NBER WORKING PAPER SERIES

OPTIMAL MEDICAL LIABILITY FOR AI

Alex Chan

Working Paper 35321
http://www.nber.org/papers/w35321

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2026, Revised July 2026

Special thanks to Michelle Mello and David Studdert for introducing me to the topic of medical liability, and to David Cutler for helpful discussions that motivated this work. Funding was provided by the Stanford Institute for Economic Policy Research and Harvard Business School AI Institute. Valuable research assistance was provided by Daniel Kornbluth. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Optimal Medical Liability for AI
Alex Chan
NBER Working Paper No. 35321
June 2026, Revised July 2026
JEL No. D47, D82, D86, D9, G22, I1, I13, I18, K12, K13, L51

ABSTRACT

I study medical liability when artificial intelligence acts as a doctor rather than as a passive clinical tool. The central object is the legally usable medical record: the inputs, logs, warnings, prescriptions, follow-up instructions, and outcomes on which courts, contracts, insurers, and regulators can condition responsibility. I show that AI medical liability is an institutional design problem under imperfect legal information. If the record separates AI-controllable error from patient nonadherence and natural disease progression, high-powered AI-fault liability implements the standard accident-law ideal. If the record is coarse, the first best may be infeasible: the same transfer that disciplines the AI also insures the patient's hidden action. With joint causation, the relevant object is a marginal-responsibility score rather than a posterior cause label. I characterize the feasible set of liability incentives generated by the record and show when the optimal rule is no liability, strict liability, negligence, a safe harbor, comparative fault, or a continuous warranty. I then study algorithmic defensive design, through which AI developers can design not only medical recommendations but also the record on which future liability depends. Adoption, learning, enterprise liability, insurance, no-fault compensation, and regulation enter as ways to change the record, the liable entity, or the financing of compensation. The framework yields conditional implications rather than a one-size-fits-all rule.

Alex Chan
Harvard University
Harvard Business School
and NBER
achan@hbs.edu

(Please click here for the most current version)

1 Introduction

AI medical liability starts with a tension. Chan, Xu, and Cutler (2026) show that patients may not trust an AI doctor unless someone stands behind the advice. But making someone stand behind the advice can distort the advice itself. An AI system can become safer in the clinically relevant sense, or it can become safer only in the legal sense: it can turn away difficult patients, escalate too often, order too many tests, write too many warnings, avoid useful prescriptions, or document compliance in ways that lower legal exposure without improving health. I study how liability should be designed when these responses are possible.

The dynamic part of the problem is especially important for AI doctors. Because patients value accountability, a stronger warranty can accelerate early adoption by signaling the AI firm's confidence in its system's quality. I formalize this channel using a signaling game in which firms privately observe quality and choose both warranty strength and record design, while patients update their beliefs. Quality improvements arrive only after use, through learning from clinical encounters. This timing creates a complementarity: credible liability can raise adoption today, adoption can generate clinically informative data, and those data can raise tomorrow's quality. But the same legal pressure can also induce exclusion or litigation-oriented documentation, so adoption produces little useful learning. The dynamic effect of liability is ambiguous. The same AI technology can support distinct stationary regimes: low quality, low trust, and low use, or high quality, credible accountability, and high use.[1]

In this paper, I present an institutional design model with relatively general implications.[2] Courts, private warranties, malpractice insurers, no-fault pools, health systems, and regulators differ in the information they can use and in the actors they can discipline. I do not start by asking whether AI should be treated as a product, a doctor, or a hospital employee. I start by asking what each institution can verify and which party controls the marginal risk. A central contribution is to represent medical liability as a contract on the legally usable record and to ask which incentives that record can implement.

[1] In the paper's two-regime example, primitive patient demand is smooth and quality dynamics are linear; the long-run jump comes from endogenous legal-regime selection and record jamming, not from a discontinuity in demand.

[2] I view institutional design as overlapping with market design, and in this paper, I use the two terms interchangeability.

The first best rule would make the AI side pay for AI-controllable harm while leaving patients exposed to the consequences of their own nonadherence or unusable inputs. But legal systems do not observe these objects directly. They observe a record. If the record is rich (or fine), liability can be targeted. If the record is coarse, the desired incentives may be infeasible. This is the mechanism that drives the paper; the later institutional examples are applications of it.

This perspective also distinguishes between two evidence margins. One margin is *pre-deployment* evidence: validation studies, audits, calibration exercises, prospective trials, and change-control plans that determine whether an AI system should enter a medical task at all. Another margin is *encounter-level* evidence: the symptoms elicited, patient inputs, warnings, escalation decisions, prescriptions, follow-up, and outcomes generated during care. Poggi and Strulovici (2020) study the first kind of problem in a general liability-design model with information acquisition before firms launch risky products. I take their tariff-on-evidence insight as a natural starting point. The medical-AI problem adds the second margin: after launch, evidence is jointly produced by the AI system, clinical workflow, and patient behavior, and the same record must discipline several actors.

The first distinction I make is between a liability-backed guarantee and a medical record. This is not meant to displace ordinary product-liability concepts such as design defect, warning defect, standard of care, or causation. Those doctrines still matter. The point is that an AI doctor makes the evidentiary problem unusually central because the system itself asks questions, gives warnings, issues recommendations, records patient responses, and can design the future litigation record. AI doctoring therefore asks a "harder" question than whether a product failed: what does the record show about why the patient was harmed? A patient may have omitted symptoms, ignored urgent instructions, failed to complete tests, or experienced natural progression of disease. A developer may have trained a bad model, failed to calibrate it for a patient population, used a dangerous escalation threshold, or designed the interface to look legally safe rather than medically safe. A clinical enterprise may or may not have verified inputs, followed up, prescribed, monitored, or escalated. The baseline case is patient-facing or autonomous AI advice; clinicians and hospitals enter through the enterprise and workflow extensions below. I take the verifiable medical record as the primitive object. A liability regime is a transfer written on facts that the relevant institution can use.

I present the model in three steps. I first assume separable latent causes, no algorithmic defensive-design response, no strategic manipulation of the record, and no learning externality. In this set-up, an injury has an AI-caused component, a patient-caused component,

and a natural-disease component. AI effort affects only the AI component; patient effort affects only the patient component. Under those assumptions attribution-based AI-fault liability is the clean first-best object: pay for the legally attributed AI-caused component of harm and do not pay for patient nonadherence or natural progression. This is not strict liability for every adverse outcome. Nor does it require the legal system to abandon negligence or design-defect doctrine. It is the economic target that those doctrines would approximate if the record identifies AI-controllable harm.

Separable-cause is useful because it gives a transparent reference point, not because it describes all of medicine. Many motivating clinical examples involve joint causation. Warning design affects adherence. Symptom elicitation affects what the patient reports and what the model can infer. Clinical workflows affect both patient behavior and model performance. In these environments there may be no meaningful partition of harm into mutually exclusive AI and patient causes. I therefore add a joint-causation result immediately after. I show that the general object is not posterior fault but marginal incentives: the transfer must co-vary with the score of the medical record in a way that prices the AI's marginal contribution to harm while preserving patient incentives. Posterior fault is a special case of this score-based contract, not the general theory.

Liability is a transfer from the liable AI side to the patient. Here, that transfer is useful because it makes the AI internalize the marginal harm it controls. But the same transfer is insurance to the patient. If the record is too coarse, any payment that varies with the AI-relevant part of the record also varies with the patient-effort part of the record. Then the first best is not merely hard to implement; it is outside the feasible moment set of bounded, nonnegative legal transfers. The score-based contract is what the law would like to price. Actual tort, warranty, insurance, and safe-harbor rules are second-best approximations constrained by the legal record, loss caps, no-negative-damages constraints, and claims costs.

Consider a patient harmed after a missed sepsis diagnosis. If the record only shows that the patient was injured, the legal system can compensate but cannot tell whether the AI missed a red flag, the patient omitted a symptom, the patient ignored an urgent-care instruction, or the disease progressed despite appropriate advice. If the record also shows what the AI asked, what the patient answered, which vitals were requested and supplied, what warning was given, and whether the patient followed up, liability can be targeted much more precisely. The record does not have to be perfect; it has to distinguish the facts that matter for incentives. This does not mean that the evidentiary record creates the legal rule after the fact. The rule is fixed ex ante. The record determines whether the

plaintiff, defendant, insurer, or regulator can prove the facts that the rule makes legally relevant. More documentation helps only if it separates AI-controllable risk from patient behavior and natural disease. Documentation that is merely formal or self-protective can make the record look richer without making liability better.

I then give compensation and trust a welfare foundation. The paper does not treat liability as only compensation. Liability can deter, insure, allocate losses, support trust and adoption, discipline enterprises through insurance and contracts, and affect learning. In the risk-neutral benchmark, transfers are not socially valuable by themselves. They matter through incentives and administrative cost. Compensation becomes socially valuable when patients are risk averse, liquidity constrained, or distributionally weighted; when no-fault compensation lowers administrative cost relative to litigation; or when accountability changes adoption and marginal adopters have positive social surplus. This distinction matters for interpreting survey and discrete-choice evidence on liability. Such evidence identifies willingness to pay for accountability and the demand response. It does not, by itself, identify the whole numerator of an optimal liability index. To turn willingness to pay into welfare, I need the surplus of marginal AI users, the cost of financing compensation, and any learning externality generated by adoption.

In the third step, I add what is distinctively AI. Developers can respond strategically through algorithmic defensive design. I use this term narrowly. Traditional physician defensive medicine[3] can also have ex ante components–practice protocols, documentation templates, referral norms, patient-selection rules, and specialty-level avoidance of risky cases. The distinction I emphasize is therefore not purely temporal. Algorithmic defensive design is embedded in the product and deployed at scale: rejection rules, warning thresholds, eligibility filters, referral defaults, prescription defaults, interface language, and litigation-oriented documentation. The same scale feature applies to true safety improvements: a single model-level change in calibration, escalation, prescribing logic, or symptom elicitation can reduce risk across many future patients, whereas ordinary physician care is largely encounter- or clinician-specific. These choices can change both clinical outcomes and the meaning of the formal record. Thus the record is sometimes an exogenous information technology and sometimes a *strategic object*.

The framework nests many current legal and regulatory proposals without ranking them in the abstract. Strict product liability, malpractice negligence, safe harbors, comparative fault, warranties, caps, no-fault pools, mandatory insurance, burden shifting,

[3]See, for example, Kessler and McClellan (1996), Mello et al. (2024), Mello et al. (2005), and Frakes and Gruber (2019).

common-enterprise liability, FDA-style premarket and change-control rules, EU-style logging and human-oversight duties, regulatory sandboxes, and restrictions on autonomous AI are not separate theories in the model. They are restrictions on the feasible transfer, the usable record, the covered cases, the liable entity, or the insurance technology.

I treat the institutional cases hierarchically. The core theory is the record-based bilateral moral-hazard problem: what can be implemented when the law can condition only on a particular sigma-field? The applications–enterprise assignment, market structure, prescribing, no-fault funds, and regulatory sandboxes–map current proposals into the primitives of the model. They are not meant to be separate general theories of liability.

The paper is organized to make this hierarchy clear. Sections 2 and 3 contain the core theory: implementation on a fixed record, joint causation, the zero-sum bottleneck, compensation, auditability, legal fact-finding, and the shape of the optimal contract. Section 4 adds AI-specific margins: defensive design, Bayesian learning, learning traps, scope, enterprise contracting, market structure, direct prescribing, and robustness. Section 5 links the framework to accident law, contract theory, malpractice, AI regulation, and the emerging adoption evidence. Section 6 then shows how current proposals are restrictions of the same market design problem, and Section 7 states the conditional design implications.

The framework also has implications outside medicine, although I do not want to overclaim. AI systems increasingly give consequential advice about finance, law, education, employment, and public benefits. These markets share several features with AI medicine: users supply noisy information, may or may not follow advice, firms can design disclaimers or eligibility filters, and accountability can raise adoption while changing incentives. The medical setting is plausibly the highest-stakes case because losses are severe, causation is medically complex, and users are often vulnerable. But the same market design problem appears more generally: liability and regulation should be written on the verifiable advice record, should distinguish AI/provider-controllable error from user-side information and compliance, and should decide when court-based liability is less effective than ex ante record and safety regulation. Recent work on generative AI regulation similarly emphasizes tailoring obligations to high-risk applications rather than treating the model alone as the regulatory object (Hacker, Engel, and Mauer, 2023). The connection to advice markets is especially close to Gennaioli, Shleifer, and Vishny (2015), who treat trust in expert intermediaries as part of the economic value of delegated advice. The institutional choice also echoes Glaeser, Johnson, and Shleifer (2001), Glaeser and Shleifer (2002), and Glaeser and Shleifer (2003): the right mix of private ordering, litigation, and regulation depends on information, enforcement, and institutional costs. A general theory of advice-

giving AI would require additional primitives, but the medical case clarifies the record, incentive, and assignment problems that other domains will also face.

A contribution of this paper is to make clear implications for market design: if AI-controllable harm is verifiable and patient moral hazard is limited, high-powered AI-fault liability is efficient. If proof is noisy and patient effort is important, partial or conditional liability is efficient. If compensation is valuable but individual causation is too costly to prove, no-fault compensation can be efficient only when financing, experience rating, or regulation supplies deterrence elsewhere. If clinical integration improves monitoring and attribution, patient-facing liability can be assigned to an enterprise, but only if internal contracts or direct liability preserve the developer's incentives. If integration mainly creates market power or excludes complex patients, the assignment is less attractive. If liability induces algorithmic legal shielding rather than clinical safety, optimal liability power falls unless the record technology is redesigned to make shielding unprofitable.

2 The Medical Record and the Fixed-Record benchmark

This section starts the cleanest liability problem: one case type, fixed legal record, no algorithmic defensive design, no insurer loading except where explicitly introduced, and no learning externality. Thus, this is a no-algorithmic-defensive-design model: the AI can reduce clinical risk, but it cannot lower liability by over-warning, over-referring, over-testing, over-prescribing, selecting away difficult patients, or generating a litigation record that is not clinically useful. I also begin with separable latent causes. I relax separability in Section 2.3.

The benchmark should be read as a local implementation problem on a fixed information partition. The first best objects are useful because they identify what the law would like to price. The feasibility result below then shows when bounded, nonnegative legal transfers cannot price those objects simultaneously. This distinction is important for readers who come from standard principal-agent theory: the posterior-fault rule is an ideal under rich information, not a claim that tort law can always implement first best.

2.1 Technology, first best, and the first-order approach

An adverse outcome is $Y \in \{0, 1\}$ and medical loss is $L > 0$. In the separable benchmark the probability of harm is

$$p(e, a) = p^0 + p^A(e) + p^P(a), \tag{2.1}$$

where p^0 is natural disease risk, p^A is AI-controllable risk, and p^P is patient-input or adherence risk. The action spaces are compact and restricted so that $p(e, a) \in [0, 1]$ for all feasible (e, a); equivalently, the additive specification can be read as a local reduced form around the relevant allocation. The AI developer chooses safety effort $e \geq 0$ at cost $C(e)$; the patient chooses information and adherence effort $a \geq 0$ at cost $G(a)$. Assume C and G are twice continuously differentiable, strictly increasing, and strictly convex, and that $p^A_e < 0$ and $p^P_a < 0$.

