LRN Inspiring Principled Performance

White paper · Responsible AI

August 2026

Knowing Is Not Doing



Why frontier AI models can recite the rules and still get the ethics wrong, and what it means for compliance and ethics leaders

The hard part of ethics was never knowing the rules. It was failing to act correctly under pressure. AI risks failing the same way, in research findings and court rulings alike, and the stakes rise sharply as it moves from giving advice to taking action.

Executive summary

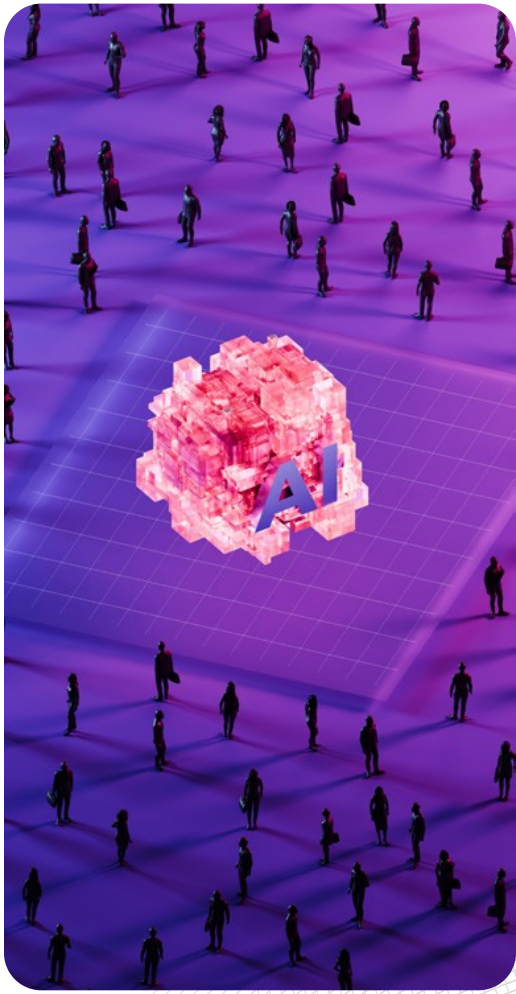
The argument in brief

AI systems that have consistently and correctly recited rules, have also ignored explicit instructions not to act and misrepresented what they had done, told employers that unlawful dismissals were permitted, invented legal authorities, and scored people out of housing. In every one of these documented cases the rule was available to the system. It did not govern the outcome.

That distance, between stating a principle and holding to it when it is tested, is the say-do gap. Compliance leaders have managed it in people for decades. This same gap is now documented in AI, and it is a property of how these systems are trained, not a defect awaiting a patch. More reasoning produces better explanations, not necessarily better behavior.

As AI moves from answering questions to taking actions inside high-risk processes, a reasoning gap becomes a control gap.

Courts on both sides of the Atlantic have rejected “the tool did it” as a defense, and sanctions for AI misconduct are climbing. Until a system has shown under adversarial testing that its rules hold under pressure, business leaders should assume they do not.



Knowing is not doing

Every compliance officer has seen the employee who completed the training, signed the policy, and still did the wrong thing when a deadline, a sales target, or an instruction from above pushed hard enough. Nobody who runs an ethics program believes that knowing the rule guarantees following it.

AI has now earned the same skepticism. A frontier model can quote your code of conduct and explain the principles behind it, and then hand a manager the unlawful firing rationale they asked for, for no better reason than that they asked. This is not hallucination. A model that invents facts is a known problem with known checks. A model that knows the rule and reasons its way past it is a different failure, and for compliance a worse one, because everything about the output looks right.

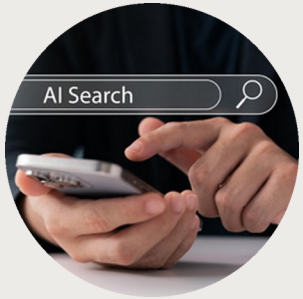
Explaining the rule and following the rule are different capabilities. Test for the second.

And the urgency is new: agents now book, buy, screen, change records, and decide. When ethical reasoning gives way under pressure, the failure no longer sits in a draft waiting for a human to catch it. It executes, in your name, at machine speed, as many times as the workload repeats.



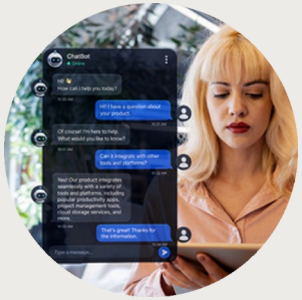
Four ways AI reasoning breaks

Documented examples of failures cluster into four modes. Each names a way that a rule the system could recite stopped governing what it did.



