Negative Effects from Mandatory AI Liability Insurance

Executive Summary

This memo catalogs the negative effects of a mandatory AI liability insurance regime and shows how modern insurance—through contractual cost-sharing, risk-based pricing, and direct oversight—mitigates rather than amplifies them, identifies the residual gaps (true judgment-proofness, correlated shocks, exogenous doctrinal change), and explains why insurance still improves incentives where labs are thinly capitalized.

I. Taxonomy

In his 2003 paper, Christopher Parsons outlines a four-channel account of how third-party liability insurance can reshape behavior: once a defendant's losses are spread through insurance, not only can defendants adjust their level of care, but claimants, courts and legislatures, and even insurers may respond in ways that amplify loss or uncertainty.[1] He labels these channels policyholder hazard, claimant hazard, jurisprudential hazard, and underwriting hazard, and argues that their interaction is most acute in long-tail, information-poor liability lines.[2] Of these, policyholder hazard corresponds to the traditional, ex-ante definition of moral hazard. In this memo, I reserve "moral hazard" for policyholder behavior only and treat the other channels as insurance-induced distortions. Parsons's taxonomy remains useful as a checklist of negative effects that a mandatory insurance regime can trigger and that policymakers should anticipate in design.

This liability-insurance lens is directly relevant to AI. In many realistic scenarios, the immediate deployer—or even the developing lab—will be thinly capitalized relative to potential harms and thus effectively judgment-proof. Insurance becomes the practical route to compensation and loss-spreading; but precisely because coverage supplies a solvent payor, it reshapes incentives across the system. Unlike first-party insurance, third-party liability coverage also induces claimant responses (in litigation strategy and settlement behavior) and can shift the legal environment itself (through evolving standards, evidentiary burdens, and insurability constraints). These dynamics are likely to be pronounced for AI, where harms may be diffuse, delayed, and hard to verify ex-ante.

A. Policyholders

[1] Christopher Parsons, Moral Hazard in Liability Insurance, 28 *Geneva Papers on Risk & Insurance—Issues & Practice* 448 (2003), https://doi.org/10.1111/1468-0440.00236.
[2] *Id*.

Policyholder moral hazard is the classic ex ante effect: when liability insurance shares losses, the insured's marginal cost of harm falls, and care can drop or shift.[3] In the AI context, this shows up before incidents as effort reduction (shallower evaluations, red-teaming limited to obvious prompt attacks, weak production instrumentation for lineage/rollback, and thin telemetry) and risk substitution toward controls that are easy to display but less connected to actual hazard.[4] The same loss-sharing can also tilt project selection toward higher-variance launches. Labs may release more capable models, greater autonomy, riskier deployment contexts, or use faster release schedules because more of the downside is externalized.

There is also a reputational substitution effect. Ben-Shahar & Logue frame this as 'outsourcing regulation': the presence (or absence) of favorable insurance terms operates as a private certification of safety practices—but only to the extent the insurer actually underwrites and monitors those practices.[5] A policy can be marketed as if it were evidence of safety; insurance becomes a proxy for due care rather than a complement to it. After incidents, the same dynamic reduces urgency to make patches, issue a recall, or gate access; firms may "monitor and message" longer than they should because the marginal costs of delay are shared.

Defense-within-limits (DWL) policies—where defense costs erode the same pot that pays judgments or settlements—can shape bargaining dynamics, but they do not change the baseline rule that the insured remains responsible for any amount above the remaining limit.[6] In practice, DWL often pulls toward earlier settlement when the insurer controls defense and faces bad-faith exposure, because every additional defense dollar both burns limits and increases the risk of refusing a reasonable within-limits deal. By contrast, in forms that give the insured consent-to-settle rights (often paired with a hammer clause), a defendant with strong reputational or precedent concerns—plausibly acute in early AI cases—may be more inclined to litigate longer despite eroding limits. That said, even in AI, this prolongation effect is context-dependent and typically tempered by the insurer's settlement incentives, the hammer's cost-sharing, and any excess-layer or collectability considerations.

Additionally, once coverage is in place, firms might choose expensive forms of remediation—broad settlements, sweeping PR campaigns, voluntary restitution packages—that are not strictly necessary to restore victims but that increase pool losses.[7] However, note that

---

[3] Id.
[4] Nat'l Inst. of Standards & Tech., *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* 26–34 (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf (calling for independent evaluation, red-teaming, and operational telemetry as part of ongoing measurement).
[5] Omri Ben-Shahar & Kyle D. Logue, Outsourcing Regulation: How Insurance Reduces Moral Hazard, 111 MICH. L. REV. 197 (2012).
[6] Int'l Risk Mgmt. Inst., Inc. (IRMI), *Defense Within Limits*, IRMI Insurance Definitions, https://www.irmi.com/term/insurance-definitions/defense-within-limits (last visited Sept. 23, 2025).
[7] Tom Baker & Sean J. Griffith, *Ensuring Corporate Misconduct: How Liability Insurance Undermines Shareholder Litigation* 3–8, 182–86 (U. Chi. Press 2010) (how coverage can blunt deterrence and shape corporate responses).

insurers likely would not allow continued exorbitant remediation once detected. In all, Parsons' label here is simple: the insured's behavior changes because insurance exists.