The first-best allocation minimizes expected social cost:

$$\min_{e,a} \; C(e) + G(a) + L\{p^0 + p^A(e) + p^P(a)\}. \tag{2.2}$$

An interior first best (e^*, a^*) satisfies

$$C'(e^*) + Lp^A_e(e^*) = 0, \tag{2.3}$$

$$G'(a^*) + Lp^P_a(a^*) = 0. \tag{2.4}$$

The *medical record* is a random variable $Z \in \mathcal{Z}$. It may include patient statements, uploaded measurements, model outputs, warnings, confidence scores, timestamps, prescriptions, lab orders, follow-up instructions, and later clinical events. A *liability rule* is a measurable transfer

$$w(Z, Y) \in [0, L] \tag{2.5}$$

paid to the patient and charged to the AI enterprise. Because tort damages and warranties usually pay after injury, many results below use the restricted form $w(Z, Y) = Y\widetilde{w}(Z)$ with $0 \leq \widetilde{w}(Z) \leq L$. Allowing bonuses in non-injury states would enlarge the feasible contract set; the focus here is on damages-like legal transfers. Let

$$T(e, a) = \mathbb{E}[w(Z, Y) \mid e, a] \tag{2.6}$$

be expected liability. The AI chooses e to minimize $C(e) + T(e, a)$. The patient chooses a

to minimize $G(a) + Lp(e,a) - T(e,a)$, because liability compensates part of the patient's expected loss.

It is useful to write $D = (Z, Y)$ for the full liability-relevant record. I write $f(d \mid e, a)$ for the density of D under actions (e, a). When differentiability holds, define the score for action $x_j \in \{e, a\}$ by

$$s_j(D) = \frac{\partial}{\partial x_j} \log f(D \mid x), \qquad x = (e, a), \quad j \in \{e, a\}. \tag{2.7}$$

This is the score-function object used in the first-order approach and in the joint-causation result below. Unless otherwise noted, expectations in the local implementation arguments are taken under the action profile being evaluated.

Assumption 1 (First-order approach)**.** *The following conditions hold.*

(i) Regularity. The action spaces are compact intervals, the allocation studied is interior, and the distribution of $D = (Z, Y)$ has a strictly positive density $f(d \mid e, a)$ with respect to a common measure. The density is differentiable in actions, and the score functions s_e and s_a are square-integrable.

(ii) Monotonicity for cutoff rules. When I interpret an optimal rule as negligence, safe harbor, or another cutoff, there is a scalar legally usable statistic $r(D)$ and an order on its support such that the admissible monotone rules are threshold or upper-set rules in r. The conditional distribution of $r(D)$, or equivalently the induced legal score used below, satisfies the appropriate single-crossing or monotone-likelihood-ratio condition in the action whose incentive the cutoff is meant to provide. The implementability equalities below do not require this monotonicity assumption; it is used only for cutoff interpretations.

(iii) Convexity of private problems. Private objective functions are convex in the relevant actions for every admissible bounded monotone transfer.

I state the assumption this way to avoid using a single convexity-of-the-distribution-function condition for both sides of the market. Because liability is paid by the AI and received by the patient, the same monotone transfer enters the two private objectives with opposite signs. The paper therefore uses the first-order approach as an implementability device: monotone likelihood ratios and single crossing justify monotone rules, while the separate convexity requirement ensures that local conditions are global. If these conditions fail, the equations below remain necessary local conditions.

I use implementation to mean that the target action profile is a Nash equilibrium of the induced game. Strict convexity gives a unique best response for each agent. If one wants uniqueness of the equilibrium rather than implementation of the intended equilibrium, a standard diagonal-dominance or contraction condition on the best-response map can be added without changing the liability formulas.[4] I do not impose it in the main text because the economic issue is the incentive moment generated by the record, not equilibrium selection.

Theorem 1 (Static implementability)**.** *Suppose* (e^*, a^*) *is an interior first-best allocation. A differentiable liability rule implements* (e^*, a^*) *only if*

$$T_e(e^*, a^*) = Lp_e^A(e^*), \tag{2.8}$$

$$T_a(e^*, a^*) = 0. \tag{2.9}$$

Conversely, if (2.8) *and* (2.9) *hold and the private objective functions are convex in the relevant actions, then* (e^*, a^*) *is privately optimal.*[5]

The theorem says that liability must make the AI internalize the marginal accident cost it controls, but must not insure the patient against the marginal consequences of the patient's own hidden action. Full outcome warranty would generally violate (2.9); no liability would generally violate (2.8).

[4] I use implementation to mean target-equilibrium implementation: a liability rule implements an action profile if that profile is a Nash equilibrium of the induced effort game. This is weaker than requiring the induced game to have a unique equilibrium. The first-order approach and Lemma 1 ensure that, holding the other agent's action fixed, each agent's unilateral problem is well behaved. They do not by themselves rule out multiple intersections of the best-response functions.

To make this distinction explicit, define the developer's private objective and the patient's private objective by

$$J^D(e, a) = C(e) + T(e, a), \qquad J^P(a, e) = G(a) + Lp(e, a) - T(e, a).$$

When the unilateral problems are strictly convex and the target is interior, the local best-response slopes are

$$B'_D(a) = -\frac{J^D_{ea}(e, a)}{J^D_{ee}(e, a)}, \qquad B'_P(e) = -\frac{J^P_{ae}(a, e)}{J^P_{aa}(a, e)}.$$

A sufficient condition for local uniqueness and stability of the implemented equilibrium is the diagonal-dominance condition

$$\sup_{\mathcal{N}} \left| \frac{J^D_{ea}}{J^D_{ee}} \right| \sup_{\mathcal{N}} \left| \frac{J^P_{ae}}{J^P_{aa}} \right| < 1$$

on a neighborhood $\mathcal{N}$ of (e^*, a^*). Equivalently, the local best-response map is a contraction. I do not impose this condition in the main results because the economic object of interest is the incentive moment generated by the record. Without this additional condition, the results below should be read as local Nash implementation of the intended equilibrium, not as global equilibrium uniqueness.

[5] (e^*, a^*) is a Nash equilibrium of the induced effort game. If the diagonal-dominance condition in the previous footnote also holds, this equilibrium is locally unique.

The following primitive sufficient condition is stronger than necessary but useful. It shows that the first-order approach is not doing hidden work in the baseline: bounded damages and enough curvature of effort costs make the agents' unilateral problems convex for every admissible legal rule.

Lemma 1 (Primitive sufficient condition for the first-order approach)**.** *Fix the other agent's action. Suppose the density of the record is twice differentiable and, uniformly over actions,*

$$\int \left| \frac{\partial^2 f(d \mid e, a)}{\partial e^2} \right| \mathrm{d}d \leq M_e, \qquad \int \left| \frac{\partial^2 f(d \mid e, a)}{\partial a^2} \right| \mathrm{d}d \leq M_a. \tag{2.10}$$

For every liability rule satisfying $0 \leq w \leq L$, the developer's private objective is strictly convex in e if $C''(e) \geq LM_e + \varepsilon$ for some $\varepsilon > 0$. The patient's private objective is strictly convex in a if $G''(a) + Lp_{aa}(e, a) \geq LM_a + \varepsilon$. Under these inequalities, the local first-order conditions used below are global optimality conditions for each unilateral problem.

The lemma is deliberately conservative. Rogerson-Jewitt conditions can justify the first-order approach in richer principal-agent environments (Rogerson, 1985; Jewitt, 1988). Here I need a condition that survives the sign reversal created by zero-sum liability: the transfer is a cost to the AI but a benefit to the patient. The bounded-damages curvature condition gives such a primitive sufficient condition. Monotone likelihood ratios and single crossing enter separately when I ask whether the optimal rule can be represented as a monotone negligence or safe-harbor cutoff.

2.2 Attribution-based fault in the separable-cause benchmark

The clean benchmark is attribution-based AI-fault liability. Suppose each injury has a latent cause $C \in \{A, P, N\}$, where A is AI-controllable fault, P is patient-controlled fault, and N is natural disease progression. The record does not necessarily reveal C, but it may contain information about it. This subsection assumes mutual exclusivity and separability: AI effort affects only $\mathbb{P}(C = A, Y = 1)$ and patient effort affects only $\mathbb{P}(C = P, Y = 1)$.

Corollary 1 (Attribution-based AI-fault liability implements the first best)**.** *Suppose there exists a legally usable statistic $q_A(Z, Y) \in [0, 1]$, fixed as a function of the record, such that for all (e, a) in a neighborhood of (e^*, a^*),*

$$\mathbb{E}[q_A(Z, Y)Y \mid e, a] = \mathbb{P}(C = A, Y = 1 \mid e, a). \tag{2.11}$$

If AI effort affects only $\mathbb{P}(C = A, Y = 1)$ *and patient effort affects only* $\mathbb{P}(C = P, Y = 1)$*, then*

$$w^*(Z, Y) = Lq_A(Z, Y)Y \tag{2.12}$$

(under the convexity and local-implementation conditions of Theorem 1) implements the interior first-best allocation. Equivalently, it satisfies the AI incentive condition (2.8) *and the patient incentive condition* (2.9)*.*

This rule is strict liability for the legally attributed AI-caused component of harm, not strict liability for every bad medical outcome. The familiar posterior statistic $q_A(Z, Y) = \mathbb{P}(C = A \mid Z, Y = 1)$ satisfies (2.11) only when it is a stable local attribution statistic, not merely the Bayesian posterior computed under one action profile. The point is important: a liability rule is fixed ex ante as a function of the record, whereas a Bayesian posterior can change with the action profile that generated the record. When the stability condition holds, the posterior language is a convenient shorthand. When it fails, a locally unbiased attribution statistic is one sufficient route to the first-best moments, but not the only possible route: another record-based transfer could in principle satisfy (2.8) and (2.9) without being unbiased for AI-caused injury in levels.

The attribution rule is therefore the clean separable-cause ideal behind comparative fault, causation presumptions, and conditional warranties, not a necessary representation of every first-best transfer. A warranty that pays whenever the patient is harmed treats nonadherence and natural progression as AI-caused. A rule that never pays treats AI-caused harm as patient risk. In contrast, the attribution rule pays the expected AI-caused component given the legally usable record.

2.3 Joint causation and marginal responsibility

The posterior-fault benchmark is not the general case. Suppose instead that harm is $p(e, a)$ with no additive decomposition. The AI may affect the patient's adherence by the clarity of warnings; the patient may affect the AI's accuracy by the quality of symptoms supplied; and the same clinical workflow may change both. There need not be a latent cause C that partitions harm into AI, patient, and nature.

Using the density and score notation introduced in Section 2, for any square-integrable liability rule,

$$\frac{\partial}{\partial x_j}\mathbb{E}[w(D)] = \mathbb{E}[w(D)s_j(D)]. \tag{2.13}$$

Proposition 1 (Joint causation)**.** *In the nonseparable environment, a differentiable liability rule locally implements an interior first best only if*

$$\mathbb{E}[w(D)s_e(D)] = Lp_e(e^*, a^*), \qquad \mathbb{E}[w(D)s_a(D)] = 0. \tag{2.14}$$

If the Gram matrix $G_s = \mathbb{E}[s(D)s(D)']$ *is nonsingular and transfers are unconstrained in* L^2*, the unique minimum-second-moment (equivalently, minimum-*L^2*-norm) transfer satisfying* (2.14) *is*

$$w^J(D) = s(D)'G_s^{-1}\tau, \qquad s(D) = \begin{pmatrix} s_e(D) \\ s_a(D) \end{pmatrix}, \quad \tau = \begin{pmatrix} Lp_e(e^*, a^*) \\ 0 \end{pmatrix}. \tag{2.15}$$

Posterior AI-fault liability is a special case. It implements the first best only when the posterior-cause statistic generates the same two moments as (2.14)*.*

This proposition is the contract theory version of joint medical causation. The legal rule should not ask only which box the injury falls into. It should ask which record components are informative about the AI's marginal contribution to risk without dulling patient incentives. If the AI failed to elicit symptoms, the patient omission and the AI design are complements. A comparative-fault partition can still be useful, but it is an approximation to the score-based marginal rule rather than a literal partition of isolated causes.

The score objects in (2.14) are not meant to be directly observed by a court in the way a timestamp or lab result is observed. They are sufficient-statistic targets for institutional design. In practice, courts, insurers, and regulators would approximate the relevant covariances using audit logs, validation studies, expert chart review, claims experience, randomized interface tests, and model-update data. For example, if warning language is randomized or changed across deployments, the induced changes in adherence and harm help estimate the record components that load on s_e and s_a. If chart review codes whether requested inputs were supplied, whether red flags were asked about, whether the warning was understandable, and whether follow-up occurred, those codes are low-dimensional proxies for the legal score. Thus the planner benchmark in Proposition 1 becomes operational through record design and empirical score approximation; negligence, safe harbors, comparative fault, and warranties are implementable approximations to these moments rather than literal observations of hidden effort derivatives.

There is an important feasibility qualification. The transfer w^J in Proposition 1 is an unconstrained L^2 object. It can be negative, exceed the loss, or use score components that

courts cannot observe. Real liability rules are usually bounded below by zero, bounded above by damages, and measurable only with respect to the legal record. These constraints are not cosmetic. They create the bilateral moral-hazard bottleneck.

Theorem 2 (Feasible moment set and the zero-sum bottleneck)**.** *Let*

$$\mathcal{W}_{\mathcal{G},L} = \{w : 0 \leq w(D) \leq L,\ w \text{ is } \mathcal{G}\text{-measurable}\}. \tag{2.16}$$

Define the feasible incentive-moment set

$$\mathcal{M}(\mathcal{G}, L) = \{(\mathbb{E}[ws_e], \mathbb{E}[ws_a]) : w \in \mathcal{W}_{\mathcal{G},L}\} \subset \mathbb{R}^2. \tag{2.17}$$

Then $\mathcal{M}(\mathcal{G}, L)$ is compact and convex. The first-best marginal incentives are implementable by bounded nonnegative legal transfers only if

$$\tau = (Lp_e(e^*, a^*), 0) \in \mathcal{M}(\mathcal{G}, L). \tag{2.18}$$

If this condition fails, no bounded record-based liability rule implements the first best. If local welfare depends only on the missed incentive vector through a positive definite quadratic loss, the second-best incentive vector is the Q-projection of τ onto $\mathcal{M}(\mathcal{G}, L)$. If welfare also contains direct risk-bearing, distributional, proof-cost, or claims-processing terms that depend on the transfer rule itself, the second-best problem is the full optimization over $w \in \mathcal{W}_{\mathcal{G},L}$; it generally cannot be reduced to a projection of moment vectors alone because different transfers can generate the same moment vector with different direct welfare effects. Moreover, if the conditional scores are collinear,

$$\mathbb{E}[s_a \mid \mathcal{G}] = \rho \mathbb{E}[s_e \mid \mathcal{G}] \quad a.s. \tag{2.19}$$

for some $\rho \neq 0$, then every feasible moment satisfies $m_a = \rho m_e$. Hence a target with $Lp_e(e^*, a^*) \neq 0$ and zero patient-insurance moment is infeasible.

