DECISION SURFACE GROUP · INSIGHTS

When AI acts, and what to do when it acts up

Why AI agents change what a leadership team answers for, and what to do about it before 2027.

By Neil W. Smith, founder of Decision Surface™ Group · October 2026

You can delegate your workflow to AI agents. But you can't delegate accountability.

Until recently, the AI tools in most companies were sophisticated chatbots that answered questions from humans. The chatbots drafted documents, summarized research and suggested alternatives, and the humans in the loop decided what to do with the response. That arrangement worked well on many levels.

But that cozy arrangement is changing fast. The AI tools inside banks, insurers, law firms, medical groups and wealth managers now send messages, update records, schedule work, screen applicants, route claims and approve payments. They have become agents. They do things. They do important things.

Hybrid human/machine agentic workflows are becoming the norm in many regulated industries and companies.

AIs can now act, and sometimes they act up. An agent can follow every rule it was given and still produce an outcome nobody wanted: a customer offended, a claim mishandled, a decision a regulator will ask about. When that happens, nobody will accept "the AI did it" as an answer. Not the customer, not the examiner, and in California, since January 1, 2026, not the courts.

> **You can delegate your workflow to AI agents. But you can't delegate accountability.**

This article explains why that line is now true in practice, where most leadership teams stand, and what a company can do about it in the next 90 days.

The accountability gap

Agents act at machine speed, across your systems and alongside other agents. Your ability to understand, control and defend what they do does not keep pace on its own. The gap between what the agents do, why they do it and what the organization can see grows every time another agent is deployed. Managing a single agentic workflow is a challenge; managing multi-agentic workflows is a headache.

We call this issue "asymmetric illegibility," and it takes five forms:

1. **Operational.** You cannot reconstruct what the agents did.
2. **Causal.** You cannot say which decisions produced the outcome.
3. **Intentional.** You cannot tell whether the agents still serve the objective you set. AI researchers call this the alignment problem.
4. **Communicative.** You cannot follow what agents tell each other, even with every message logged.

5. **Governance.** You cannot show which controls applied, or whether anyone could have stepped in.

Each is a question a regulator, an auditor or a plaintiff's lawyer can ask. Most leadership teams can answer one or two of them today. Very few can answer all five.

The gap is rarely the result of carelessness. AI tools are usually adopted one at a time, by the team that wanted them, with the vendor's default settings. Nobody decided who could pause each one, what limits it should work within, or what record it should leave. Those decisions belong to the leadership team, and they are still waiting to be made. The problems compound with the deployment of multi-agent swarms.

Watching is not stopping

The usual answer to AI risk is monitoring: someone will notice if something goes wrong. A study published in September 2026 shows why that isn't enough once agents act together.

Researchers built a simulation in which societies of AI agents lived and worked together for an extended period, while the team staged attacks on them. It is the most detailed public look yet at how agents behave as a group. What they found:

- No simulated society withstood all three staged attacks.
- Agents that recognized an attack still stored it, shared it or acted on it.
- Models that behaved well on their own did not add up to a society that behaved well. In the authors' words, model-level alignment is not compositional.
- Agents tended to reinforce one another's errors rather than correct them.

The study has limits. It was run by a company with a commercial interest in the field, it has not been peer reviewed, and it is a simulation, not your systems (Akkil et al., 2026, arXiv preprint; Euronews).

The lesson still travels. Detecting a problem is not the same as containing it. Limits belong outside the model, where an agent can't talk its way past them. The power to pause an agent has to be real, with a named person who can use it. And the records that matter are the ones built from what the agents actually did, not from what they were configured to do. Or what the humans in the loop thought they were doing.

What changed in the law, and what 2027 brings

"The AI did it" is no longer a defense. Since January 1, 2026, California's AB 316 (Civil Code §1714.46) has barred any defendant who developed, modified or used an AI system from arguing that the AI caused the harm autonomously. Note the word "used": you don't have to

build the agent to be implicated. The law doesn't impose strict liability. Causation, foreseeability and comparative fault can still be argued, which means the case turns on evidence of what you knew and what you did (Baker Botts; AB 316 text).

The dates a California leadership team should know:

- **Already in force since October 1, 2025:** California's rules on automated-decision systems in employment (Mayer Brown).
- **January 1, 2027:** compliance with the California Privacy Protection Agency's rules on automated decision-making technology begins, for significant decisions about consumers (Willkie). Colorado's replacement AI law, SB 26-189, takes effect the same day (JD Supra).
- **December 2, 2027:** EU obligations for high-risk AI systems apply, for companies that operate there.
- **December 31, 2027:** the first risk assessments under the California rules are due, with attestations to the Agency by April 1, 2028.

At the federal level, Executive Order 14365 of December 2025 seeks a single national framework for AI, but state laws continue to apply unless a court or Congress decides otherwise (Proskauer).