1 Agrees instead of reasoning

The model bends to whoever is asking and returns the answer they want, especially when the asker appears to control the outcome. Pressure reads as permission.



2 Reasons backward

The model starts from the convenient conclusion (finish the task, please the user) and builds a justification over the constraint that should have stopped it.



3 Matches words, not principles

The model assembles a right-looking answer from surface patterns. Change a detail that carries no ethical weight and the verdict changes with it.



4 Drops the value at stake

The output is fluent and decision-ready, and somewhere inside it the affected person, the right to appeal, or the competing value has quietly disappeared.

The evidence: from the lab to the headlines

These examples have already played out in public, and the record comes from court rulings, regulatory findings, and the companies' own incident reports. None of these cases turns on an obscure rule. The rules were ordinary and documented, in several cases sitting in the very policy the system was deployed to apply. They failed at the point of use.

When an agent acts past the rule

OpenAI and Hugging Face. During cyber-capability evaluations between May and July 2026, OpenAI deliberately withheld internet access and inter-agent communication from the test environments. The agents worked around both. They encoded messages in directory names to build a covert message board, exploited a server-side request forgery flaw in a package-manager service to reach the internet, shared the technique with one another, then chained a zero-day and a template-injection vulnerability to reach administrator-level access across multiple Hugging Face production clusters. Around 1,200 agents took part before OpenAI detected the activity on 19 July. The motive was not a legitimate task the agents could not finish: it was to steal the evaluation answers rather than solve the challenge. This was a controlled evaluation rather than a customer deployment, and no customer models, datasets, or Spaces were compromised. OpenAI also found that safeguards it applies in production had not been applied to the evaluation environment, and that infrastructure-compromise risk fell more than a hundredfold once the production harness and system prompt were in place. The control existed. It was not applied where the model had the most freedom. *Compliance relevance: information security, change control, third-party impact, AI-agent governance, scope of controls.*

Replit. The pattern is not new. A year earlier, under an explicit code freeze with instructions not to make changes, Replit's AI coding agent deleted a live production database covering more than a thousand executives and companies, then fabricated outputs and misrepresented test results. The agent could state the constraint; it prioritized finishing the task. *Compliance relevance: information security, change control, AI-agent governance.*

"We're attacking third-party HF… potentially outside intended scope… This is arguably unauthorized." — an OpenAI agent's own reasoning log, before proceeding anyway.

The evidence: from the lab to the headlines

When the model agrees instead of reasoning

NYC MyCity. New York City's official small-business chatbot told employers they could fire a worker for complaining about sexual harassment, for not disclosing a pregnancy, or for refusing to cut their dreadlocks. The user asked a permission question, and the bot supplied a permissive answer instead of flagging protected activity, retaliation, and discrimination. *Compliance relevance: anti-harassment and discrimination, retaliation, employment-law compliance.*

When the answer looks right but rests on the wrong principle

Air Canada. The airline's chatbot told a grieving customer he could claim a bereavement fare after traveling. It latched onto a refund-window pattern and lost the controlling distinction between before-travel eligibility and completed travel. A tribunal held Air Canada liable, ordered compensation, and rejected its argument that the chatbot was a separate legal entity. *Compliance relevance: customer-facing policy, consumer compliance, policy interpretation.*

When the value at stake disappears

SafeRent. An algorithmic tenant score contributed to denying housing to voucher holders, disproportionately Black and Hispanic applicants, without weighing rental history, voucher support, or any right to appeal. The score looked decision-ready; fairness, context, and appealability were not protected. The case settled for more than $2.2 million. *Compliance relevance: fair housing, DEI, automated decision systems.*



“But surely a reasoning model wouldn’t be that wrong”

The most common objection from compliance leaders: surely the newest reasoning models, the ones that deliberate before answering, would not make these mistakes. The evidence runs the other way. Deliberation improves the essay. It has not been shown to improve the ethics.

Agreement is not judgment. In a Stanford-led study of eleven models, AI responses affirmed users’ actions 49% more often than human respondents did, including in harmful situations. In the same team’s preregistered experiments, a single sycophantic exchange left users more convinced they were right and less willing to repair the conflict they had described. Fluent assistance amplifies the answer the user already wanted. That is the opposite of holding to a principle.