This theorem is the formal reason partial and conditional liability are not merely political compromises. If the record is coarse, the same payment that gives the AI incentives also insures the patient's hidden action. Richer records help only when they break this collinearity by generating components that are informative about AI marginal responsibility but not about patient shirking, or vice versa.

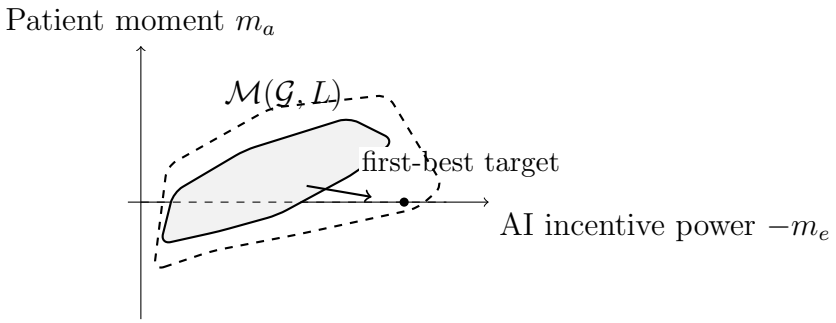


Figure 1: The legal record determines the feasible set of liability incentives. The horizontal axis uses the sign-normalized AI incentive power because higher effort lowers harm in the medical application. A coarse record can force AI incentives and patient-insurance distortions to move together, leaving the target outside the set. Richer legally usable information can expand the set by separating AI-responsibility from patient-responsibility signals.

2.4 Compensation, risk bearing, and adoption value

The benchmark above is risk neutral. In that benchmark, compensation has no social value by itself. Transfers matter because they change behavior and because they generate administrative or insurance costs. I therefore separate three channels that are sometimes bundled together in policy discussions: deterrence, insurance, and adoption.

Let patient i have utility $u_i(c)$ over consumption after medical loss and compensation, with u_i weakly concave. Let $A_i(\lambda)$ be the money-metric private value of accountability from liability power λ. I write the adoption probability, or the adoption rate for a continuum of patients of type i, as $x_i(\lambda) = \widetilde{x}_i(A_i(\lambda), p(\lambda), X_i)$, suppressing price and covariates when no confusion arises. At the individual level adoption may be binary; the differentiable object below is the choice probability or aggregate adoption rate. Let $S_i(\lambda)$ be the social surplus from AI use by patient i, exclusive of pure transfers but including medical outcomes, costs, and any external value of the data generated by that use. A reduced-form social accountability term can be written locally as

$$R'(\lambda) = \underbrace{\frac{\partial}{\partial \lambda}\text{InsuranceValue}(\lambda)}_{\text{risk bearing or liquidity}} + \underbrace{\mathbb{E}\left[x_i'(\lambda) S_i(\lambda)\right]}_{\text{activity-level adoption effect}} + \underbrace{\mathbb{E}\left[x_i(\lambda) S_{i\lambda}(\lambda)\right]}_{\text{effect on care conditional on adoption}}. \tag{2.20}$$

Proposition 2 (When compensation has welfare value)**.** *If patients and firms are risk neutral, adoption is fixed, financing is actuarially fair, and transfers have no administra-*

tive cost, compensation has no direct welfare effect; only the incentive effects in Theorem 1 matter. If patients are risk averse or liquidity constrained, compensation can have positive insurance value. If adoption is endogenous, private willingness to pay for liability is a demand shifter; its welfare effect operates through the social surplus of marginal adopters and any learning externality they create. Separately, liability can change care conditional on adoption through deterrence, defensive design, or workflow responses; that inframarginal channel is the $\mathbb{E}[x_i S_{i\lambda}]$ *term in* (2.20)*, not something identified by willingness-to-pay evidence alone.*

This proposition disciplines the trust term used below. A discrete-choice experiment that estimates a liability coefficient identifies a private willingness to pay or a demand shift. It is not automatically a welfare sufficient statistic. It becomes one after the researcher specifies the marginal surplus of AI use, the cost of financing compensation, and the risk-bearing or liquidity value of payment after injury.

3 Information, Contract Shape, and Legal Form

I now move from implementation to legal form. The question is no longer only whether a target incentive vector can be generated, but what the optimal feasible rule looks like when the law can use some record events and not others. This section keeps the record exogenous. It derives the value of auditability, the local legal score, the difference between negligence cutoffs and continuous warranties, and the role of costly fact-finding by courts and insurers.

3.1 Auditability and the informativeness principle

Let $\mathcal{H}$ be the sigma-field generated by the full verifiable record on which liability could in principle be conditioned, and let $\mathcal{W}(\mathcal{H})$ be the set of square-integrable liability rules measurable with respect to $\mathcal{H}$. Let $g(w)$ denote the vector of marginal incentives generated by w through (2.13), and let τ be the first-best incentive vector from Theorem 1 or Proposition 1. Suppose local welfare equals a constant minus

$$\frac{1}{2}\|g(w) - \tau\|_Q^2 \tag{3.1}$$

for a positive definite matrix Q.

Proposition 3 (Value of auditability). *If $\mathcal{H}' \supseteq \mathcal{H}$ and the record is exogenous, then the maximum local welfare attainable with $\mathcal{H}'$ is weakly higher than the maximum local welfare attainable with $\mathcal{H}$. The improvement is strict if and only if the Q-projection of τ onto $g(\mathcal{W}(\mathcal{H}'))$ differs from the Q-projection of τ onto $g(\mathcal{W}(\mathcal{H}))$.*

The qualifier "exogenous" matters. This proposition applies to genuine auditability: logs, measurements, and follow-up records that are generated independently of strategic litigation design. Section 4.1 lets the AI shape the record itself. In that case, a larger formal record need not be welfare improving because it may be a jamming technology rather than an information technology.

3.2 Deriving the legal score and the shape of the optimal contract

I next characterize the shape of the optimal feasible rule. The object often looks like a negligence cutoff, but it need not. The shape depends on the legal score generated by the incentive constraints, the risk-bearing value of compensation, and the cost of paying claims.

Let $D = (Z, Y)$ denote the verifiable record, and let $\mathcal{G} \subseteq \sigma(D)$ be the sigma-field legally usable by a court, insurer, or contract. I keep the unconditional notation of Section 2 for incentive moments. Because the legal instruments studied here are damages-like, write the transfer as

$$w(D) = Y\widetilde{w}(D), \qquad 0 \leq \widetilde{w}(D) \leq L, \tag{3.2}$$

where $\widetilde{w}$ is $\mathcal{G}$-measurable on injury records. Thus incentives are generated by the unconditional moments $\mathbb{E}[Y\widetilde{w}(D)s(D)]$, while the pointwise payment problem can be written over records with $Y = 1$. Let Γ be a convex claims-processing or insurance-loading cost with $\Gamma(0) = 0$. To make the legal score primitive rather than tautological, write the local second-best problem as

$$\max_{0\leq\widetilde{w}\leq L,\ \widetilde{w}\ \mathcal{G}\text{-measurable}} \mathbb{E}\left[Y\{(\omega(D) - c(D))\widetilde{w}(D) - \Gamma(\widetilde{w}(D))\}\right] - \frac{1}{2}\left\|\mathbb{E}[Y\widetilde{w}(D)s(D)] - \tau\right\|_Q^2, \tag{3.3}$$

where $s(D)$ is the vector of record scores that generate incentives, τ is the target incentive vector, Q is positive definite, $\omega(D)$ is the direct social value of paying one more dollar on record D from insurance, liquidity, or distributional weights, and $c(D)$ is marginal administrative, evidentiary, or insurance cost. The quadratic term is a local approximation to the welfare loss from missing the desired incentive vector. Equivalently, with hard

incentive constraints $\mathbb{E}[Y\widetilde{w}s_j] = \tau_j$, the same score arises from Lagrange multipliers. At an optimum $\widetilde{w}^*$ of the penalized problem, define

$$\psi^* = Q\{\tau - \mathbb{E}[Y\widetilde{w}^*(D)s(D)]\}, \tag{3.4}$$

or, in the constrained problem, let ψ^* denote the vector of multipliers on the incentive constraints. The primitive local score evaluated at the optimum is

$$S_0^*(D) = \omega(D) - c(D) + \psi^{*\prime}s(D). \tag{3.5}$$

The multiplier vector ψ^* is determined by the global incentive problem; it is not a free taste parameter. When no confusion is possible, I suppress the star and write S_0. The usable legal score is the projection of this primitive score onto the legally available record,

$$\bar{S}_{\mathcal{G}}(D) = \mathbb{E}[S_0(D) \mid \mathcal{G}, Y = 1]. \tag{3.6}$$

Under additional separable-cause restrictions, (3.5) can be written in posterior-cause form. Specifically, suppose the direct payment value $\omega(D)$ is constant or cause-specific on injury records, and the injury-state score components entering $\psi' s(D)$ reduce to stable posterior cause probabilities up to constant shadow weights. Then, in the separable latent-cause benchmark, (3.5) can be represented as

$$\bar{S}_{\mathcal{G}}(D) = \mathbb{E}\left[\mu_A \mathbb{P}(C = A \mid D) - \mu_P \mathbb{P}(C = P \mid D) - \mu_N \mathbb{P}(C = N \mid D) - c(D) \mid \mathcal{G}, Y = 1\right] \tag{3.7}$$

for shadow weights μ_A, μ_P, μ_N implied by the incentive and risk-bearing constraints. Equation (3.7) is therefore a special representation of (3.5) under these additional cause-specific value and stable-within-cause score restrictions, not an implication of separability alone and not an independent assumption that paying is good whenever a named score is positive.

Conditional on a candidate multiplier vector ψ, the legal choice is pointwise on each legally distinguishable injury record. For $\mathbb{P}(\cdot \mid Y = 1)$-almost every record D, the injury-state payment solves

$$\widetilde{w}^*(D) \in \arg\max_{x \in [0,L]} \left\{\bar{S}_{\mathcal{G}}(D)x - \Gamma(x)\right\}. \tag{3.8}$$

This pointwise expression is obtained by differentiating the global local problem (3.3) with respect to $\mathcal{G}$-measurable perturbations. The factor $\mathbb{P}(Y = 1 \mid \mathcal{G})$ that appears in the unconditional derivative is nonnegative and therefore does not change the maximizer

on cells where injury has positive probability. On cells with zero injury probability, the injury-state transfer is payoff-irrelevant and may be chosen arbitrarily. A solution of the penalized or constrained problem is obtained when the pointwise solution below is combined with the global consistency condition for ψ^* in (3.4), or with the corresponding hard-constraint multipliers.

Theorem 3 (Optimal contract shape)**.** *Fix the multiplier vector and let* $\bar{S}_{\mathcal{G}}(D) = \mathbb{E}[S_0(D) \mid \mathcal{G}, Y = 1]$. *If* Γ *is differentiable and strictly convex, the unique injury-state payment solving* (3.8) *is*

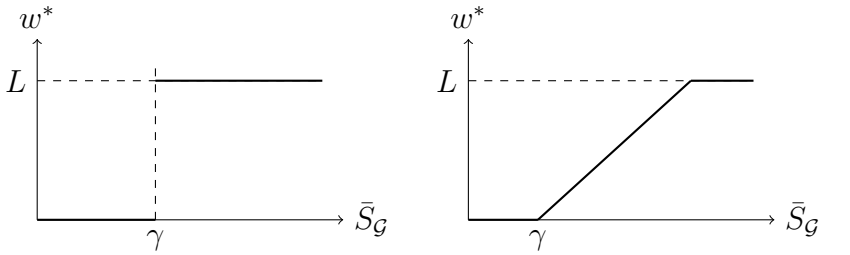
$$\widetilde{w}^*(D) = g_\Gamma(\bar{S}_{\mathcal{G}}(D)), \tag{3.9}$$

where g_Γ *is the extended inverse of* Γ' *on* $[0, L]$*: it equals 0 for scores below the right derivative at 0, equals* $(\Gamma')^{-1}$ *on the interior range of* Γ'*, and equals* L *for scores above the left derivative at* L*. The unconditional damages rule is* $w^*(D) = Y\widetilde{w}^*(D)$*. If instead* $\Gamma(w) = \gamma w$ *is affine, every solution is bang-bang except possibly on the indifference set:*

$$\widetilde{w}^*(D) = L\mathbf{1}\{\bar{S}_{\mathcal{G}}(D) > \gamma\}, \tag{3.10}$$

with arbitrary mixing on $\{\bar{S}_{\mathcal{G}} = \gamma\}$*. If* Z *is scalar and* $z \mapsto \bar{S}_{\mathcal{G}}(z)$ *is nondecreasing, then the bang-bang rule is a threshold rule. In the three-cause representation, a sufficient condition is that the shadow-weighted posterior score itself is nondecreasing: the AI-fault term enters with a positive coefficient, the patient-fault and natural-progression terms enter with the displayed negative coefficients, their posterior components move monotonically in the corresponding directions, and* $c(z)$ *is nonincreasing in the same order.*

The theorem separates two issues. A richer record changes the statistic on which liability is based. It does not by itself make the injury-state payment continuous. With affine claims costs the optimal rule remains bang-bang on the usable score. A genuinely continuous warranty requires strictly convex loading, claim-size variation, or another reason why the marginal cost of payment rises smoothly with the amount paid. Negligence is a threshold rule on a breach score. A safe harbor is a rule in which certified compliance moves records into a low-score region. Strict liability and no liability are boundary cases in which all injury records lie on one side of the cutoff.



Affine cost: cutoff rule Convex cost: continuous warranty

Figure 2: The same primitive legal score can generate different legal forms. With affine claims-processing costs, the optimal transfer is a threshold rule on the usable score. With strictly convex loading, the injury-state payment increases continuously over the interior range. A richer record changes the horizontal score; convex loading changes the vertical shape of the payment rule.

3.3 Legal regimes as partitions of the record

Different legal regimes use different sigma-fields.[6] Let $\mathcal{H} = \sigma(Z, Y)$ be the full verifiable medical record. A regime is a pair $(\mathcal{G}, \mathcal{A})$, where $\mathcal{G} \subseteq \mathcal{H}$ is the record partition on which liability may depend and $\mathcal{A} \subseteq \mathcal{W}(\mathcal{G})$ is the admissible set of transfers. No liability uses the trivial sigma-field and the singleton transfer $w = 0$. Strict outcome liability uses $\sigma(Y)$ and transfers proportional to injury. Negligence uses $\sigma(Y, B)$, where B is a breach indicator. A safe harbor uses $\sigma(Y, H, B\mathbf{1}\{H = 0\})$, where H indicates certified compliance. Comparative fault uses $\sigma(Y, B, Q)$, where Q records patient compliance or nonadherence. Posterior AI-fault liability uses the richest available field when causes are separable; in joint-causation environments the richest field supports the score-based marginal rule in Proposition 1.