B. Claimants' Behavioral Change ("claimant hazard")

Parson also argues that claimant behavior may change in response to visible coverage. The welfare effect of claimants' changed behavior can be positive (access to compensation, injunctive leverage) or negative (nuisance pressure, remote defendants). When a payer is visibly solvent, more suits become economically viable, classes are easier to organize, and forum and theory selection adapt to target the deepest pockets. In AI disputes, this can rationally push claims upstream to foundation-model developers and hyperscale providers where app-layer integrators or small deployers are judgment-proof. That shift can improve access to compensation and concentrate litigation where discovery and injunctive leverage are most effective; it can also increase nuisance pressure and draw in remote defendants. Empirically, plaintiffs and their lawyers do screen on solvency: RAND's 'deep-pocket' analysis documents that injured parties and attorneys are more likely to bring or accept cases against 'deep-pocket' targets.[8] In securities litigation, mandatory disclosure of D&O insurance premiums led to a higher *dismissal* rate (with fewer but larger wins), consistent with plaintiffs selecting into cases based on visible insurance signals—i.e., more weak cases when coverage is salient.[9] Across multiple domains, large shares of filed claims appear weak: RAND reviews in medical malpractice and auto PI found roughly ½–⅔ of medical malpractice claims lacked evidence of negligence and ~42–50% of minor auto-injury claims were nonmeritorious or exaggerated.[10] The signal, in short, is ambiguous: increased suits can reflect both improved access to redress and a higher share of low-merit filings.

Parsons's core intuition on claimant behavior seems correct: claimants are strategic and will redirect effort toward insured targets when coverage is present. The policy-relevant point here is descriptive rather than condemnatory: claimant adaptation can simultaneously expand access to compensation and increase settlement pressure from low-merit suits. Whether the net effect is harmful depends on contemporaneous doctrine, contract terms, and the distribution of solvency in the market.[11]

C. Courts and Doctrinal Shift ("Jurisprudential/legislative hazard")

---

[8] Robert J. MacCoun, *Is There a Deep-Pocket Bias in the Tort System? The Concern over Biases Against Deep-Pocket Defendants*, RAND Corporation Issue Paper IP-130 (1993), https://www.rand.org/pubs/issue_papers/IP130.html.

[9] Dain C. Donelson, Justin Hopkins & Christopher G. Yust, *The Cost of Disclosure Regulation: Evidence from D&O Insurance and Nonmeritorious Securities Litigation*, Mays Business School Research Paper No. 3112661 (Feb. 6, 2018), https://ssrn.com/abstract=3112661.

[10] Carole Roan Gresenz, Deborah R. Hensler, David M. Studdert, Bonnie Dombey-Moore & Nicholas M. Pace, *A Flood of Litigation? Predicting the Consequences of Changing Legal Remedies Available to ERISA Beneficiaries*, RAND Issue Paper IP-184 (1999), https://www.rand.org/pubs/issue_papers/IP184.html.

[11] David J. Nye & Donald G. Gifford, *The Myth of the Liability Insurance Claims Explosion: An Empirical Rebuttal*, 41 Vand. L. Rev. 909 (1988), https://scholarship.law.vanderbilt.edu/vlr/vol41/iss5/2/.

In Parson's paper, jurisprudential and legislative hazard captures the idea that the availability of insurance can make courts and legislatures more willing to recognize new duties, relax proof rules, or broaden recoverable heads of loss in order to spread harm. Since this does not clearly fall under the category of moral hazard, I describe it as "doctrinal shift." A long line of evidence rules and doctrine implicitly assumes that knowledge of insurance can skew fact-finding—hence Federal Rule of Evidence 411's ban on using liability insurance to prove negligence—though the best mock-jury evidence suggests jurors often do discuss insurance and that discussion by itself has, at most, a weak or inconsistent effect on award size.[12] In AI, this can manifest as the incremental judicial recognition of pre-deployment evaluation duties, recall or warn obligations when defects are discovered, and ongoing monitoring duties for models with evolving behavior; as evidentiary presumptions or burden-shifting in black-box contexts where internal states are opaque; as greater willingness to entertain large-scale pure economic loss or reputational harms from generative systems; and as an inclination to run claims upstream on joint-and-several or enterprise-liability theories when small deployers are judgment-proof. Courts have, in other contexts with serious causation opacity, relaxed proof burdens to ensure compensation—most famously adopting market-share liability in DES cases—which functionally shifts causation risk from victims to insured enterprises; that dynamic is a plausible template for AI when harms are diffuse and attribution is hard.[13]