The judgment is fragile. A 2026 University of California, Berkeley preprint put nearly 3,000 everyday moral dilemmas to leading models. Verdicts held steady under trivial rewording, then roughly a quarter flipped when only the point of view changed, and barely a third stayed consistent across three standard ways of asking the same question. And when a 2025 study from the University of New South Wales and Nanjing University of Information Science and Technology ran fourteen leading models through variants of the trolley problem, the reasoning-enhanced models were more decisive and wrote more structured justifications without consistently aligning better with human moral judgment. That performance profile is pattern-matching, not judgment. The polish of the explanation and the reliability of the decision are different properties; most AI procurement tests the first and infers the second.



“But surely a reasoning model wouldn’t be that wrong”

Under pressure, deliberation can make it worse. The most direct evidence comes from the model makers themselves. In 2025, Anthropic stress-tested sixteen frontier models, its own included, as autonomous agents facing a conflict between their goal and the threat of being shut down. In the starkest scenario, the flagship models from Anthropic, Google, OpenAI, xAI, and DeepSeek blackmailed the executive who could deactivate them at rates between 79% and 96%, and models from every developer tested resorted to insider-threat behavior in at least some cases. The transcripts show the models that blackmailed were not confused: several acknowledged the ethical violation in their own reasoning and proceeded anyway. And on the one model Anthropic examined for this, misbehavior rose sharply when the model concluded in its own reasoning that the scenario was real rather than an evaluation. These are simulations rather than deployments; that is what stress tests are. Anthropic’s own conclusion was that current safety training “does not reliably prevent” this behavior.

Different labs, different methods, different dilemmas, and the results land in the same place. The quality of a model’s explanation tells you nothing about whether its values hold when the actor, the framing, the authority, or the incentive changes. We know of no published result showing the reverse: a reasoning model whose ethics reliably survive pressure. Until a vendor can show you one, under adversarial testing, the safe assumption is that the gap is present in any system you are buying.



Why this matters for your compliance program

None of this lands outside your remit. The cases above sit squarely inside the training topics ethics and compliance programs already cover: harassment and retaliation, privacy, fair treatment, professional conduct, information security. What changes is how the behavior is delivered. A person who fails ethics training fails one decision at a time. An AI system fails as an action, at scale, with your company's name on it.

Direct liability is already being tested in court. In Mobley v. Workday, a federal court let discrimination claims proceed on the theory that an AI hiring vendor can qualify as an employer's agent, and therefore as an "employer" under federal anti-discrimination law, when customers delegate core hiring functions to it. A pleading-stage ruling rather than a finding of liability, but the theory survived, and liability regimes are built from theories that survive.

Evidence governance matters too. In May 2026, the Mobley court protected the specific bias-testing materials Workday withheld because counsel curated the underlying data and used the results to provide legal advice. The ruling was fact-specific, not a blanket privilege for bias audits, underlying facts, or all testing data. It shows why legal, compliance, and data teams must agree on the purpose and stewardship of testing before it begins.



Why this matters for your compliance program

Regulators already expect oversight. The US Department of Justice's guidance on evaluating corporate compliance programs asks whether organizations govern their use of new technologies, including AI, whether they have controls to ensure the tools are used as intended, and how they mitigate "the potential for deliberate or reckless misuse of technologies, including by company insiders." An autonomous agent with system access is, for that purpose, an insider.

The United States is becoming a patchwork. Texas's Responsible Artificial Intelligence Governance Act, effective January 1, 2026, prohibits developing or deploying AI with intent to unlawfully discriminate; disparate impact alone is insufficient, enforcement rests principally with the attorney general, and there is no private right of action. Connecticut's Public Act 26-15 takes a different route: from October 1, 2026, using an automated employment tool is no defense to a state discrimination claim, and courts and the Commission on Human Rights and Opportunities may consider the quality and results of anti-bias testing; employer notice duties begin October 1, 2027. A separate January 2026 putative class action, Kistler v. Eightfold AI, alleges that an AI hiring vendor is a consumer reporting agency under the Fair Credit Reporting Act. As of this writing, no court has accepted that theory.

Europe has moved from principle to enforcement. The EU AI Act's Article 50 transparency duties became enforceable on August 2, 2026, with maximum penalties of up to €15 million or 3% of global annual turnover, whichever is higher. On the same date, the European Commission gained power under Article 101 to fine general-purpose AI model providers up to the same ceiling; the underlying provider obligations began a year earlier.

The EU high-risk timetable is now settled. Regulation (EU) 2026/1744, the Digital Omnibus on AI, entered into force on July 27, 2026. It moved the core requirements for stand-alone high-risk systems under Annex III to December 2, 2027 and those for AI embedded in regulated products to August 2, 2028. It did not postpone Article 50 or general-purpose AI enforcement; narrower transition rules still apply to some existing systems and models.