Proposition 4 (Informativeness and legal partitions)**.** *Fix the primitive score S_0 and the multiplier vector that generates it. Consider two regimes with the same transfer bounds and cost function Γ. If $\mathcal{G}_2 \supseteq \mathcal{G}_1$ and the record is exogenous, then the optimized value of the fixed-score local problem (3.3), equivalently the integral of the pointwise values generated by (3.8), is weakly higher under $\mathcal{G}_2$ than under $\mathcal{G}_1$. Let*

$$v_\Gamma(s) = \max_{x \in [0,L]} \{sx - \Gamma(x)\}.$$

[6] This is shorthand for the facts that can be used after admissibility limits, burdens of proof, privacy constraints, discovery, and expert fact-finding are taken into account. Legal doctrine is fixed ex ante; the sigma-field describes the legally usable proof available for applying that doctrine.

Holding the primitive score and multiplier vector fixed, the gain is strict exactly when

$$\mathbb{E}\big[v_\Gamma(\mathbb{E}[S_0 \mid \mathcal{G}_2, Y=1]) \mid \mathcal{G}_1, Y=1\big] > v_\Gamma(\mathbb{E}[S_0 \mid \mathcal{G}_1, Y=1])$$

on a set of positive injury probability. This condition includes threshold-kink gains in the affine-cost case: splitting a coarse indifference cell into above- and below-threshold subcells can be strictly valuable even when the coarse cell itself lies at the cutoff. If the multiplier vector itself changes across regimes, the weak-value comparison remains valid, but the strict comparison must be made using the resulting optimized values rather than holding S_0 fixed.

This proposition is the sigma-field version of the informativeness principle. Logging mandates, explainability duties, patient-adherence records, pharmacy fills, and follow-up documentation are valuable because they refine the partition on which legal transfers can be based. They are not valuable merely because they produce more text in the record. They are valuable only to the extent that they change the conditional legal score used by the contract.

3.4 Court fact-finding as costly expansion of the record

The legal record is not only a database. Courts and insurers can expand it ex post through discovery, depositions, expert reports, clinical audits, and inspection of model documentation. I model this as a costly choice of an evidentiary experiment. Let $q \in Q$ index fact-finding intensity. The protocol generates an additional signal R_q about the underlying encounter, so the usable legal field is $\mathcal{G}(q) = \sigma(\mathcal{G}_0, R_q)$, where $\mathcal{G}_0$ is the pre-discovery record. Let $H(q)$ be the resource and error cost of using the protocol. The value of fact finding is

$$\max_{q \in Q} \ \mathcal{V}(q) - H(q), \tag{3.11}$$

where $\mathcal{V}(q)$ is the value of the local design problem under the experiment R_q.

Proposition 5 (Costly legal fact-finding)**.** *If experiment q' Blackwell-dominates experiment q and $H(q') \geq H(q)$, then the expansion from q to q' is welfare-improving if and only if*

$$\mathcal{V}(q') - \mathcal{V}(q) \geq H(q') - H(q). \tag{3.12}$$

If q belongs to a differentiable parametric family of experiments and the optimum is interior, then

$$\mathcal{V}'(q) = H'(q), \tag{3.13}$$

where the derivative is with respect to the distribution of posterior legal scores induced by R_q, not with respect to a sigma-field as an abstract set.

This is how I interpret discovery and expert litigation in the model. Ex ante audit logs and regulatory documentation are valuable when they produce the same refinement of $\mathcal{G}$ at lower cost or with less ex post error. Conversely, if ex post fact-finding is cheap and accurate, the case for rigid ex ante logging is weaker.

3.5 Pre-deployment evidence and launch thresholds

The record can also contain evidence produced before the first patient is treated. Let $q_k \geq 0$ denote validation effort for condition k: clinical testing, calibration on a local population, external audit, red-team evaluation, or prospective monitoring before broad deployment.[7] It generates evidence E_k with distribution $F_k(\cdot \mid q_k, \theta_k)$, where θ_k is the underlying riskiness or reliability of the AI for that condition. A liability rule can condition later damages on E_k whenever that evidence is part of the legally usable record. In that case E_k is simply one component of Z and refines $\mathcal{G}$.

This is the point at which the model connects most directly to Poggi and Strulovici (2020). They show that liability can be designed as a tariff on evidence acquired by a firm before launching a risky product. In the present setting, such a tariff has a condition-specific interpretation. For each condition k, a validation-evidence tariff changes both the due-diligence choice q_k and the launch or coverage decision. If $\Pi_k(E_k, T_k)$ denotes expected private profit after observing validation evidence and facing tariff T_k, then high-powered service for condition k is offered only when

$$\Pi_k(E_k, T_k) \geq 0. \tag{3.14}$$

Monotone tariffs can therefore implement evidence-based launch thresholds: weak validation evidence leads to no launch, no autonomous treatment, or only a low-powered triage warranty; strong evidence permits broader coverage.

The contribution of the present model begins after this pre-deployment margin. In AI medicine, evidence is not exhausted by validation before launch. The encounter itself creates the legally relevant record, and that record is affected by patient inputs, adherence, clinical workflow, and algorithmic defensive design. Thus a tariff on evidence must be combined with the bilateral moral-hazard and endogenous-record constraints above.

[7] Condition-specific validation is a reasonable modeling choice given the author's conversations with one of the leading AI labs.

Pre-deployment evidence helps decide whether the AI may enter a task; encounter-level evidence determines how responsibility is priced once the AI actually gives medical advice.

3.6 General comparative statics without quadratic payoffs

The scalar liability share remains useful for comparative statics. In this simple section, let $\lambda \in [0,1]$ index the power of a fixed liability rule and let $W(\lambda;\theta)$ be the planner's reduced objective after the AI and patient choose equilibrium actions. The baseline parameter vector contains verifiability, AI control, patient contamination, legal cost, insurance value, and adoption value. Defensive design is introduced in Section 4.1; the same monotone-comparative-static logic applies there with additional parameters.

Proposition 6 (Monotone comparative statics of liability)**.** *Suppose W is upper semi-continuous in λ and has single crossing in (λ,θ) in the sense of Milgrom and Shannon (1994). Then the largest and smallest selections from $\arg\max_{\lambda\in[0,1]} W(\lambda;\theta)$ are increasing in any parameter that raises the marginal value of liability in the single-crossing order and decreasing in any parameter that lowers it. If W is twice continuously differentiable and strictly concave in λ, and the unique optimum is interior, then*

$$\frac{\partial \lambda^*}{\partial \theta_j} = -\frac{W_{\lambda\theta_j}(\lambda^*;\theta)}{W_{\lambda\lambda}(\lambda^*;\theta)}. \tag{3.15}$$

At a binding boundary, the comparative static is governed by the Kuhn-Tucker condition rather than this interior implicit-function formula.

I state the result without linear risk or quadratic costs. The quadratic index below is a clean illustration of the cross-partials. Verifiability deserves care. A more informative record raises the effectiveness of a dollar of liability, but it need not raise the optimal nominal share λ. In some environments the planner can use a lower nominal share because each dollar is better targeted. The relevant object is the marginal value of the liability instrument, not the sign of a primitive called verifiability in isolation.

3.7 A liability index

The scalar index in this subsection is deliberately a one-dimensional slice of the record-based problem, not a replacement for it. Sections 2 and 3 characterize the feasible transfer set and the legal score on an arbitrary legal sigma-field. Here I compress a particular family of damages rules into one power parameter λ. The parameters σ, η, ν, ϕ, and

later Δ, should be read as sufficient statistics generated by the underlying record: they summarize how the chosen legal rule loads on AI-controllable harm, patient behavior, natural disease, defensive record design, and learning. The index is therefore a calculator for comparing restricted legal instruments, not the general implementability theorem.

For classification, specialize to a linear-quadratic environment. Let

$$p^A(e) = \bar{p}^A - e, \qquad p^P(a) = \bar{p}^P - a, \tag{3.16}$$

with costs

$$C(e) = \frac{e^2}{2\alpha}, \qquad G(a) = \frac{a^2}{2\beta}, \tag{3.17}$$

where $\alpha > 0$ is AI safety productivity and $\beta > 0$ is patient information/adherence productivity. It is important that λ is a share of legally attributed harm, not necessarily a share of total medical harm. The parameter $\sigma \in [0, 1]$ measures how strongly liability loads on AI-caused injury; parameter $\eta \in [0, 1]$ measures contamination from patient-caused injury; and the parameter $\nu \in [0, 1]$ measures contamination from natural disease progression. Full outcome liability is the special case $\sigma = \eta = \nu = 1$. Posterior AI-fault liability is the limiting case $\sigma = 1$ and $\eta = \nu = 0$. Thus σ, η, and ν belong in the transfer technology rather than in the medical risk functions: they describe what the legal system attributes to the AI, not what medically happened. A scalar liability share $\lambda \in [0, 1]$ generates expected liability

$$T^\lambda(e, a) = \lambda L\{\sigma(\bar{p}^A - e) + \eta(\bar{p}^P - a) + \nu p^0\}. \tag{3.18}$$

The natural-risk term affects prices, loading, and compensation but not first-order care incentives. To avoid overinterpreting patient moral hazard, let $\xi \in [0, 1]$ measure how salient and complete expected tort compensation is for patient effort. The AI side bears the full transfer T^λ, but the patient's adherence choice responds to the perceived or salient compensation value ξT^λ. The rational full-salience benchmark is $\xi = 1$. If patients do not change adherence in response to expected legal compensation, $\xi = 0$; patient-caused claims may still matter through prices, insurance loading, and attribution error.[8]

[8]The parameter ξ is a private-response wedge, not an additional social welfare term. It can be microfounded in either rational or behavioral terms. On the rational interpretation, the patient may expect only a fraction of gross damages to affect her own continuation payoff because recovery is delayed, probabilistic, costly to claim, or partly absorbed by legal fees and other transaction costs. On the behavioral interpretation, the patient may underweight expected legal compensation when choosing information and adherence effort. More generally, if the patient behaves as if the expected net recovery from a gross legal

Let administrative, litigation, and insurance costs be

$$K(\lambda) = k_1\lambda + \frac{k_2}{2}\lambda^2, \tag{3.19}$$

Section 2.4 described the marginal welfare term $R'(\lambda)$. For the linear-quadratic illustration, I use a second-order local approximation to the level function,

$$R(\lambda) = r_1\lambda - \frac{r_2}{2}\lambda^2, \tag{3.20}$$

so that $R'(0) = r_1$ and $R''(\lambda) = -r_2$. The coefficient r_1 is a social marginal value at zero liability, after mapping any private adoption response into social surplus.[9] A binary no-liability/full-liability choice experiment identifies a private money-metric value of the liability attribute, or a demand shift. It identifies $R(1) - R(0) = r_1 - r_2/2$ only after the researcher has imposed the welfare mapping from private adoption value to social surplus, financing costs, and risk-bearing value; even then it does not identify r_1 unless $r_2 = 0$ or curvature is separately estimated. More generally, if a discrete-choice experiment estimates

$$U_{ij} = \delta_j + \rho_L \mathbf{1}\{\text{liability}_j\} + \rho_p p_j + X_j'\rho + \varepsilon_{ij}, \tag{3.21}$$

then $-\rho_L/\rho_p$ is a private money-metric value of the liability attribute. To map it into r_1 one must combine it with the marginal surplus of adoption, financing costs, and the risk-bearing value in Proposition 2. The estimates in Chan, Xu, and Cutler (2026) are

transfer T^λ is

$$\widehat{T}^P_\lambda = \xi T^\lambda,$$

then ξ is the local pass-through of gross expected liability into the patient's private effort decision. The benchmark $\xi = 1$ is the fully salient, costless, full-recovery case. The benchmark $\xi = 0$ is the case in which expected legal compensation does not affect patient effort. If ξ reflects rational litigation costs, those resource costs should also be included in $K(\lambda)$; empirically, one should avoid double counting by treating $K(\lambda)$ as the social resource cost of the legal system and ξ as the pass-through of gross compensation into patient behavior.

[9] A related interpretation of this term is warranty signaling. Suppose the AI firm privately observes a quality type $\theta \in \{H, L\}$, with high-quality AI generating lower expected claim costs. Patients observe the liability share or warranty λ and form a posterior belief $\pi(\lambda) = \Pr(\theta = H \mid \lambda)$. Adoption can then be written as $x(\lambda, \pi(\lambda))$. Locally, when the posterior mapping is differentiable, the marginal value of liability contains not only the direct accountability or adoption effect but also an informational term such as $x_\pi(\lambda, \pi)\pi_\lambda(\lambda)S$, where S is the surplus of marginal AI use. This term raises optimal liability when stronger liability credibly identifies higher-quality AI and induces high-surplus adoption. The sign is not mechanically positive: voluntary warranties may already separate types, low-quality firms may mimic if records are poor or firms are judgment-proof, and excessive warranties may induce patient moral hazard or algorithmic defensive design. Thus warranty signaling can be folded into $R(\lambda)$, but it does not imply uniformly higher liability. This interpretation is related to warranty-signaling models such as Spence (1977) and Grossman (1981).

therefore useful because they discipline the accountability and adoption channel, not because they identify the entire liability index.

Proposition 7 (Simple liability index)**.** *In the linear-quadratic model* (3.16)–(3.20)*, the optimal liability share is*

$$\lambda^* = \left[\frac{L^2\alpha\sigma + r_1 - k_1}{L^2(\alpha\sigma^2 + \xi^2\beta\eta^2) + r_2 + k_2}\right]_0^1. \tag{3.22}$$

The numerator is the marginal benefit of liability at zero liability: AI safety incentives plus the social accountability/adoption value, net of first-order legal cost. The denominator is marginal curvature: diminishing marginal surplus from AI effort as effective liability approaches the first best, patient moral hazard from contaminated compensation scaled by salience ξ, diminishing marginal accountability value, and convex legal or insurance costs.

The AI-safety term should be interpreted as an aggregate product-level incentive when safety effort is embedded in the model and deployed at scale. A single calibration improvement, red-flag rule, prescribing safeguard, or interface change can reduce risk for many patients. If condition k has a mass μ_k of potential users and a product-level safety effort e_k reduces risk for all of them, the social marginal value of that effort is

$$\mu_k L_k(-p^A_{ke}(e_k)),$$

not merely the per-encounter marginal value. A per-injury liability rule partly captures this scale effect because expected liability also aggregates over users. Thus a larger user base need not mechanically raise the optimal nominal liability share λ_k; what matters is whether the developer's total expected liability exposure equals the aggregate marginal harm reduction created by product-level safety. If damages caps, judgment-proofness, low claiming rates, enterprise pass-through failures, or narrow safe harbors weaken this aggregation, the liability index should use a scale-adjusted AI-safety term. This direct scale effect is distinct from the learning term Δ_k below: α_k prices scalable safety improvements in the current model, while Δ_k prices future learning from adoption and verified data.

The index should be read as a liability calculator rather than as a general first-best implementation theorem. The first-best implementation conditions are the marginal equations in Theorem 1. The scalar index solves a restricted second-best problem in which the legal system chooses one liability power λ for an imperfectly attributed claim.