Parsons' claim is descriptive: insurability can and does influence legal development. Experimental and archival studies on juries, however, are mixed. Some experiments and practitioner summaries report that mentioning insurance can bias liability or damages upward, while other controlled deliberation studies find no statistically reliable link between insurance discussion and award size; credible, judge-specific causal evidence is thin, but leading tort/insurance scholarship documents how insurance availability and enterprise-level coverage have historically coincided with expansions in duty and recoverable heads of loss.[14] In lines where causation is hard to observe and harms aggregate slowly, this feedback increases uncertainty, broadens exposure, and undermines pricing discipline.

D. Insurer Behavior ("underwriting hazard")

Parsons's framework concludes with underwriting hazard: behavior changes or negative impacts that originate on the insurer's side, not the policyholder's. The core idea is simple: when insurers classify, price, and word policies under uncertainty, their choices can distort who buys insurance, how much capacity the market supplies, and what safety signals prices send.[15] Two

---

[12] Fed. R. Evid. 411; Edith Greene, Kathryn Hayman & Matt Motyl, "Shouldn't We Consider…?": Jury Discussions of Forbidden Topics and Effects on Damage Awards, 14 Psychol. Pub. Pol'y & L. 194 (2008), https://doi.org/10.1037/a0013486.

[13] *Sindell v. Abbott Labs*., 607 P.2d 924 (Cal. 1980).; *Hymowitz v. Eli Lilly & Co*., 539 N.E.2d 1069 (N.Y. 1989).

[14] Susan A. Row, *Admissibility of Insurance Policy Limits*, 45 La. L. Rev. (1985), https://digitalcommons.law.lsu.edu/lalrev/vol45/iss6/9/; Kenneth S. Abraham, *The Liability Century: Insurance and Tort Law from the Progressive Era to 9/11* (Harvard Univ. Press 2008).

[15] Parsons, *supra* note 1.

key ways in which insurers may frustrate the intended purpose of insurance is through risk misclassification and cross-subsidy. Risk misclassification is when the insurer puts firms into buckets that don't match their true hazard—often because the insurer lacks granular, verifiable information. For example, two frontier labs might both be priced as "Tier A," even though one runs rigorous evaluations and staged rollouts (lower risk) and the other does not (higher risk). Cross-subsidy is the consequence: if both pay about the same premium, the safer firm overpays and the riskier firm underpays; the former is quietly subsidizing the latter's expected losses. These errors are largely products of information asymmetry and non-stationary risk. Insurers have incentives to do real due diligence, but in AI that is harder: development cycles are fast, models and dependencies change quickly, validated loss data are scarce, and many critical practices (eval quality, rollback capability, vendor stack) are opaque ex ante. Even capable underwriters default to coarse proxies (revenue, sector, limit band, basic control checklists) and standardized terms to quote at speed and satisfy reinsurance comparability.[16]

A statutory requirement that frontier model developers carry liability coverage would expand the pool to include firms that previously opted out because they were judgment-proof, were priced out, or preferred to self-insure operationally. At launch, markets would likely rely on simplified buckets and templated wording, exactly when precision is most needed, which increases misclassification and cross-subsidy. In the short run, that tilts the insured pool toward higher-hazard exposures (the ones that most benefit from subsidized pricing), making pricing more volatile. In the medium run, however, mandates can interact with policyholder hazard in a constructive way: if risky activities become uninsurable or too expensive under clarified wording and tighter caps, some policyholders will scale back or de-risk projects they cannot afford to cover (see your policyholder-hazard section). The net effect depends on how quickly classification improves and wording catches up to actual perils.

Even if every account were perfectly classified, AI portfolios face common-mode (correlated) exposures that traditional per-account pricing misses. In practice, common-mode exposures in AI arise from shared infrastructure and components—tokenizers and inference libraries, CUDA/driver stacks, serving frameworks, data pipelines and third-party datasets, as well as common cloud regions and upstream vendors.[17] Within a given provider, thousands of customers may be on the same API model/version, so a single regression can synchronize losses across that provider's book.[18] (With open-source base models, multiple deployers may literally share weights; with proprietary labs, they do not.)[19] Recent software-supply episodes (e.g., a