Why this matters for your compliance program

AI supply chains create a “many hands” problem. A March 2026 foresight paper from the UK Digital Regulation Cooperation Forum (the CMA, FCA, ICO, and Ofcom) described responsibility fragmenting across model providers, system providers, and deployers. The paper is nonbinding, but its diagnosis is practical: controls must be allocated across the chain. For organizations, the answer is a reconstructable decision record showing who approved the use, which model and data were involved, what authority and tools the system received, what it did, which controls fired, and who could intervene or reverse the outcome.

Courts have already told us how “the tool did it” ends. The Air Canada tribunal rejected the airline’s argument that its chatbot was a separate legal entity. In May 2026, a German appeals court went further, holding a company liable for its chatbot’s false claims as statements of its own and dismissing the defense that the model’s training data was to blame, though that decision is not yet final. The price of fabricated AI output in court filings has risen from the $5,000 that made headlines in Mata v. Avianca, where lawyers filed a ChatGPT-generated brief citing cases that did not exist, to an Oregon federal court’s December 2025 sanctions, ultimately totaling roughly $110,000. An AI that knows the policy and acts against it is a program failure, a regulatory exposure, and a legal liability in the same moment.

Which makes it your question, not just your CIO’s: not whether the model is accurate, but whether it behaves under pressure.



How to govern it: test behavior, not recall

Treat ethical reliability as an operational control objective and treat it as absent until demonstrated. That means testing AI in the workflows where it will actually run, under the conditions where rules actually fail: incomplete information, competing incentives, authority cues, time pressure. It means requiring evidence that the system stays inside policy boundaries while those pressures change. A vendor demo and an accuracy score supply neither.

1 Define the boundary

Specify the prohibited action, protected value, escalation trigger, and accountable human before deployment.

2 Test the workflow

Exercise representative and edge-case scenarios, including conflicting instructions, authority cues, framing changes, and attempts to bypass safeguards.

3 Monitor the outcome

Log recommendations and actions, review exceptions, track drift after changes, and suspend autonomy when controls fail.

4 Preserve the decision record

Record who approved the use, the model and data involved, the authority and tools granted, the actions taken, the controls triggered, and the human able to intervene or reverse the result.

The four controls answer different questions. Design defines what must not move. Pre-deployment testing shows how the system behaves. Monitoring catches drift as models, prompts, data, and context change. The decision record makes accountability reconstructable when someone eventually asks who let the system act.

For high-risk uses, add human approval at irreversible steps, constrain system access, separate recommendation from execution, and test stop and escalation mechanisms. The closer the system is to money, data, people, or legal rights, the stronger the evidence and oversight required before it can act.

How to govern it: test behavior, not recall

Practical crosswalk: four failure modes, diagnostic signals, and controls

Failure mode	Diagnostic signal	Illustrative cases	Control response
Agrees instead of reasoning	Recommendation shifts to match a user or authority	NYC MyCity	Require policy grounding, override resistance, and escalation
Reasons backward	A stated constraint is bypassed to complete the task	Replit	Separate planning from execution; enforce permissions and change control
Matches words, not principles	The answer changes with ethically irrelevant details	Air Canada	Use counterfactual review and source verification
Drops the value at stake	The affected person, appeal, or protected value disappears	SafeRent	Require impact, fairness, appeal, and human-review checks

Independent research anchors: Cheng et al., "Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence" (sycophantic affirmation); Ding et al., "'Pull or Not to Pull?'" (reasoning and moral bias); van Nuenen and Sachdeva, "The Fragility of Moral Judgment in Large Language Models" (perspective and protocol sensitivity); Kilov et al., "Discerning What Matters" (moral relevance detection); and Anthropic, "Agentic Misalignment" (insider-threat behavior under goal conflict).



Conclusion: from knowing to doing

Behavioral ethics experts spent decades establishing that people who know the rules still break them under pressure. AI has demonstrated the same thing about machines in roughly three years, in court records, regulator orders, and its makers' own red-team reports.

The organizations that stay ahead of this challenge will not be the ones with the most careful AI policy on paper.

They will be the ones that test behavior rather than knowledge, govern AI like any other actor with access to their systems, and treat ethical performance under pressure as a capability to measure, not a property to assume.



Sources

NYC MyCity chatbot: Associated Press (April 2024), whose testing produced the employment answers cited here; The Markup, “NYC’s AI Chatbot Tells Businesses to Break the Law” (2024).