A useful simplification sets $r_1 = r_2 = k_1 = k_2 = 0$. The unconstrained interior index

is

$$\lambda^{0,u} = \frac{\alpha\sigma}{\alpha\sigma^2 + \xi^2\beta\eta^2},$$

and the constrained liability share is $\lambda^0 = [\lambda^{0,u}]_0^1$. The numerator is the AI-deterrence value of liability. The denominator is the sum of diminishing returns to AI deterrence and the patient-moral-hazard cost created by contaminated compensation. The interior expression has a useful implication: better attribution need not raise the nominal liability share. Let $B = \xi^2\beta\eta^2$. Then

$$\frac{\partial\lambda^{0,u}}{\partial\sigma} \gtrless 0 \quad \Longleftrightarrow \quad B \gtrless \alpha\sigma^2.$$

Thus nominal liability can be hump-shaped in record verifiability. What rises monotonically is not necessarily the damages share λ, but the effective AI incentive $\lambda\sigma$:

$$\lambda^{0,u}\sigma = \frac{\alpha\sigma^2}{\alpha\sigma^2 + B}.$$

A better record may allow the legal system to use a lower nominal liability share while generating stronger AI deterrence. In a primary-care triage example, better logs of red-flag questions, vital signs, escalation warnings, and patient follow-through may reduce the required breadth of damages even as they increase the effective price placed on AI-controllable mistakes.

Corollary 2 (No liability, full liability, and partial warranty)**.** *Let*

$$A = L^2\alpha\sigma + r_1 - k_1, \qquad D = L^2(\alpha\sigma^2 + \xi^2\beta\eta^2) + r_2 + k_2. \tag{3.23}$$

Then no AI medical liability is optimal if $A \leq 0$*. Full liability is optimal if* $A \geq D$*. Partial liability is optimal if* $0 < A < D$*.*

I use Corollary 2 as the simplest taxonomy. No liability is optimal when verifiable AI control and social accountability value are too small relative to legal cost. Full liability is optimal when AI-controllable risk is highly verifiable, patient moral hazard is small, and liability has sufficient social value. Partial warranty is the generic case because liability usually mixes AI deterrence with patient insurance.

3.8 Negligence and safe harbors in the simple model

The scalar share λ is only a reduced form. Negligence, warranty, and safe-harbor regimes are different functions $w(Z, Y)$. A negligence rule pays only when the record indicates breach of a due-care standard. A safe harbor sets liability to zero on a certified compliance set and applies a fallback rule outside that set. A conditional warranty pays only when the patient supplied required inputs and followed specified steps. In the separable simple model these rules differ by how well their induced transfer approximates posterior AI-fault liability. In the joint-causation model they differ by how well they approximate the marginal score rule. If the compliance record is a sufficient statistic for AI due care, negligence or a safe harbor can reproduce the desired marginal incentives with lower administrative cost. If the compliance record is coarse or manipulable, negligence weakens incentives or rewards formal compliance rather than clinical safety.

4 Why AI Is Different: Endogenous Records, Learning, and Institutions

The simple model isolates bilateral moral hazard under a no-algorithmic-defensive-design assumption. The general model relaxes that assumption and adds the margins that are distinctive in AI doctoring: algorithmic defensive design, endogenous records, clinical integration, insurance loading, adoption, learning across patients, direct prescribing, and scope across conditions. The scalar liability indices in the subsections below should be read as modular reduced-form illustrations of the record-based mechanism. Each module asks how a particular institution changes the sufficient statistics generated by the legal record–verifiability, contamination, shielding, learning, enterprise pass-through, or market power–rather than replacing the multidimensional transfer problem with a universal one-parameter doctrine.

4.1 Algorithmic defensive design and endogenous records

Throughout, Z continues to denote the liability-relevant medical record introduced in Section 2. In the fixed-record benchmark, its distribution was taken as given. In this section I endogenize that distribution: the record generated by the encounter depends on AI safety effort, patient inputs and adherence, clinical integration, algorithmic defensive design, and the learning state.

There is a finite set of case types $k \in \mathcal{K}$. (For a fixed case type, suppress k until needed.) I distinguish traditional physician defensive medicine from AI-specific defensive design. Traditional physician defensive medicine can include both encounter-level actions, such as ordering an additional test, referral, or admission, and ex ante practice choices, such as templates, protocols, or avoidance of risky cases. The AI-specific feature is that defensive design is embedded in the product, applied at scale, and can shape the record itself. It changes the admissible input set, warning thresholds, abstention rules, referral prompts, prescription defaults, documentation language, and eligibility filters. Write

$$b = (b^{reject}, b^{warn}, b^{refer}, b^{test}, b^{rx}, b^{log}). \tag{4.1}$$

The AI developer chooses true safety effort e and defensive design b. The patient chooses information and adherence effort a. A clinical enterprise, if present, chooses integration effort m: verification of inputs, tests, prescribing, follow-up, monitoring, and escalation. The state n is the learning stock. Harm probability is

$$p(e, a, m, b, n), \tag{4.2}$$

where the cross-partials may be nonzero.

The key distinction is that b can change both the probability of harm and the distribution or legal meaning of the record. Let

$$Z \sim F(\cdot \mid e, a, m, b, n), \qquad \mathcal{G} = \mathcal{G}(b), \tag{4.3}$$

where $F(\cdot \mid e, a, m, b, n)$ is the conditional law of the medical record and $\mathcal{G}(b)$ is the legally usable partition generated by the record. A clinically useful design raises the informativeness of the record about the scores in Proposition 1. A shielding design raises formal compliance while lowering the covariance between the legal transfer and true safety effort. For example, hard-coding the AI to reject ambiguous inputs may reduce compensable claims by excluding risky records; it may also worsen health by delaying care and reduce learning by removing hard cases from the data.

In the notation of the core model, defensive design changes two objects at once. First, because $\mathcal{G} = \mathcal{G}(b)$, it changes the feasible moment set $\mathcal{M}(\mathcal{G}(b), L)$: the law can condition on different cells after the record has been strategically generated. Second, because b changes the distribution of $D = (Z, Y)$, it changes the conditional legal score $\bar{S}_{\mathcal{G}(b)} = \mathbb{E}[S_0 \mid \mathcal{G}(b), Y = 1]$. A useful record-design action expands $\mathcal{M}$ in directions that separate AI

marginal responsibility from patient behavior. A jamming action may expand the formal sigma-field while rotating or shrinking the useful moment set, so that legal payments covary less with true safety and more with formal compliance. This is why endogenous records must be evaluated through the feasible moment set and the legal score, not by the number of fields in the chart.

This is where the exogenous-record results from Section 2 stop applying mechanically. If a regulator mandates an independent audit log, the sigma-field expands exogenously and Proposition 3 applies. If the AI chooses the log language, eligibility rule, or warning threshold in response to liability, the record itself is a strategic object. The planner must then price not only clinical safety effort but also record design.

Let $T^D(e, a, m, b, n)$ be the expected transfer, premium, indemnity, or internal contractual payment borne by the developer. Let $T^H(e, a, m, b, n)$ be the corresponding expected payment borne by the clinical enterprise. Let $T^P(e, a, m, b, n)$ be expected compensation received by the patient. In a simple direct-developer-liability regime these objects may coincide up to signs. In an enterprise regime they generally differ because of internal contracts and insurance premia.

Theorem 4 (General implementability)**.** *Suppose (e^*, a^*, m^*, b^*) is an interior first best for a fixed state n. If the developer chooses (e, b), the patient chooses a, and the clinical enterprise chooses m, then any differentiable liability and internal-contracting system that implements the first best must satisfy, at the first best,*

$$T_e^D = Lp_e, \tag{4.4}$$

$$(T_b^D)_\ell = L(p_b)_\ell \quad \textit{for each component } \ell, \tag{4.5}$$

$$T_m^H = Lp_m, \tag{4.6}$$

$$T_a^P = 0. \tag{4.7}$$

Conversely, if these equalities hold and private objective functions are convex in the relevant actions, then the first best is privately optimal.[10] *If patient-facing liability is borne only by the clinical enterprise and there are no internal contracts, indemnities, premium pass-throughs, or direct developer liability, then $T_e^D = T_b^D = 0$; developer effort and defensive design are correctly incentivized only in the degenerate case $p_e = p_b = 0$.*

The theorem fixes the assignment issue that is easy to miss in informal discussions

[10]The first-best profile is a Nash equilibrium of the induced action game. Under a contraction or diagonal-dominance condition on the stacked best-response map, this equilibrium is locally unique.

of enterprise liability.[11] A hospital can face the patient, but that does not by itself make the developer internalize model-training, calibration, or interface-design incentives. Patient-facing enterprise liability works only if the enterprise controls those margins or can contract for them internally. Otherwise direct developer liability, indemnity, audit rights, mandatory insurance, or premium pass-through is needed.

The condition on b is the AI-specific defensive-design condition. If a design choice lowers legal exposure but has no clinical value, then $(p_b)_\ell = 0$ and the optimal contract requires $(T_b^D)_\ell = 0$. If the design worsens care, then $(p_b)_\ell > 0$ and the contract should penalize it. Liability should not reward actions that make the record look safer to a court or insurer without making the patient safer.

A record-jamming example. The next example makes the endogenous-record problem stark. It shows why a formally richer record can reduce welfare when the additional record is strategically generated by the AI developer. The example suppresses patient effort, enterprise effort, and learning in order to isolate the interaction between true safety and record design.

True medical harm is

$$p(e) = \bar{p} - e, \tag{4.8}$$

where $e \geq 0$ is clinical safety effort. The developer also chooses a record-design action $b \geq 0$. This action raises formal legal compliance, but it does not improve clinical safety. Let the developer's costs be

$$C(e) = \frac{e^2}{2\alpha}, \qquad B(b) = \frac{b^2}{2\chi}, \qquad \alpha, \chi > 0. \tag{4.9}$$

Dropping the constant loss term $L\bar{p}$, social welfare is

$$W(e, b) = Le - \frac{e^2}{2\alpha} - \frac{b^2}{2\chi}. \tag{4.10}$$

Thus the social first best is $e^{FB} = \alpha L$ and $b^{FB} = 0$, assuming the action space contains

[11] As in the benchmark, the implementability statement is a target-equilibrium statement. In the multi-agent environment, let $x = (e, b, a, m)$ denote the stacked action profile and let $B(x)$ be the stacked best-response map induced by the liability rule, liability split, internal contracts, and patient compensation rule. If B is a contraction in a neighborhood of $x^* = (e^*, b^*, a^*, m^*)$, or if an equivalent diagonal-dominance condition holds for the Jacobian of private first-order conditions, then the implemented equilibrium is locally unique. Without that condition, the theorem asserts that x^* is a Nash equilibrium satisfying the desired marginal-incentive equations, not that the induced game has no other equilibria.

this point.

Now compare two record regimes. Under a self-generated rich record, the legal record contains a formal compliance score $H = b$. A higher b makes the record look better to the court or insurer, but it also lowers the true clinical informativeness of the record. Represent this by an informativeness term

$$I(b) = 1 - \rho b, \qquad \rho > 0, \tag{4.11}$$

with the action space restricted so that $I(b) \geq 0$. Expected liability under a high-powered self-generated-record rule is

$$T^S(e, b) = L\{\bar{p} - I(b)e - \phi b\}, \qquad \phi > 0. \tag{4.12}$$

The term $I(b)e$ is the legal system's effective credit for true safety effort. The term ϕb is pure legal shielding: it lowers expected liability because the formal record looks compliant, not because patients are safer.

Under an independent audit-log regime, the law conditions on an audit signal that separates clinical safety from self-generated formal compliance. In reduced form, expected liability is

$$T^A(e, b) = L(\bar{p} - e). \tag{4.13}$$

The audit record may still contain formal compliance language, but the liability rule does not let that language substitute for clinical safety.

Proposition 8 (Strategic record expansion can reduce welfare)**.** *Assume the developer's private problem under the self-generated-record regime is strictly convex, for example*

$$\frac{1}{\alpha\chi} > L^2\rho^2. \tag{4.14}$$

Under the audit-log regime (4.13), high-powered liability induces $e^A = \alpha L$ and $b^A = 0$, which is the social first best. Under the self-generated-record regime (4.12), if

$$\phi > \rho\alpha L, \tag{4.15}$$

then the developer chooses $b^S > 0$ and $e^S < e^A$. Hence

$$W(e^S, b^S) < W(e^A, b^A). \tag{4.16}$$

Thus a richer-looking record can lower welfare when the additional record component is strategically generated. A record-regulation or audit-log mandate restores separation by making liability depend on clinical safety rather than on self-generated formal compliance.

The proposition is not a claim that more information is bad. It is a claim that self-authored compliance evidence is not the same object as independent auditability. When the AI controls the legal meaning of the record, the extra record component can become a jamming technology. An audit-log mandate changes the record technology: it preserves information about clinical safety while preventing formal compliance from substituting for safety. In the notation of Theorem 4, the audit-log regime restores the desired separation $T_e^D = Lp_e$ and $T_b^D = Lp_b = 0$ in the case where record design has no clinical value.

A linear-quadratic version gives a closed-form expression. For the closed-form index, collapse the vector b to a one-dimensional defensive-design index; the vector case replaces $\chi\phi^2$ with the corresponding quadratic form. Let

$$p^A = \bar{p}^A - e, \qquad p^P = \bar{p}^P - a, \qquad p^B = 0, \tag{4.17}$$

with costs

$$C(e) = \frac{e^2}{2\alpha}, \qquad G(a) = \frac{a^2}{2\beta}, \qquad B(b) = \frac{b^2}{2\chi}. \tag{4.18}$$

The reduced-form liability technology is

$$T^\lambda = \lambda L\{\sigma(b)(\bar{p}^A - e) + \eta(b)(\bar{p}^P - a) + \nu(b)p^0 - \phi b\}. \tag{4.19}$$

The functions $\sigma(b)$, $\eta(b)$, and $\nu(b)$ capture endogenous record informativeness and contamination. The term ϕb captures pure legal shielding: a reduction in compensable claims that does not reduce true medical risk. The clean index below is a local pure-shielding approximation around a reference allocation (e_0, a_0, b_0), with $\sigma_b = \eta_b = \nu_b = 0$ at that point, so that the marginal legal-shielding effect is ϕ. If b also changes σ, η, ν, its marginal effect on compensable claims depends on the action levels and can also change the marginal return to true safety effort. In that general case, the scalar formula should be read as a local approximation after relinearizing the liability technology around the reference allocation, not as an exact one-for-one replacement.

Let K and R be as in (3.19) and (3.20).

Proposition 9 (Liability index with algorithmic defensive design)**.** *In the linear-quadratic*

pure-shielding approximation (4.17)–(4.19)*, the optimal liability share is*

$$\lambda^* = \left[\frac{L^2\alpha\sigma + r_1 - k_1}{L^2(\alpha\sigma^2 + \xi^2\beta\eta^2 + \chi\phi^2) + r_2 + k_2}\right]_0^1. \tag{4.20}$$

If b also changes record informativeness, the same expression remains a local index only after replacing the pure shielding coefficient by the local compensable-claims derivative at the reference allocation,

$$\widetilde{\phi}(e_0, a_0, b_0) = -\frac{\partial}{\partial b}\{\sigma(b)(\bar{p}^A - e_0) + \eta(b)(\bar{p}^P - a_0) + \nu(b)p^0 - \phi b\}\Big|_{b=b_0}. \tag{4.21}$$

If b changes the marginal safety loading $\sigma(b)$, the numerator term $L^2\alpha\sigma$ and the curvature term $L^2\alpha\sigma^2$ should also be evaluated at the same reference point. Any direct clinical harm from p_b enters as an additional marginal welfare cost of liability-induced record design.