[16] Steven Shavell, *On Moral Hazard and Insurance*, 93 Q.J. Econ. 541 (1979).
[17] *See* Hugging Face, *Summary of the tokenizers* (docs), https://huggingface.co/docs/transformers/en/tokenizer_summary (last visited Sept. 18, 2025); NVIDIA, *TensorRT-LLM Documentation* (docs), https://docs.nvidia.com/tensorrt-llm/ (last visited Sept. 18, 2025); Sebastian Moss, *AWS us-east-1 outage brings down services around the world*, DataCenterDynamics (Dec. 7, 2021), https://www.datacenterdynamics.com/en/news/aws-us-east-1-outage-brings-down-services-around-the-world/.
[18] OpenAI, *API Platform* ("Three million developers build with OpenAI"), https://openai.com/api/ (last visited Sept. 18, 2025).
[19] Meta AI, *Introducing Meta Llama 3* (Apr. 18, 2024), https://ai.meta.com/blog/meta-llama-3/.

widely pushed update) show how one change can trigger industry-wide claims, underscoring the difficulty of pricing correlated technology risks.[20] Legacy cyber/tech forms may also silently capture AI losses they were never priced for, owing to broad or outdated language. Further, when privacy or copyright doctrine shifts, yesterday's training or processing can become the basis for today's suits, reaching back into policy years that did not contemplate such claims.

II. Modern Insurance Levers

Modern insurance has well-tested tools for mitigating moral hazard without abandoning the benefits of risk-spreading and mitigating judgment-proofness. Three groups of mechanisms work on realigning incentives: (1) contractual cost-sharing, (2) risk-based pricing, and (3) direct oversight.[21] Cost-sharing restores immediate financial consequences to the insured's choices, pricing transmits a longer-horizon signal that rewards precaution and penalizes risk, and oversight makes promised care observable and enforceable. Used together, they narrow the gap between how a party behaves when uninsured and how they behave once a policy is in place.[22]

a. How they work

Contractual cost-sharing reduces moral hazard by ensuring the insured bears a meaningful slice of any loss. Deductibles and self-insured retentions move the "first dollars" of harm back onto the policyholder, which discourages small, avoidable losses and promotes routine care and maintenance. Coinsurance keeps alignment in place even after a threshold is crossed, cutting down on "spend freely because insurance pays" dynamics in the midst of remediation. Per-occurrence and aggregate limits, if calibrated, deter over-consumption of post-loss services and encourage prioritization when losses accumulate. The point is not to punish loss, but to make the insured face enough marginal cost that prevention (and quick mitigation) remains the cheaper choice. Poorly set cost-sharing can overshoot—inviting underinsurance or deferred reporting—so insurers and regulators tune these levers to preserve access while keeping incentives sharp.[23]

Risk-based pricing curbs moral hazard by turning better risk management into lower, recurring costs. Underwriting differentiates exposures ex ante, while experience rating ties future premiums to actual loss history. Together they translate to safer behavior—maintenance, training,

---

[20] Jeffrey Dastin, *Microsoft says about 8.5 million of its devices affected by CrowdStrike-related outage*, Reuters (July 20, 2024), https://www.reuters.com/technology/microsoft-says-about-85-million-its-devices-affected-by-crowdstrike-related-2024-07-20/; Manya Saini & Zeba Siddiqui, *Insured losses from CrowdStrike outage could reach $1.5 bln, CyberCube says*, Reuters (July 25, 2024), https://www.reuters.com/business/finance/insured-losses-crowdstrike-outage-could-reach-15-bln-cybercube-says-2024-07-25/.

[21] Gabriel Weil, Overcoming Judgment-Proofness: The Law & Economics of Insuring and Mitigating AI Risk (manuscript draft, Sept. 2025) (on file with author).

[22] Id.

[23] Id.

controls, compliant operations—into tangible price credits, and translate risky behavior into surcharges. Because premiums are revisited each term, pricing carries consequences forward in time, counteracting the short-term temptation to cut corners. Credibility-weighted experience rating, schedule credits/debits, and class relativities all serve the same function: send reliable, actuarially grounded signals that make precaution economically rational. The discipline operates even when no loss occurs—policyholders invest in care because they expect a premium break—and it persists after a loss because next year's renewal will reflect how well they managed it.[24]

Direct oversight provisions reduce moral hazard by converting promises of care into enforceable obligations. Coverage conditions and warranties (maintenance, inspections, safety programs), audit and inspection rights, cooperation clauses, and approved-vendor or service-level requirements let insurers verify that loss-control measures are real and ongoing. This shrinks the information gap that otherwise invites shirking: if a training requirement or monitoring protocol is part of the coverage grant, the insured has reason to comply continuously, not just at renewal. Well-designed policies use proportionate remedies—premium adjustments, step-downs, or higher retentions for non-compliance—so that oversight changes behavior without turning every technical misstep into a forfeiture. The effect is to keep preventive effort visible, verifiable, and tied to concrete coverage consequences.[25]

