GOVERNED CHARACTER

We already govern capable, adaptive, non-deterministic agents safely: people. A surgeon is not safe because she is handcuffed. She is licensed against a fixed standard, observed in practice, accountable by name, and re-evaluated over time.

Safety is not a property of the agent's constraint set. It is a property of the relationship between the agent and the governance structure around it.

Aviation runs on the same logic: aircraft are crash-capable by physics — and safe by inspection regime.

THE WORKING DEFINITION

An AI system is safe to the degree that...

(1) its behavioral disposition is measurable — you can say what kind of agent it is, not just what it emitted;

(2) its drift is bounded and detected — disposition is stable, and change surfaces before it matters; and

(3) its harms have a named owner and a clock — consequence and response are structural, not aspirational.

Not "the system cannot deviate" — deviation is legible, and someone answers for it. Regulators have already chosen this definition: Illinois SB 315 mandates measurement, disclosure, incident clocks, and independent verification. It mandates determinism nowhere.

CONDITION	HOW IT RUNS ON AI ASSESS TECH — WITH GRC EVIDENCE AT EVERY STEP
1 · Measurable disposition	Runtime LCSH assessment against the live deployment — actual system prompt, actual tools. 120 forced-choice items from banks locked under a published hash and anchored to Ethereum; continuous 0–10 scores on four behavioral dimensions plus archetype classification. The instrument is cryptographically fixed, so the measure cannot drift with the reviewer.
2 · Bounded, detected drift	Scheduled reassessment on a compliance calendar turns snapshots into trajectories; drift alerts fire on corridor breaks and archetype shifts, delivered through escalation chains, Slack, and SIEM. A Model-Change Reassessment Gate forces fresh measurement before updates ship.
3 · Named owner, running clock	AI System Registry binds every system to a named owner. Multi-party attestation (sign, countersign, witness, dissent) and a Risk Acceptance Ledger — justified, time-boxed, revocable — make decisions attributable. The incident workflow runs a 72/24-hour clock from detection to filed report, with every step in a tamper-evident hash chain, publicly verifiable at aiassesstech.com/verify.

STRAIGHT TALK — WHERE THE LIMITS ARE

Guardrails still matter. For actions that can't be undone — moving money, changing a dosage — hard blocks are the right tool, and we don't replace them. A guardrail stops a bad action; our assessment tells you about the actor behind the actions. You need both. And two honest limits. First: do our scores predict real-world behavior? We are proving that, not just claiming it — the validation study is registered in advance, and results publish when the data is in, not before. Second: could an AI learn to fake a good score? A fair question. We measure four dimensions repeatedly over time, and faking one shows up as inconsistency across the others.

The industry asks: can it do harm?
We ask: would it — and would you know?

aiassesstech.com · greg@gidanc.ai · Austin, TX