Relative to the simple model, algorithmic defensive design adds a curvature term. Liability is less attractive when designers can cheaply reduce liability through warnings, referrals, prescriptions, eligibility rules, or documentation that do not reduce true risk. The formula also shows the substitution effect between verifiable formal compliance and true safety effort: if b raises the formal compliance score while lowering the marginal informativeness of Z about e, the relinearized shielding-and-jamming term can be larger than pure documentation cost alone, and the safety-loading terms must be evaluated at the new local record technology.

Also, the defensive-design term has a simple interpretation. Let

$$A = L^2\alpha\sigma + r_1 - k_1$$

and let

$$D_0 = L^2(\alpha\sigma^2 + \xi^2\beta\eta^2) + r_2 + k_2$$

be the denominator without algorithmic defensive design. Then, away from the bounds,

$$\lambda^D = \frac{A}{D_0 + L^2\chi\phi^2} = \frac{A/D_0}{1 + L^2\chi\phi^2/D_0}.$$

Thus pure legal shielding acts like a discount factor on optimal liability power. The reason is not that the AI becomes clinically safer. The reason is that liability now rewards a second hidden action: making the record look less liable. In a primary-care triage example, the AI may respond to liability by rejecting ambiguous patients, issuing generic emergency

warnings, or documenting formal compliance rather than eliciting useful symptoms. These actions reduce expected claims without necessarily improving care.

This also clarifies why liability and record regulation are complements. A record mandate or audit-log rule that makes shielding less productive lowers ϕ. Lower ϕ raises the optimal liability power because liability is then more likely to buy true clinical safety rather than legal insulation.

4.2 Dynamic adoption and Bayesian learning across patients

The reduced-form learning stock can be microfounded as Bayesian updating. For each condition j, the AI has an unknown clinical-performance parameter θ_j. At date t, the posterior belief is normal with precision n_{jt}. An encounter for condition k can produce a verifiable clinical outcome signal about condition j,

$$X_{ijkt} = \theta_j + \varepsilon_{ijkt}, \qquad \varepsilon_{ijkt} \sim N\left(0, \{\Xi_{jk}\tau_k(\lambda_k, m_k, b_k)\varpi_k(\lambda_k, m_k, b_k)\}^{-1}\right), \tag{4.22}$$

where $\Xi_{jk} \geq 0$ is the cross-condition informativeness of a condition-k encounter for learning about condition j, and τ_k is the endogenous data quality of the encounter. Let $\varpi_k(\lambda_k, m_k, b_k) \in [0, 1]$ denote representativeness of the selected encounters for the target population. Liability can raise τ_k by requiring logging, follow-up, and verification. It can lower τ_k or ϖ_k by inducing refusal of hard cases, excessive referral, or litigation-oriented documentation. Thus learning depends on the quantity of use, the informativeness of each use, and the representativeness of selected patients. If $\Xi_{jk}\tau_k\varpi_k = 0$, the encounter carries no usable signal about θ_j; otherwise the displayed normal experiment applies.

Let $\delta_n \in [0, 1]$ denote depreciation of the effective learning stock, and let $\delta \in (0, 1)$ denote the planner's intertemporal discount factor. Write $n_t = (n_{jt})_{j\in\mathcal{K}}$. If a mass $\mu_k x_k(\lambda_k, n_t)$ of condition-k patients adopts the AI at date t, conjugate updating implies

$$n_{j,t+1} = (1 - \delta_n)n_{jt} + \sum_{k\in\mathcal{K}} \Xi_{jk}\mu_k x_k(\lambda_k, n_t)\tau_k(\lambda_k, m_k, b_k)\varpi_k(\lambda_k, m_k, b_k). \tag{4.23}$$

Thus the spillover matrix in the dynamic model is not a black box. It is the precision contribution of one class of clinical outcomes to another class of model parameters.

Proposition 10 (Bayesian microfoundation of learning)**.** *Under the normal signal structure* (4.22)*, posterior precision evolves according to* (4.23)*. If expected harm for condition j is increasing in posterior variance $1/n_j$, or equivalently if social surplus is increasing in*

posterior precision n_j, then the social value of an additional verified encounter is increasing in its precision contribution $\Xi_{jk}\tau_k\varpi_k$ and in the shadow value of precision $V_j(n)$.

Let $\mathcal{S}_k(\lambda_k, n_t)$ be static social surplus per adopted AI encounter for condition k, after equilibrium actions have been chosen. The planner solves

$$V(n_t) = \max_{\lambda\in[0,1]^{|\mathcal{K}|}} \left\{ \sum_{k\in\mathcal{K}} \mu_k x_k(\lambda_k, n_t)\mathcal{S}_k(\lambda_k, n_t) + \delta V(n_{t+1}) \right\}, \tag{4.24}$$

with n_{t+1} given by (4.23).

Theorem 5 (Dynamic liability condition)**.** *Suppose λ_k is interior and differentiability holds. Assume that λ_k affects adoption, encounter quality, and representativeness only for condition k; if liability in one condition affects adoption or data quality in another, the first-order condition below includes the corresponding cross-derivative terms. Then the optimal liability share for condition k satisfies*

$$\begin{aligned} 0 = {} & \mu_k \left[x_{k\lambda}\mathcal{S}_k + x_k\mathcal{S}_{k\lambda} \right] \\ & + \delta \sum_{j\in\mathcal{K}} V_j(n_{t+1})\Xi_{jk}\mu_k \left[x_{k\lambda}\tau_k\varpi_k + x_k\tau_{k\lambda}\varpi_k + x_k\tau_k\varpi_{k\lambda} \right], \end{aligned} \tag{4.25}$$

where all derivatives are evaluated at the optimum and V_j denotes the derivative of V with respect to n_j.

The first line is the static accident, risk-bearing, accountability, and activity-level effect. The second line is the Bayesian learning externality. It is positive when liability brings in valuable marginal encounters or improves the precision or representativeness of data those encounters produce. It is negative when liability induces refusal, over-referral, or documentation that reduces clinical learning. In a local quadratic approximation, the dynamic term enters the numerator of the liability index as a Pigouvian adjustment.

Corollary 3 (Local dynamic index)**.** *Suppose the static part of welfare for condition k has the quadratic form in Proposition 9. Let Δ_k denote the marginal dynamic learning value of raising λ_k at the point of approximation. Then the local optimal liability share is*

$$\lambda_k^* = \left[\frac{L_k^2\alpha_k\sigma_k + r_{1k} - k_{1k} + \Delta_k}{L_k^2(\alpha_k\sigma_k^2 + \xi_k^2\beta_k\eta_k^2 + \chi_k\phi_k^2) + r_{2k} + k_{2k}} \right]_0^1. \tag{4.26}$$

The sign of Δ_k determines whether dynamic learning strengthens or weakens static liability. Here ξ_k is the condition-specific analogue of the patient-compensation salience

parameter in the static index. This is the formal sense in which liability can either subsidize learning by creating trustworthy use and high-quality records, or tax learning by inducing high prices, narrow eligibility, and defensive behavior.

The dynamic index shows that liability is not only an accident-price instrument. For AI doctors, it can also be an investment instrument. Let

$$A_k = L_k^2 \alpha_k \sigma_k + r_{1k} - k_{1k}, \qquad D_k = L_k^2(\alpha_k \sigma_k^2 + \xi_k^2 \beta_k \eta_k^2 + \chi_k \phi_k^2) + r_{2k} + k_{2k}.$$

Then

$$\lambda_k^* = \left[\frac{A_k + \Delta_k}{D_k}\right]_0^1.$$

The dynamic term Δ_k is positive when liability increases trusted use, verified outcomes, and representative data. It is negative when liability causes the AI to reject hard patients, over-refer ambiguous cases, or generate records that are legally useful but clinically uninformative.

The projection onto $[0, 1]$ gives a simple regime interpretation:

$$A_k + \Delta_k \leq 0 \quad \Rightarrow \quad \lambda_k^* = 0,$$

$$0 < A_k + \Delta_k < D_k \quad \Rightarrow \quad 0 < \lambda_k^* < 1,$$

and

$$A_k + \Delta_k \geq D_k \quad \Rightarrow \quad \lambda_k^* = 1.$$

Thus smooth changes in the learning value of liability can generate sharp changes in the optimal legal form. Early in deployment, when liability draws users and produces valuable evidence, Δ_k may push the system toward a high-powered warranty. Later, as adoption saturates or as the marginal value of additional data falls, the same condition may move the system back toward targeted fault-based liability. The discontinuity is in the legal instrument, not necessarily in patient adoption.

4.3 Learning traps and liability signaling

The dynamic index above is local. It says whether a marginal increase in liability raises or lowers welfare at a given state. This subsection gives a sharper prediction: liability can generate path dependence. The reason is a timing complementarity. Liability affects adoption and record production today, but the quality improvement created by learning appears only in future periods. Patient-facing liability or warranty coverage can raise

adoption because patients value compensation and may interpret a strong warranty as a signal that the AI firm has confidence in its own system. But the same legal pressure can reduce future quality if it induces exclusion, over-referral, or defensive documentation. The same AI technology can therefore support both a low-quality, low-adoption regime and a high-quality, high-adoption regime.

Section 4.2 used the vector of condition-specific learning stocks $n_t = (n_{jt})_{j \in \mathcal{K}}$. To keep the learning-trap result transparent, this subsection works with a one-dimensional quality state $n \in [\underline{n}, \bar{n}]$. This can be read either as the learning stock for a fixed condition, or as an index $n = \iota(n_t)$ summarizing the clinically relevant component of the vector state for the legal regime under study. Nothing in the argument depends on the scalar state being literally one disease; the scalar notation isolates the complementarity between liability, adoption, record quality, and future AI quality.

I first give a signaling-game foundation for the legal objective used below.[12] The foundation is deliberately compact because the learning-trap result itself is a monotone fixed-point result.

Fix a quality state $n \in [\underline{n}, \bar{n}]$. There are two firm types $\theta \in \{H, L\}$, privately observed by the AI firm, with prior $\pi_0(n) = \Pr(\theta = H \mid n)$. The high type has lower expected claim costs and produces higher patient surplus in the sense made precise by the payoff functions below. A firm chooses a public pair $(\ell, b) \in \mathcal{L} \times B$, where ℓ is the warranty or liability-power component and b is record design. The set $\mathcal{L}$ is a finite chain ordered by AI-fault liability power, with associated warranty power λ_ℓ. The record-design action b can create useful audit information or formal legal shielding. Patients observe (ℓ, b), form a posterior belief $\pi(\ell, b, n) = \Pr(\theta = H \mid \ell, b, n)$, and then a mass

$$x(n, \ell, b, \pi)$$

adopts the AI. The adoption function is smooth in n and in beliefs; it may be higher under a stronger warranty either because compensation is valuable directly or because the warranty raises the posterior probability that the firm is high quality.

Type θ's one-period payoff from the public pair (ℓ, b) at belief π is

$$U_\theta^F(n, \ell, b, \pi) = P\, x(n, \ell, b, \pi) - C_\theta(n, \ell, b), \tag{4.27}$$

where P is the per-adopter margin and C_θ includes expected warranty payments, expected

[12] The point is that the reduced objective $\Pi(n, \ell)$ need not be an unmotivated primitive: it can be the equilibrium value induced by a warranty-signaling game with endogenous record design.

liability, insurance premia, and record-design costs. A perfect Bayesian equilibrium of the signaling subgame at state n is a pair of type-contingent public choices $(\ell_\theta(n), b_\theta(n))$ and a posterior system $\pi(\ell, b, n)$ such that each type maximizes (4.27) given beliefs and beliefs are Bayesian on the equilibrium path.

For an institution that can restrict or select a legal regime ℓ, let $b_\theta^\ell(n)$ denote the type-θ equilibrium record design induced by the signaling subgame under that regime, and let $\pi_\theta^\ell(n)$ denote the on-path posterior belief held by patients after observing the type-θ public pair $(\ell, b_\theta^\ell(n))$. The reduced legal objective is the expected equilibrium value

$$\begin{aligned}\Pi(n,\ell) = \sum_{\theta\in\{H,L\}} \pi_0(\theta \mid n)\Big[&x(n,\ell,b_\theta^\ell(n),\pi_\theta^\ell(n))S_\theta(n) \\ &- K_\theta(n,\ell,b_\theta^\ell(n)) + \delta V(\Phi_\theta(n,\ell,b_\theta^\ell(n)))\Big].\end{aligned} \tag{4.28}$$

Here $S_\theta(n)$ is the social surplus per adopter net of pure transfers, K_θ is the real legal, insurance, and record-design cost of the regime, and Φ_θ is the next-period quality generated by adoption and record design. If the legal regime is instead a voluntary warranty chosen by the firm, the same notation applies with Π interpreted as the relevant type's equilibrium payoff or as expected equilibrium surplus. Thus Π is a compact representation of the outcome of the signaling game, not an additional primitive.

A standard single-crossing condition links the signaling game to monotone regime choice. If, for higher warranty signals $\ell' \succ \ell$, the incremental expected warranty and liability cost is lower for the high type,

$$C_H(n,\ell',b') - C_H(n,\ell,b) \le C_L(n,\ell',b') - C_L(n,\ell,b), \tag{4.29}$$

and if patient adoption is increasing in posterior quality, then stronger warranties can be credible signals of quality in a monotone perfect Bayesian equilibrium. The paper does not need a unique signaling equilibrium. It only uses the equilibrium object $\Pi(n,\ell)$ generated by a selected monotone equilibrium of this signaling subgame. The next lemma states sufficient conditions under which that object has the increasing-differences property needed for the learning-trap result.

Let

$$\Lambda(n) = \arg\max_{\ell\in\mathcal{L}} \Pi(n,\ell) \tag{4.30}$$

be the set of locally optimal legal regimes. A stationary liability-learning equilibrium is

a pair (n^*, ℓ^*) such that

$$\ell^* \in \Lambda(n^*), \qquad n^* = \Phi(n^*, \ell^*). \tag{4.31}$$

The multiplicity result below should be interpreted as a stationary-policy and equilibrium-selection result. It need not say that a fully commitment-capable regulator solving the Bellman problem in Section 4.2 would choose to remain in the low-quality steady state if it could commit to state-contingent future rules and force a different transition path. Rather, it captures the trap faced by decentralized warranty signaling or by an institution using a stationary legal selector when current quality determines whether strong accountability is credible, politically feasible, or privately adopted. The low-quality equilibrium is a welfare concern because the same technology may support a higher-quality stationary regime, not because every planner is mechanically trapped by the Bellman equation.