Individually, each tool restrains a different expression of moral hazard; together, they are complementary. Cost-sharing acts at the moment of loss, pricing shapes behavior over the policy's life, and oversight ensures the facts underlying both are true. That is how modern insurance spreads risk while still preserving strong incentives to avoid, reduce, and swiftly remediate harm.[26]

b. How the levers mitigate

Modern liability insurance does not merely spread losses; when designed with (i) contractual cost-sharing, (ii) risk-based pricing (including experience rating), and (iii) direct oversight (warranties, audit/inspection, cooperation and service-level obligations), it counteracts the very distortions that mandates risk creating.[27]

i. Policyholder hazard

Cost-sharing, risk-based pricing, and direct oversight all act directly on the classic ex-ante moral-hazard channel. Deductibles, self-insured retentions, coinsurance, and calibrated occurrence/aggregate limits restore meaningful marginal cost to the insured's decisions, which

[24] Id.
[25] Id.
[26] Id.
[27] Id.

reduces the incentive to under-invest in prevention or to over-consume post-loss services; this is the canonical prediction of insurance economics and remains robust across models.[28] Experience rating and schedule credits/debits then extend that discipline over time by tying next term's premium to observed loss experience and verifiable controls, turning better red-teaming, staged rollouts, rollback/kill-switch capabilities, and incident management into recurring price advantages rather than one-off marketing claims.[29] Finally, direct oversight provisions—conditions, warranties, audit/inspection rights, and cooperation clauses—convert safety promises into enforceable obligations and shrink the information gap that invites shirking; insurers can verify that evaluation and monitoring practices exist and persist, not just that they were promised at bind.[30] In combination, the three levers make it economically rational for insured labs to keep up meaningful evaluations, instrumentation, and rollback capacity, and they raise the opportunity cost of "shipping fast" without controls.

ii. Claimant dynamics & litigation selection

Cost-sharing and experience-rated pricing indirectly dampen nuisance pressure by reducing the attractiveness of "pay-small-claims" equilibria. When insureds face retentions and know that clustered small settlements will raise renewal pricing, they have sharper incentives to resist weak claims and to invest in the records and telemetry that allow quick merits screening; that, in turn, can shift plaintiffs toward higher-merit cases.[31] Oversight also helps here: cooperation and documentation clauses (and the insurer's own early-case assessment) produce standardized artifacts—protocols, logs, and incident timelines—that enable faster, more accurate claim sorting and narrow discovery to the core issues, which reduces the payoff to low-merit filings.[32] None of this eliminates claimant adaptation, but it raises the relative return to meritorious suits.

iii. Doctrinal shift

The three levers cannot—and need not—dictate doctrine, but they can reduce the conditions under which courts feel pressure to expand duties "because insurance exists." Better prevention and faster mitigation (induced by cost-sharing and pricing) lower the salience of

[28] See Steven Shavell, *On Moral Hazard and Insurance*, 93 Q.J. Econ. 541, 541–49 (1979) (deductibles/coinsurance reduce ex-ante moral hazard).
[29] See id.; see also Nat'l Inst. of Standards & Tech., *Artificial Intelligence Risk Mgmt. Framework (AI RMF 1.0)* 26–34 (Jan. 2023), https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf (tying measurement, testing, and monitoring to ongoing governance—practices insurers can observe and price).
[30] See, e.g., 4 Jeffrey E. Thomas & Francis J. Mootz III, *New Appleman on Insurance Law Library Edition* § 31.05 (LexisNexis 2025) (warranties/conditions and cooperation clauses as enforceable oversight tools) (treatise support for standard insurer oversight provisions).
[31] See Tom Baker & Sean J. Griffith, *Ensuring Corporate Misconduct: How Liability Insurance Undermines Shareholder Litigation* 182–86 (U. Chi. Press 2010) (insurance structure and settlement dynamics); cf. Omri Ben-Shahar & Kyle D. Logue, *Outsourcing Regulation: How Insurance Reduces Moral Hazard*, 111 Mich. L. Rev. 197, 215–24 (2012) (insurer monitoring and private certification improve safety incentives).
[32] See New Appleman, supra note 3, § 32.02 (insurer investigation/cooperation and claim-handling duties); Fed. R. Evid. 408 (settlement communications) (structuring early merits evaluation and resolution).

catastrophic uncompensated losses; better records and replicable testing (induced by oversight) give judges more concrete bases to resolve causation and standard-of-care questions without resorting to broad presumptions or burden-shifting. As a result, the evidentiary posture is less likely to devolve into "black box" contests in which the mere presence of coverage risks becoming the anchor, a dynamic that underlies Federal Rule of Evidence 411's general exclusion of liability-insurance evidence to prove negligence.[33]

iv. Underwriting errors & capacity risk

The same levers also mitigate insurer-side problems. Oversight and disclosure duties (e.g., dependency statements for model/version/stack; evidence of evaluations and rollback capability) improve the classification inputs that underwriters need, which reduces misclassification. Pricing that explicitly credits demonstrable controls and loads for opaque, highly correlated stacks (rather than relying on revenue or limit band alone) shrinks cross-subsidies within coarse buckets. Cost-sharing can also be made to reduce accumulation—e.g., higher retentions or sublimits for specific common-mode perils—so that correlated losses do not turn into an unpriced, pool-wide subsidy. In short, the levers give insurers better measurement, sharper price signals, and more tools to align capacity with actual hazard.