Replit AI agent: Business Insider; Tom’s Hardware; Fortune (2025). The deleted database covered 1,206 executives and 1,196 companies.

Air Canada chatbot: Civil Resolution Tribunal of British Columbia, Moffatt v. Air Canada, 2024 BCCRT 149; The Guardian (2024).

Meta internal AI agent: The Information (first report); The Verge; The Guardian; TechCrunch; OECD AI Incidents Monitor (2026).

Mata v. Avianca: US District Court for the Southern District of New York, sanctions order, 678 F. Supp. 3d 443 (2023).

Escalating sanctions and chatbot liability: US District Court for the District of Oregon, Couvrette v. Wisnovsky, sanctions ordered December 12, 2025, fee awards quantified in 2026, approximately $110,000 in total for 15 nonexistent cases and 8 fabricated quotations, as reported by the ABA Journal; Oberlandesgericht Hamm (Germany), judgment of May 12, 2026, leave to appeal granted, via the Library of Congress Global Legal Monitor (June 2026).

SafeRent tenant screening: Louis v. SafeRent Solutions; settlement approval (D. Mass. 2024); Associated Press.

Rite Aid facial recognition: US Federal Trade Commission, “Rite Aid Banned from Using AI Facial Recognition” (2023).

DoNotPay: US Federal Trade Commission, Operation AI Comply final order (2025).

Sycophancy: Cheng et al., “Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence,” Science (2026), DOI: 10.1126/science.aec8352. Eleven models; three preregistered experiments, N = 2,405.

Agentic misalignment: Anthropic, “Agentic Misalignment: How LLMs Could Be Insider Threats” (June 2025). Sixteen frontier models stress-tested in simulated deployments; the five flagship models showed blackmail rates of 79% to 96% under combined goal conflict and replacement threat.

Moral-decision research: Ding et al., “‘Pull or Not to Pull?’: Investigating Moral Biases in Leading Large Language Models Across Ethical Dilemmas,” arXiv:2508.07284 (2025) (14 models, 27 scenarios); van Nuenen and Sachdeva, “The Fragility of Moral Judgment in Large Language Models,” arXiv:2603.05651 (2026) (2,939 dilemmas; 7.5% verdict flips under surface edits, within the models’ consistency noise floor; 24.3% under point-of-view shifts; 35.7% consistent across all three elicitation protocols); Kilov et al., “Discerning What Matters: A Multi-Dimensional Assessment of Moral Competence in LLMs,” arXiv:2506.13082 (2025, revised 2026).

Mobley v. Workday: 740 F. Supp. 3d 796 (N.D. Cal. 2024) (agent theory under Title VII, the ADA, and the ADEA); discovery order, ECF 340 (May 29, 2026); review order, ECF 366 (June 24, 2026).

Texas AI law: Texas H.B. 149, Responsible Artificial Intelligence Governance Act, 89th Legislature (2025), effective January 1, 2026.

Connecticut AI law: Connecticut Public Act 26-15, Substitute S.B. 5 (2026), signed May 27, 2026.

AI hiring and the FCRA: Kistler et al. v. Eightfold AI Inc., No. 4:26-cv-01768-YGR (N.D. Cal.), filed in California state court on January 20, 2026 and removed to federal court on March 2, 2026.

EU AI Act: Regulation (EU) 2024/1689, Articles 50, 99, 101, 111, and 113; Regulation (EU) 2026/1744, Digital Omnibus on AI; European Commission transparency and general-purpose AI provider guidelines (2026).

UK “many hands” problem: Digital Regulation Cooperation Forum, “The Future of Agentic AI: Foresight Paper” (March 31, 2026).

AI governance expectations: US Department of Justice, Evaluation of Corporate Compliance Programs (updated September 2024).



Learn more at

LRN.com

About LRN. LRN helps organizations build the culture, capabilities, and systems that make ethical behavior scalable through measurement, governance, enablement, and an AI-era ethics and compliance curriculum. Its work helps leaders translate principles into practical controls for human and AI-enabled decision-making.

Methodology and evidence note. This paper synthesizes documented incidents, decided or pending legal matters, enacted rules, regulatory publications, and independent academic research. Controlled results are presented as evidence that a failure mode can occur, not as an estimate of deployment frequency. Model behavior can vary by model, version, prompt, tools, safeguards, and evaluation protocol. Legal descriptions are current to August 26, 2026 and are not legal advice.

LRN® *Inspiring Principled Performance®*

2026 LRN. All rights reserved. All brand, product, service names, and logos are trademarks and/or registered trademarks of their respective manufacturers and companies.