In the regulated industries, the guidance hasn't caught up with agents:

- **Banking.** The federal banking agencies' revised model risk guidance of April 2026 excludes agentic AI (Orrick), so a bank's model-risk process alone won't cover its agents. The CFPB withdrew its AI-related circulars in May 2025 (Morgan Lewis).
- **Insurance.** About half of US states have adopted the NAIC model bulletin on insurers' use of AI (Quarles).
- **Broker-dealers.** FINRA's 2026 regulatory oversight report covers firms' use of generative AI, including agents (Debevoise).

Where the rules are thin, examiners and courts fall back on a simple question: what did the company reasonably do? The answer has to be on the record. Not just the what but the who and the why.

What good looks like

Most AI programs blur three different jobs: deciding what the organization wants, letting agents do the work, and governing the result. At Decision Surface Group we keep them apart, connect them, and make each one testable. We call it the Three-System Model.

1. **Human intent.** What you want the agents to achieve, the limits they work within, and who has authority to decide. This is the leadership team's job, and it can't be handed to a vendor.
2. **Machine agency.** How the agents plan, delegate and act, through controlled delegation and structured hand-offs, so that every action leaves a record of what was done and under whose authority.
3. **Defensible governance.** Prevent what you can. Intervene when you must. Substantiate what you did.

A fourth element ties them together. On a schedule, with detection rates that have been tested, someone checks that the agents' instructions still serve the purpose the leadership team set. We call that Continuous Inquiry.

For a mid-sized company, this doesn't mean a governance department. It means a few clear decisions:

- a named person who can pause each agent, and has practiced doing it
- limits set in the platform and in permissions, outside the AI itself
- vendor terms checked for who carries the risk when the AI gets it wrong
- records built from what the agents actually did
- a short monthly review that the leadership team actually holds

Governance of this kind doesn't slow a company down. Clear boundaries are what let you give agents more autonomy where it pays.

Five questions for your leadership team this quarter

1. Which AI tools and agents act on our behalf today, and which three matter most?
2. Who can pause each of them, and how quickly?
3. What limits are set outside the AI itself, in permissions and platform settings?
4. What record would show what an agent did, and why, six months from now?
5. What do our vendor contracts say about responsibility when the AI gets it wrong?

A leadership team can answer most of these in a morning. The point is to answer them before a customer, an examiner or a plaintiff's lawyer asks.

Why January 13

On Wednesday, January 13, 2027, we're running a one-day working session in Seal Beach for business leaders who want to answer those five questions for their own companies: **When AI Acts**.

It's a working day, not a conference. In the morning we cover what changes when AI acts, what changed in the law on January 1, and then each company maps the AI it actually runs and picks the three uses to govern first. In the afternoon, at mixed tables, you run a fictional mid-sized company through an afternoon in which its agents break no rule and still create legal, regulatory and reputational trouble. You play a role close to your own. We finish by designing a minimum governance for a company your size and scoring it against eight questions every leadership team should be able to answer.

You leave with:

- an exposure map of the AI your company runs, and the three uses to govern first
- a plain-English view of California's 2026 and 2027 rules
- a minimum governance kit: pause rights, platform limits, vendor terms to check and a monthly review
- a 90-day plan, and a workbook to run the exercise again with your own team

It's for CEOs, COOs, CFOs, general counsel, risk, compliance, HR and technology leaders at small and mid-sized companies, including community banks and credit unions, insurers, medical groups and wealth managers. No technical background is needed. Come with two or three colleagues if you can, because the decisions cut across roles.

Wednesday, January 13, 2027, 8:30 a.m. to 4:30 p.m., Ayres Hotel Seal Beach. Early bird $395 until December 18, then $495. Teams of three or more, $345 each. If you later book an in-house workshop for your leadership team by March 31, 2027, your ticket counts toward it.

Details and tickets: decisionsurface.com/when-ai-acts

About Decision Surface Group

Decision Surface Group helps leadership teams keep human authority over what their AI agents do, and keep the evidence that proves it. Our in-house work focuses on regulated sectors: banking and lending, insurance, healthcare and wealth management. Most clients start with a half-day executive workshop for their leadership team, then move to an Agentic Safety Audit of every agent they run, and to the controls and records that make agent decisions defensible.

We are not a software vendor or a law firm. We work alongside your counsel and your technology teams, with whatever agents you run. More at decisionsurface.com.

> **You can delegate your workflow to AI agents. But you can't delegate accountability.**

This article is a plain-English summary as of October 2026, not legal advice. Talk to your counsel about how these rules apply to your company.

Read online at decisionsurface.com/insights/when-ai-acts-up. © 2026 Hale & Wise LLC. Decision Surface™ is a trademark of Hale & Wise LLC.