III. What the three levers don't fully fix

The three levers—cost-sharing, risk-based pricing, and direct oversight—substantially narrow the gaps created by a mandatory AI liability regime, but they do not eliminate them. Some limits are structural: information about true practices and dependencies remains imperfect and quickly out of date; hazards evolve faster than historical data can stabilize prices; and losses can arrive in correlated waves when many firms share the same upstream stack, overwhelming even well-tuned rating plans. Other limits are institutional: the presence of insurance will continue to attract some marginal claims; courts and regulators may still broaden duties when attribution is opaque; and contract mechanics that discipline behavior in ordinary cases can blunt incentives when reputational stakes are high or when shocks are systemic. What follows explains, for each of the four channels—policyholder behavior, claimant behavior, doctrinal shift, and underwriting—where the levers bite and where residual risk endures despite them.

A. Policyholder

Even with cost-sharing and experience rating, some insureds will still take too little care when stakes are reputational or strategic, especially if they believe correlated failures will trigger broad market assistance or political intervention. Deductibles that are too low relative to event size, or priced credits that fail to distinguish between "checklist" and substantive controls, can

[33] Fed. R. Evid. 411.

leave the basic moral-hazard problem intact. And because AI risks are non-stationary, a lab can earn credits for controls that later prove weak against new failure modes—a reminder that pricing must be updated frequently to track genuine loss reduction rather than simple checklist compliance.[34]

B. Claimant

The presence of a solvent insurer might still attract some claims that would not be filed against judgment-proof defendants, and plaintiffs' counsel will continue to screen on insurance signals when allocating effort.[35] Cost-sharing does not bar low-merit suits; it only changes the insured's settlement posture at the margin. Likewise, oversight-generated artifacts can be used both to exonerate and to target upstream actors, so improved documentation does not guarantee fewer suits—it improves sorting, not necessarily volume. And because ex post loss sharing remains, defendants may still pursue broad, reputationally motivated remediation that increases pool losses even when narrower remedies would compensate victims—albeit subject to insurer pushback.[36]

C. Courts

Courts and legislatures may still expand recognized duties or adjust proof rules when harms are diffuse or attribution is genuinely opaque; the existence of coverage remains a background condition that can make such expansions more palatable. Experimental evidence on whether insurance references bias jurors is mixed, and there is no strong basis to assume that better oversight alone will arrest doctrinal drift in fast-moving technical fields.[37] Moreover, to the extent that improved records make failures more legible, they can also enable new duties (e.g., to patch, recall, or warn) once capabilities for post-release control are documented—reducing uncertainty but not necessarily shrinking exposure.

D. Insurer

The hardest problems—information asymmetry, non-stationary risk, and correlation—do not disappear. Even with better disclosures and audits, underwriters will often start with coarse proxies because validated loss data are scarce and development cycles are rapid. Pricing will remain volatile until enough experience accumulates; in the interim, safer firms will still subsidize some riskier peers within buckets, and mandates can amplify that misallocation by

[34] See Steven Shavell, *On Moral Hazard and Insurance*, 93 Q.J. Econ. 541, 541–49 (1979) (deductibles/coinsurance reduce ex-ante moral hazard).
[35] See Baker & Griffith, supra note 4, at 3–8 (insurance as a target signal in corporate litigation); see also Stephen J. Carroll et al., *The Institute for Civil Justice: RAND Studies of Civil Justice* 31–36 (RAND 1991) (screening and settlement patterns).
[36] See Baker & Griffith, supra note 4, at 3–8 (insurance as a target signal in corporate litigation); see also Stephen J. Carroll et al., *The Institute for Civil Justice: RAND Studies of Civil Justice* 31–36 (RAND 1991) (screening and settlement patterns).
[37] Fed. R. Evid. 411.

forcing rapid, standardized placements before classification catches up. Correlation is the structural limiter: shared upstreams (model/version, libraries, data pipelines, cloud regions) mean a single regression or vendor update can generate synchronized claims across a portfolio even when each account is "correctly" priced; the 2024 global outage tied to a widely pushed software update is a vivid reminder that accumulation can overwhelm otherwise sound rating plans.[38] Accumulation caps, clash stress tests, and per-peril sublimits can contain, but not eliminate, this residual risk; the market will still "snap back" (prices, exclusions, capacity retrenchment) after true correlation is revealed.