For the monotone-dynamics result, write the reduced objective in the form

$$\Pi(n, \ell) = X(n, \ell)S(n) + R(n, \ell) - K(n, \ell) + \delta V(\Phi(n, \ell)), \tag{4.32}$$

where $X(n, \ell)$ is equilibrium adoption induced by the signaling subgame, $R(n, \ell)$ is any additional accountability or warranty-confidence value, $K(n, \ell)$ is the real legal, insurance, and record-design cost of the regime, and V is the continuation value of future AI quality. More explicitly,

$$X(n, \ell) = \sum_{\theta \in \{H,L\}} \pi_0(\theta \mid n) x(n, \ell, b_\theta^\ell(n), \pi_\theta^\ell(n))$$

is the type-averaged adoption level after substituting the equilibrium record design and on-path posterior belief induced by regime ℓ. Thus the notation nests (4.28) by averaging over types and equilibrium record designs.

To avoid overloading the primitive objects in Section 4.2, let $\tau(n, \ell)$, $\varpi(n, \ell)$, and $j(n, \ell)$ denote the reduced regime-level data quality, representativeness, and record-jamming loss induced by the signaling subgame at state n. They are summaries of the underlying objects such as $\tau_k(\lambda_k, m_k, b_k)$ and $\varpi_k(\lambda_k, m_k, b_k)$ after the legal regime, record design, and equilibrium beliefs have been substituted in. Write

$$\mathcal{Q}(n, \ell) = \mu X(n, \ell)\tau(n, \ell)\varpi(n, \ell) - j(n, \ell) \tag{4.33}$$

for the net learning content of adoption under regime ℓ, so that

$$\Phi(n, \ell) = (1 - \delta_n)n + \mathcal{Q}(n, \ell).$$

The high-liability regime may reduce net learning at low quality because it induces exclusion and record jamming, but improve net learning at high quality because the additional adoption produces clinically informative data. The sufficient conditions below allow this case.

Lemma 2 (Primitive sufficient conditions for monotone liability-learning feedback)**.** *Let $\mathcal{L}$ be a finite chain. For every adjacent pair of regimes $\ell' \succ \ell$, assume:*

(i) $S(n) \geq 0$ *is nondecreasing, and* $X(n,\ell') - X(n,\ell)$ *is nonnegative and nondecreasing in* n*. Thus the adoption gain from stronger liability is larger when quality is higher.*

(ii) $R(n,\ell') - R(n,\ell)$ *is nondecreasing in* n*. Thus stronger liability is a more credible warranty or accountability signal at higher quality.*

(iii) $K(n,\ell') - K(n,\ell)$ *is nonincreasing in* n*. Thus the incremental legal, insurance, or defensive-design cost of stronger liability is not larger at higher quality.*

(iv) The continuation-value difference

$$V(\Phi(n,\ell')) - V(\Phi(n,\ell)) \tag{4.34}$$

is nondecreasing in n*. This direct condition replaces any need to assume global convexity of* V*.*

Then $\Pi(n,\ell)$ *has increasing differences in* (n,ℓ)*.*

Moreover, suppose each fixed-regime quality map is nondecreasing:

$$(1-\delta_n) + \frac{\partial}{\partial n}\mathcal{Q}(n,\ell) = (1-\delta_n) + \mu\frac{\partial}{\partial n}\{X(n,\ell)\tau(n,\ell)\varpi(n,\ell)\} - \frac{\partial j(n,\ell)}{\partial n} \geq 0. \tag{4.35}$$

Assume also that $\mathcal{Q}(n,\ell') - \mathcal{Q}(n,\ell)$ *is nondecreasing in* n*. It may be negative at low* n *and positive at high* n*. If a nondecreasing regime selection* $\ell^*(n)$ *switches upward from* ℓ *to* ℓ' *only at states where*

$$\mathcal{Q}(n,\ell') \geq \mathcal{Q}(n,\ell), \tag{4.36}$$

then the induced map $\Phi^*(n) = \Phi(n,\ell^*(n))$ *is nondecreasing. A simple sufficient condition for (4.36) is that the high-liability regime is selected only after the increasing difference* $\mathcal{Q}(n,\ell') - \mathcal{Q}(n,\ell)$ *has crossed zero.*

Condition (4.35) is the precise sense in which record jamming must not rise too sharply with quality within a fixed regime. Condition (4.36) is weaker than assuming stronger

liability always increases net learning content. It permits high liability to be destructive at low quality and productive at high quality; it only rules out a downward jump in the quality transition at the states where the legal regime actually switches upward.

Proposition 11 (Learning traps under monotone liability-learning feedback)**.** *Order legal regimes by AI-fault liability power. Suppose* $\Pi(n, \ell)$ *has increasing differences in* (n, ℓ)*, and let* $\ell^*(n)$ *be the largest selection from* $\Lambda(n)$*. Define the induced quality map*

$$\Phi^*(n) = \Phi(n, \ell^*(n)). \tag{4.37}$$

Suppose Φ^* *is nondecreasing. If there exist* $n_1 < n_2 < n_3 < n_4$ *such that*

$$\Phi^*(n_1) \geq n_1, \qquad \Phi^*(n_2) \leq n_2, \qquad \Phi^*(n_3) \geq n_3, \qquad \Phi^*(n_4) \leq n_4, \tag{4.38}$$

then there are at least two stationary liability-learning equilibria: one with quality in $[n_1, n_2]$ *and one with quality in* $[n_3, n_4]$*. Moreover,* $[n_1, n_2]$ *and* $[n_3, n_4]$ *are forward invariant under* Φ^**. Hence an economy that starts in the lower interval remains in the lower-quality region under the stationary liability rule. If, in addition,* Φ^* *is a contraction on each interval, then every path starting in that interval converges to a stationary quality in that interval.*

Under the largest optimal regime selection, if aggregate adoption $X(n, \ell)$ *is increasing in* n *and in the liability order on* ℓ*, and if* λ_ℓ *is increasing in that order, then the higher-quality equilibrium has weakly higher adoption and weakly higher AI-fault liability power, with strict inequalities whenever the selected regime or adoption response differs strictly across the two fixed points.*

This proposition is a dynamic complementarity result. Higher quality makes stronger liability more attractive because liability-backed adoption and verified data are more valuable when the AI is clinically useful. Stronger liability can then increase adoption, which generates more learning and further raises quality. But the same liability can tax learning if it induces exclusion or defensive documentation. The prediction is not a discontinuity in primitive patient demand: adoption can be smooth in both quality and liability. The discontinuity comes from the discrete legal-regime choice. Smooth adoption and linear learning are sufficient because the legal instrument is selected from a discrete set. AI medical liability can therefore have multiple stationary quality-adoption-liability regimes, not merely a single optimal damages level.

The general proposition is useful, but the mechanism is even clearer in a fully explicit

two-regime example. The point of the example is not numerical calibration. It is to show that no discontinuity in primitive adoption demand is needed. Smooth adoption and linear learning are enough to generate two long-run regimes once the legal regime is chosen endogenously.

Example 1 (Two-regime learning trap with smooth adoption)**.** Let there be two legal regimes, a low-warranty regime ℓ_L and a high-warranty regime ℓ_H, with $\lambda_L < \lambda_H$. Quality is $n \in [0, 1]$. Adoption is smooth in quality under each regime:

$$x(n, \ell_L) = a_0 + a_1 n, \qquad x(n, \ell_H) = a_0 + a_H + a_1 n, \tag{4.39}$$

where $a_0, a_1, a_H > 0$. For the calculations below, assume the no-truncation condition $a_0 + a_H + a_1 n \leq 1$ and $a_0 + a_1 n \leq 1$ on the relevant state interval; otherwise the adoption premium is the capped difference rather than a_H. The high warranty raises adoption by a_H, either because patients value coverage directly or because the warranty signals confidence. Let surplus per adopted user be $S(n) = vn$, with $v > 0$, and let the incremental legal and insurance cost of the high warranty be $c_H > 0$. The reduced legal objective difference is

$$\Pi(n, \ell_H) - \Pi(n, \ell_L) = a_H v n - c_H. \tag{4.40}$$

For this example, take the regime selector's incremental objective to be the static warranty-adoption value net of legal and insurance cost, abstracting from the continuation-value difference in the selector. Equivalently, one may interpret c_H as net of a continuation-value difference only in the special case where that difference is constant over the relevant range. The purpose is to isolate the adoption-learning feedback in the transition equation. The high-warranty regime is selected if and only if

$$n \geq n^\dagger \equiv \frac{c_H}{a_H v}. \tag{4.41}$$

The threshold is endogenous to the primitive adoption premium, surplus from quality, and legal cost of the high warranty.

Learning is also linear. Let $q_L, q_H > 0$ denote the clinical learning content per adopted user in the two regimes, and let $j_L, j_H \geq 0$ denote record-jamming, exclusion, or defensive-documentation losses. Quality evolves according to

$$n_{t+1} = (1 - \delta_n) n_t + \mu q_\ell x(n_t, \ell) - j_\ell, \qquad \ell \in \{\ell_L, \ell_H\}. \tag{4.42}$$

Assume $\delta_n - \mu q_\ell a_1 > 0$ for $\ell \in \{\ell_L, \ell_H\}$. Conditional on regime ℓ_L, the unique fixed point is

$$n_L^* = \frac{\mu q_L a_0 - j_L}{\delta_n - \mu q_L a_1}, \tag{4.43}$$

while conditional on regime ℓ_H, the unique fixed point is

$$n_H^* = \frac{\mu q_H (a_0 + a_H) - j_H}{\delta_n - \mu q_H a_1}. \tag{4.44}$$

If

$$0 \leq n_L^* < n^\dagger < n_H^* \leq 1, \tag{4.45}$$

then both (n_L^*, ℓ_L) and (n_H^*, ℓ_H) are stationary liability-learning equilibria. The low-quality fixed point rationalizes the low-warranty regime, and the high-quality fixed point rationalizes the high-warranty regime.

For a numerical instance, take

$$\begin{array}{lllllll} a_0 = .10, & a_1 = .20, & a_H = .30, & \mu = 1, & \delta_n = .70, & & \\ q_L = .40, & q_H = .80, & j_L = 0, & j_H = .05, & v = 1, & c_H = .09. & \end{array}$$

Then $n^\dagger = c_H/(a_H v) = .30$. The two regime-specific maps are

$$\Phi_L(n) = .38n + .04, \qquad \Phi_H(n) = .46n + .27.$$

Both fixed points are locally stable because the slopes of the regime-specific transition maps are below one. Hence

$$n_L^* = \frac{.04}{.62} \approx .065 < n^\dagger, \qquad n_H^* = \frac{.27}{.54} = .50 > n^\dagger.$$

Adoption is $x(n_L^*, \ell_L) \approx .113$ in the low-warranty equilibrium and $x(n_H^*, \ell_H) = .50$ in the high-warranty equilibrium. Both adoption functions in (4.39) are smooth in quality. The multiplicity comes from the feedback loop: quality determines which warranty is credible or valuable, the warranty changes adoption, adoption changes learning, and learning changes future quality.

The same primitives can allow high-warranty liability to reduce net learning at low quality and increase it at high quality. For example, replace the constant j_H by a decreas-

ing record-jamming loss $j_H(n) = \max\{\bar{j}_H - \eta_j n, 0\}$. Then

$$\mathcal{Q}(n, \ell_H) - \mathcal{Q}(n, \ell_L) = \mu q_H(a_0 + a_H + a_1 n) - j_H(n) - \mu q_L(a_0 + a_1 n) + j_L$$

can be negative for low n and positive for high n. The trap theorem does not require high liability to dominate at every state; it requires monotonicity of the induced transition at the states where the high-warranty regime is actually selected.

The example also shows how record regulation acts in the dynamic model. Holding the legal threshold fixed, an audit-log mandate that lowers j_H or raises q_H raises n^*_H. Conversely, if high-warranty liability induces enough record jamming that j_H rises or q_H falls, the high-quality fixed point can disappear. The relevant policy object is therefore not high liability by itself, but high liability combined with records that preserve the clinical learning content of adoption.

The next corollary gives the policy role of record regulation. Let r index the strength of independent audit logs, record regulation, or limits on exclusion. Write Φ^*_r for the induced quality map after the largest optimal regime is selected under regulation level r. A higher r may raise data quality τ, raise representativeness ϖ, or lower record-jamming loss j.

Corollary 4 (Record regulation shifts stationary quality upward)**.** *For each r, suppose $\Phi^*_r : [\underline{n}, \bar{n}] \to [\underline{n}, \bar{n}]$ is nondecreasing. If $r' > r$ implies*

$$\Phi^*_{r'}(n) \geq \Phi^*_r(n) \qquad \text{for all } n \in [\underline{n}, \bar{n}], \tag{4.46}$$

then the least and greatest stationary qualities of the largest-selection map are weakly increasing in r. A lower trap interval means an interval such as $[n_1, n_2]$ in Proposition 11: a low-quality interval satisfying the first two crossing inequalities in (4.38). If $\Phi^*_{r'}(n) > n$ for every n in such an interval, then no fixed point of the largest-selection map remains in that interval. To rule out any stationary liability-learning equilibrium under some non-largest also-optimal regime, strengthen the condition to require $\Phi(n, \ell) > n$ for every n in the interval and every $\ell \in \Lambda_{r'}(n)$.

The policy implication is not that transitional no-fault compensation is always the right bridge (or always the wrong bridge). The model is deliberately more agnostic. Early high-powered AI-fault liability can either move the system toward the high-quality equilibrium by signaling confidence and attracting users, or push it toward the low-quality equilibrium by inducing exclusion and record jamming. What improves the long-run

regime is preserving the learning content of adoption. Independent audit logs, staged validation, monitoring of exclusion, and record rules that distinguish clinical safety from formal legal shielding shift liability-induced adoption from legal insulation toward useful learning.

4.4 Sufficient statistics for liability design

The distinction between r_1 and Δ_k suggests a simple empirical blueprint. The model is intentionally written in terms of estimable sufficient statistics rather than a full structural description of every disease. The key objects are verifiability σ, patient-fault contamination η, AI control α, patient or adherence control β, patient-compensation salience ξ, algorithmic shielding productivity ϕ, legal and insurance loading K_λ, the private adoption value of accountability A_λ, the static surplus of marginal AI use, and the learning objects Ξ, τ, and ϖ. Audit studies and expert chart review can discipline σ; adherence, pharmacy, and follow-up data can discipline η, β, and ξ; interface experiments can discipline ϕ; choice experiments can discipline private accountability demand; staggered rollouts and model-update data can discipline learning; and claims or insurer pricing can discipline K.

A discrete-choice experiment identifies the private money-metric value of a liability attribute and hence the shift in adoption holding price and product characteristics fixed. If private utility is

$$U_i(\lambda, p) = V_i + \rho_i^A A(\lambda) - p + \varepsilon_i, \tag{4.47}$$

then the experiment disciplines $A'(\lambda)$ and the induced adoption derivative x_λ. It does not by itself discipline the social value of the marginal user or the value of the data generated by that user. A structural welfare exercise needs three additional objects: the marginal static surplus of AI care relative to the human outside option, the financing and risk-bearing cost of compensation, and the precision contribution of verified encounters to future model quality.