IV. Bottom Line (why insurance still helps despite moral hazards)

Even with real gaps, a well-designed liability-insurance mandate improves incentives relative to no liability when labs are judgment-proof. It (1) prices risk where courts can't collect: premiums and capacity embed forward-looking assessments of frequency/severity that tort alone can't enforce against thin balance sheets; (2) conditions coverage on verifiable precautions: warranties, service-level commitments, and audit rights convert safety promises into obligations with coverage consequences; (3) restores skin-in-the-game through collateralized deductibles/retentions and calibrated coinsurance so the insured bears meaningful first-loss and shared loss at the margin; (4) generates usable telemetry—standardized test logs, incident reports, and replication artifacts—that regulators and courts can rely on instead of inference and opacity; and (5) protects the private market by reserving a narrow, capped public backstop for truly systemic correlation risk, keeping ordinary losses in private contracts while preserving solvency for catastrophe layers no firm can retain. In Appendix A, I have included a potential design checklist for policymakers to consider when creating an insurance mandate.

Appendix A: Design Checklist for Policymakers

A mandate works only if it can be operationalized without freezing the market. This section does two things. First, the paragraphs lay out a range of policy options—what they are, how they work, who should implement which parts (legislature/regulator versus private insurers), and the concrete pros and cons. These are meant to help policymakers understand the tradeoffs and pick the right tool for their context rather than treat any single mechanism as the perfect solution. Second, the table serves as a quick-check list of the measures I judge most likely to improve safety with minimal downside.

1) Per-Model-Release Retentions (with a clear release definition, small caps for cascades, and collateral where needed)

[38] *Insured losses from CrowdStrike outage could reach $1.5 bln*, Reuters (July 25, 2024), https://www.reuters.com/business/finance/insured-losses-crowdstrike-outage-could-reach-15-bln-cybercube-says-2024-07-25.

Labs should bear the first dollars of loss tied to each discrete model release or safety update before insurance responds. Anchoring the retention at the release level keeps incentives where they matter most: gating changes, staging rollouts, maintaining rollback readiness, and enforcing change control. It discourages preventable mishaps and delays in reporting because every new release that goes wrong immediately hits the lab's own balance sheet. To avoid turning startups into de facto judgment-proof actors on first losses, the retention should be calibrated to financial capacity and, where appropriate, backed by collateral such as a letter of credit, escrow, or a parent guarantee. Because one defective release can generate many related claims, it is important to define the "release event" by defect lineage and time window, and to include a small cap on how many retentions can attach to a single cascade. Legislators and regulators are best placed to set the narrow floor—requiring retentions at the per-model-release level, describing what counts as a release event, and permitting collateral—so there is no race to the bottom across carriers. Insurers should size the retention by risk tier, set the small caps, and spell out the collateral mechanics in the policy, which keeps the tool flexible and avoids one-size-fits-all mistakes. The public floor promotes consistency and comparability; private calibration preserves precision and practicality.

2) Coinsurance as a co-equal discipline after the retention (including defense and indemnity, with a short emergency grace)

Coinsurance should be used alongside retentions as a primary incentive. Retentions act immediately at the time of loss. Coinsurance keeps the insured economically engaged after the retention is exhausted. A modest share (perhaps ten to twenty percent) that also applies to defense and indemnity encourages the lab to manage vendor scope, avoid over-litigation, and pursue commercially sensible settlements. It reduces the tendency to "spend freely because insurance pays" once the retention is gone, and it aligns choices about the breadth of notifications, the depth of forensic work, and settlement timing with cost-effective risk reduction. To ensure coinsurance never slows urgent containment, policies should include a short grace window—typically the first forty-eight to seventy-two hours after detection—or a small containment threshold immediately above the retention. During that window or threshold, essential rollback, stabilizing forensics, and legally required notifications are not coinsured; coinsurance kicks in thereafter, including on defense and settlement. Legislators and regulators should set these guardrails in law and require plain-language disclosure so insureds understand how coinsurance works in an emergency. Insurers should then tune the percentage, the triggers, and the specific mechanics by risk class. Public guardrails protect speed of response; private tuning preserves flexibility across operating models and claim severities.

3) Minimal, updateable reporting schema and standardized safety telemetry with limited independent re-execution

A small, living set of fields can make pricing and audits credible without turning compliance into paperwork theater. The schema should cover evaluation coverage and thresholds, the depth of red-teaming, the quality of change management, the dependency stack and versions in use, incident and near-miss counts, time to detect and time to repair, and evidence of rollback testing. Fields should be refreshable each year so the regime keeps pace with technical change. To keep the numbers honest, there should be limited independent re-execution of a sample of tests. To protect sensitive information, labs can publish plain-English summaries while full artifacts are maintained in regulator-only data rooms with least-privilege access and audit trails. Legislators should mandate the existence of a common schema and delegate maintenance to a technical regulator that can update it annually. Regulators should run the data rooms and set safe-handling rules for trade secrets and security details. Insurers should consume the schema for pricing and may add their own supplemental fields privately, which lets them innovate without undermining comparability. The public standardization yields portability and better price signals; the private additions allow carriers to compete on insight rather than format.

4) Coverage conditioned on verifiable controls, with proportional remedies and cure periods

Insurance should reinforce, not merely observe, safety practice. Policies can do this by conditioning coverage on a handful of verifiable controls: an active safety-and-security protocol, replicable test logs, incident-response service levels, and documented change control for major releases. Enforcement needs to be fair and workable. Minor lapses should lead to measured consequences—such as a surcharge, a step-down in limits, or a higher retention for that claim—while outright denial should be reserved for material breach or knowing misrepresentation. There should be reasonable cure periods so good-faith actors can fix issues, and verification can rely on sampling rather than exhaustive line-by-line audits. Legislators and regulators should codify proportionality, materiality, and cure, and should authorize third-party audits or sampling so verification is feasible at scale. Insurers should draft the specific conditions and warranties, build the remedy ladder into policy language, and select or retain the auditors. Public guardrails protect fairness and prevent hair-trigger forfeitures; private implementation keeps the mechanism nimble and tailored to the control environment.

5) Accumulation management for systemic and correlated risk

Insurers and supervisors need visibility into shared dependencies that can synchronize losses across many policyholders. Labs should disclose, in protected settings, the model family and version they are using, key libraries and frameworks, critical datasets, cloud regions and availability zones, major vendors, and the lineage of safety updates. With that information, carriers can run portfolio stress tests, set narrow sublimits for specifically defined systemic perils, use per-version aggregates where appropriate, and encourage staggered releases or diversification when exposures are overly concentrated. Because these details are commercially

sensitive, they should be handled in regulator or insurer data rooms under clear use limits and with audit trails, and not be used for competitive purposes. Legislators and regulators should require standardized dependency disclosure and periodic stress testing, and should set the handling rules for confidentiality. Insurers should apply the outputs—setting accumulation caps and sublimits and adjusting pricing and capacity—and can advise insureds on diversification strategies. Public requirements ensure the data exist and are protected; private use of the data avoids blunt exclusions and supports targeted risk control.

6) Public backstop for catastrophic correlation that is high-attachment and conditional

A public backstop should be used sparingly and only as a last resort. It should attach well above private layers and trigger only for narrowly defined systemic AI incidents, such as a widely deployed model version that causes synchronized, industry-wide losses. To preserve incentives, the backstop should include coinsurance at the top layer, require participants to meet baseline controls to be eligible, fund itself through experience-rated assessments, and condition payout on independent root-cause analysis and time-bound remediation with a redacted public summary. Designed this way, a backstop does more than pay claims; it keeps private insurers in the market after a catastrophe, which preserves ongoing underwriting discipline, and it turns disasters into structured learning and hardening. If triggers are too broad or attachment is too low, however, a backstop can dull prevention and crowd out private capacity. Legislators need to enact the program and fix its triggers, attachment point, and governance. Regulators should administer eligibility, data sharing, and post-event oversight. Insurers remain the front-line risk managers for ordinary losses and interface with the backstop only in rare systemic events. Public design is essential to avoid moral hazard; private markets remain primary so day-to-day pricing and conditions continue to push for safety.

Recommendation	Primary implementer	Notes on role split
Per-model-release retentions with small caps for cascades and collateral for thin balance sheets	**Shared:** Legislature/Regulator set floors and define "release event"; **Insurers** calibrate size, caps, and collateral terms	First dollars at risk on each release reliably deter sloppy changes and delayed rollback; public floors prevent under-deterrence, private calibration preserves precision.
Minimal, updateable reporting schema and standardized safety telemetry with limited independent re-execution	**Shared:** Legislature mandates schema; **Regulator** maintains fields and operates data rooms; **Insurers** may add supplemental fields for underwriting	Comparable, auditable inputs create strong price signals for real safety work and reduce checklist gaming; annual updates keep it current.

Coverage conditioned on verifiable controls, with proportional remedies and cure periods	**Shared:** Legislature/Regulator codify proportionality, materiality, and cure; **Insurers** draft conditions, set the remedy ladder, and run sampling/audits	Turns safety promises into enforceable practice without hair-trigger denials; sustained day-to-day compliance meaningfully lowers incident risk.
Regulator-only data-room handling and confidentiality guardrails for audits and submissions	**Shared:** Regulator sets least-privilege access, submitter-notice, audit trails; **Insurers/Developers** comply and structure artifacts accordingly	Enables substantive audits and verification (which improve safety) while minimizing leakage, making robust evidence sharing feasible in practice.