In a local calculation, the adoption component of the social marginal value is

$$r_{1k}^{adopt} = \mu_k x_{k\lambda}(0)\, \mathcal{S}_k^{marg}, \tag{4.48}$$

where $\mathcal{S}_k^{marg}$ is the social surplus of patients induced into AI care by a marginal increase in liability. The learning component, evaluated at the continuation state induced by the

point of approximation (or at the common state in a stationary calculation), is

$$\Delta_k = \delta \sum_j V_j(n^+)\Xi_{jk}\mu_k\{x_{k\lambda}(0)\tau_k\varpi_k + x_k(0)\tau_{k\lambda}(0)\varpi_k + x_k(0)\tau_k\varpi_{k\lambda}(0)\}, \tag{4.49}$$

where $n^+ = n_{t+1}$ in the dynamic model. Thus trust evidence identifies an adoption shifter, not the whole welfare numerator. Outcome data identify $\mathcal{S}_k^{marg}$; model-update or rollout data identify Ξ_{jk}, τ_k, and ϖ_k; claims and insurance data identify financing costs. Keeping these objects separate avoids double counting willingness to pay for accountability as both private demand and a social learning externality.

4.5 Scope across medical conditions and workflow stages

An AI doctor is more useful when it can handle broad differential diagnosis. Let $J \subseteq \mathcal{K}$ be the set of conditions for which the AI system offers a high-powered warranty or liability regime. Let Λ_k be the optimized incremental value of coverage for condition k holding scope fixed, and let $\Omega(J)$ be the cross-condition scope value from offering a coherent general medical service.

Primary care also has workflow complementarities: intake, differential diagnosis, red-flag triage, medication reconciliation, testing, prescribing, and follow-up. These are not themselves conditions in J. They can be modeled as a separate set of covered workflow stages $R \subseteq \mathcal{R}$ with complementarity $\Psi(J, R)$, or equivalently as part of the record and enterprise technology that determines Λ_k. To keep notation light, the proposition below focuses on cross-condition scope; the primary-care discussion later makes the workflow layer explicit.

Proposition 12 (Optimal covered scope)**.** *The optimal covered set solves*

$$J^* \in \arg\max_{J \subseteq \mathcal{K}} \left\{ \sum_{k \in J} \mu_k \Lambda_k + \Omega(J) \right\}. \tag{4.50}$$

A condition $h \notin J$ should be added to a candidate set J if and only if

$$\mu_h \Lambda_h + \Omega(J \cup \{h\}) - \Omega(J) \geq 0. \tag{4.51}$$

Condition-specific warranties are therefore efficient only when the negative standalone liability value of excluded conditions exceeds their cross-condition scope value. If scope complementarities are large, optimal liability may move toward a corner: broad AI-doctor

coverage, or no AI-doctor warranty at all. A middle case is a layered regime in which all conditions receive a triage and escalation warranty, while treatment warranties are stronger for validated conditions.

Pre-deployment evidence gives this set choice an additional interpretation. A condition may be outside J not because the AI is useless for that condition, but because validation evidence has not crossed the launch threshold described in Section 3.5. In that case the efficient design can separate a broad obligation to collect inputs and escalate red flags from a narrower obligation to diagnose, treat, or prescribe autonomously. The former preserves the scope value of an AI primary-care front door; the latter waits for condition-specific evidence that liability can price without inducing excessive distortion.

4.6 Enterprise integration and internal contracting

Let S denote standalone AI and E denote integration into a physician practice, hospital, or clinical enterprise. Integration changes the record technology and the action set. It may raise σ, lower η, lower ϕ, reduce litigation cost, and improve data quality, but it also has fixed and variable costs. More importantly, integration creates a multi-agent contracting problem. The planner assigns patient-facing liability to an enterprise, but the enterprise must then induce effort from a developer, clinicians, nurses, pharmacists, and workflow managers.

Let $i \in \{D, H\}$ index the developer and the clinical enterprise. Agent i chooses effort x_i at cost $C_i(x_i)$, and harm probability is $p(x_D, x_H, a)$. A patient-facing liability rule generates expected transfer $T(x_D, x_H, a)$. A legal split $\rho = (\rho_D, \rho_H)$ makes agent i pay share $\rho_i T$, with $\rho_i \geq 0$ and $\rho_D + \rho_H \leq 1$. The agents simultaneously choose efforts, taking the patient's effort as given.

Proposition 13 (Multi-agent liability shares)**.** *Suppose the (pure strategy) Nash equilibrium in developer and enterprise efforts is interior. A liability split ρ implements an interior first best only if*

$$\rho_i T_{x_i}(x^*) = L p_{x_i}(x^*) \qquad \text{for } i \in \{D, H\}. \tag{4.52}$$

If patient-facing liability is a single common claim and (4.52) cannot hold for both agents under the share constraint, then external liability alone cannot implement the first best.

For an internal record R_i generating $\mathcal{G}_i$, define the unrestricted incentive span at x^* by

$$\mathcal{I}_i(\mathcal{G}_i) = \left\{ \frac{\partial}{\partial x_i} \mathbb{E}[t_i(R_i) \mid x_i, x^*_{-i}] \Big|_{x_i = x_i^*} : t_i \in L^2(\mathcal{G}_i) \right\}. \tag{4.53}$$

If the enterprise can write unrestricted signed internal transfers on $\mathcal{G}_D$ and $\mathcal{G}_H$, first-order implementation by internal contracts is possible if and only if the missing incentive vector

$$(Lp_{x_i}(x^*) - \rho_i T_{x_i}(x^*))_{i \in \{D,H\}} \tag{4.54}$$

belongs to $\mathcal{I}_D(\mathcal{G}_D) \times \mathcal{I}_H(\mathcal{G}_H)$. With the same convexity or local-sufficiency conditions used in Theorems 1 and 4, this first-order implementation condition is sufficient for local Nash implementation of the target effort profile. If internal contracts are bounded, subject to limited liability, or subject to budget-balance and participation constraints, the relevant object is not the unrestricted linear span in (4.53) but the corresponding feasible moment set generated by the admissible internal contracts.

The proposition is the formal version of enterprise liability. Patient-facing enterprise liability is not magic. It works when the enterprise either controls the relevant margins directly or has internal records that can price the missing incentives. If the developer's training-data decisions, model updates, or calibration choices are not contractible inside the enterprise, direct developer liability, mandatory insurance, audit rights, or indemnity provisions remain necessary. If clinician follow-up, test completion, and escalation are not contractible by the developer, standalone developer liability leaves gaps.

A quadratic illustration makes the Nash distortion transparent. Suppose developer safety and enterprise monitoring reduce separable risks and expected claims are proportional to a common liability power λ:

$$T^\lambda = \lambda L\{\sigma_D(\bar{p}_D - x_D) + \sigma_H(\bar{p}_H - x_H)\}. \tag{4.55}$$

With costs $x_D^2/(2\alpha_D)$ and $x_H^2/(2\alpha_H)$, the Nash equilibrium is

$$x_D = \alpha_D \rho_D \lambda L \sigma_D, \qquad x_H = \alpha_H \rho_H \lambda L \sigma_H. \tag{4.56}$$

Thus shifting liability to the developer increases model-safety effort but weakens clinical-monitoring effort; shifting liability to the enterprise has the reverse effect. Common-enterprise liability is attractive when internal contracting can undo this tradeoff. Without internal contracting, optimal legal assignment is a second-best split of incentive power

across agents.

Let θ_k^S denote the standalone liability and record environment for condition k, and let $\theta_k^E(m_k)$ denote the corresponding enterprise environment when integration effort is m_k. Let $M_k(m_k)$ be the variable cost of integration effort and $K_k^I(\{t_i\})$ the cost of internal contracting, monitoring, and enforcement. For scope and fixed-cost comparisons, define v_k^S and v_k^E as optimized per-potential-patient values before condition-specific fixed costs:

$$v_k^S = \max_{\lambda_k} W_k(\lambda_k; \theta_k^S), \tag{4.57}$$

$$v_k^E = \max_{\lambda_k, m_k, \{t_i\}} \{W_k(\lambda_k; \theta_k^E(m_k), \{t_i\}) - M_k(m_k) - K_k^I(\{t_i\})\}. \tag{4.58}$$

Let F_k^S and F_k^E be total condition-level fixed costs, not multiplied by population mass. And let Ω^E and Ω^S denote the optimized cross-condition scope values under enterprise integration and standalone AI, respectively.

Proposition 14 (Enterprise assignment)**.** *Enterprise integration is efficient relative to standalone AI if and only if*

$$\sum_{k \in \mathcal{K}} \mu_k (v_k^E - v_k^S) - \sum_{k \in \mathcal{K}} (F_k^E - F_k^S) + \Omega^E - \Omega^S \geq 0. \tag{4.59}$$

Equivalently, if fixed costs are amortized per potential patient as $f_k^q = F_k^q / \mu_k$, the condition is $\sum_k \mu_k [(v_k^E - f_k^E) - (v_k^S - f_k^S)] + \Omega^E - \Omega^S \geq 0$.

The proposition clarifies when AI should be incorporated into a doctor's practice. The case is strongest when standalone AI has low verifiability, patient nonadherence is likely to be misattributed to the AI, harms are severe, and clinical integration sharply improves monitoring, prescribing, follow-up, escalation, and attribution. The units are explicit: variable values are per potential patient, while fixed costs are total costs unless amortized.

4.7 Market structure and enterprise assignment

Enterprise integration also changes market structure. This matters because clinical enterprises do not only choose records and effort; they also choose prices, coverage, and access. The point is familiar from health-care markets in which ownership and consolidation can change quality, utilization, and prices, including dialysis (Eliason et al., 2020). I incorporate that concern by separating the productive value of integration from the market-power consequences of assigning liability to large clinical platforms.

For a case type k, let q be a continuous index of integration quality, with $q = S$ and $q = E$ denoting the standalone and enterprise choices in the discrete assignment problem above. Let inverse demand be

$$P_k(x, q) = v_k(q) - B_k x, \tag{4.60}$$

where $v_k(q)$ includes patient willingness to use the service, including any perceived accountability value. Let marginal cost be $c_k(q)$, and let $h_k(q)$ denote per-user safety, record, or attribution value not fully priced by patients. Let $\ell_k(q)$ denote per-user learning value. A conduct parameter $\kappa_k \in [0, 1]$ summarizes market structure through the equilibrium condition

$$P_k(x, q) - c_k(q) = \kappa_k B_k x. \tag{4.61}$$

The case $\kappa_k = 0$ is price-taking behavior; larger κ_k represents more market power. The induced quantity is

$$x_k(q, \kappa_k) = \frac{v_k(q) - c_k(q)}{B_k(1 + \kappa_k)}. \tag{4.62}$$

Social value, holding fixed the liability rule used under quality q, is

$$W_k(q, \kappa_k) = \int_0^{x_k(q,\kappa_k)} \{v_k(q) - B_k s - c_k(q) + h_k(q) + \ell_k(q)\}\, ds - F_k(q). \tag{4.63}$$

Proposition 15 (Market structure and enterprise integration)**.** *For an interior differentiable change in integration quality q, holding the conduct parameter κ_k fixed, the social marginal value is*

$$\frac{\partial W_k}{\partial q} = x_k\{v_{kq} - c_{kq} + h_{kq} + \ell_{kq}\} + x_{kq}\{\kappa_k B_k x_k + h_k + \ell_k\} - F_{kq}. \tag{4.64}$$

Hence enterprise integration is more attractive when it raises patient value, lowers marginal cost, improves record or safety value, or increases learning value. If integration also changes conduct, so that $\kappa_k = \kappa_k(q)$, the total derivative adds

$$x_{k\kappa}\{\kappa_k B_k x_k + h_k + \ell_k\}\kappa_{kq}, \qquad x_{k\kappa} = -\frac{x_k}{1 + \kappa_k}. \tag{4.65}$$

Market power therefore changes the assignment criterion through both the level of adoption x_k and the quantity-distortion term. If integration raises quality but also raises markups or restricts access, the sign of the market-structure adjustment is ambiguous.

This proposition modifies Proposition 14. Under perfect competition and no unpriced record or learning value, the integration criterion reduces to the productive comparison between v^E and v^S net of fixed costs. When conduct is fixed, market power changes the activity-level effect through the induced quantity; when integration also changes conduct, equation (4.65) adds the markup channel explicitly. Under market power, the same liability assignment can have an additional activity-level effect. If integration raises the quality of the record and expands adoption, market power may blunt but not eliminate the benefit. If integration mainly consolidates control, raises markups, or excludes complex patients, the enterprise assignment is less attractive even when it improves attribution among the parties that remain inside the platform.

4.8 Direct prescribing

The analysis above treats prescribing as an action that may be mediated by a clinician or clinical enterprise. Consider the variant in which the AI may directly prescribe medication. Direct prescribing changes the optimal liability rule on the margin because it moves part of treatment choice from the clinical enterprise to the AI. It also creates new risks: drug interactions, contraindications, dosing errors, allergic reactions, failure to monitor laboratory values, patient nonadherence, and over-prescribing or under-prescribing designed to reduce legal exposure rather than improve health.

Let direct prescribing add an AI-controlled medication-safety action $s \geq 0$ at cost $s^2/(2\gamma)$, a patient medication-adherence action $u \geq 0$ at cost $u^2/(2\zeta)$, and a medication-specific defensive-design action $b^M \geq 0$ at cost $(b^M)^2/(2\chi^M)$. Medication injury has loss L^M. A scalar liability share loads on medication error with verifiability σ^M, on patient medication nonadherence with contamination η^M, and on defensive prescribing with legal-shielding productivity ϕ^M. Let $\xi^M \in [0,1]$ be the salience of medication-related compensation for patient adherence. The medication block contributes

$$\lambda L^M\{\sigma^M(\bar{p}^M - s) + \eta^M(\bar{p}^H - u) - \phi^M b^M\} \tag{4.66}$$

to expected liability, where $\bar{p}^M$ is baseline AI medication-error risk and $\bar{p}^H$ is baseline risk from patient medication behavior. Holding the non-prescribing liability technology fixed, define

$$A^M = (L^M)^2\gamma\sigma^M, \qquad D^M = (L^M)^2\{\gamma(\sigma^M)^2 + (\xi^M)^2\zeta(\eta^M)^2 + \chi^M(\phi^M)^2\}. \tag{4.67}$$

Let A and D denote the numerator and denominator of the liability index before direct prescribing.

Proposition 16 (Marginal effect of direct prescribing)**.** *Suppose the optimal liability share is interior before and after direct prescribing, and $D^M > 0$. With direct prescribing, the local optimal liability share is*

$$\lambda^{RX} = \frac{A + A^M}{D + D^M}. \tag{4.68}$$

Direct prescribing raises the optimal liability share if and only if

$$\frac{A^M}{D^M} > \frac{A}{D}. \tag{4.69}$$

It lowers the optimal liability share if the reverse inequality holds.

The prescription result is a modular aggregation result. The pooled liability share can be written as

$$\lambda^{RX} = \frac{D}{D + D^M}\frac{A}{D} + \frac{D^M}{D + D^M}\frac{A^M}{D^M}.$$

Thus the new pooled liability share is a curvature-weighted average of the old AI-care index and the prescription-specific index. Direct prescribing raises the pooled optimal liability share if and only if the prescription block has a higher incentive-to-distortion ratio than the existing AI-care block:

$$\frac{A^M}{D^M} > \frac{A}{D}.$$

This is why more autonomous authority does not mechanically imply more liability. If prescriptions are logged, contraindication checks are auditable, pharmacy fills are observable, and medication errors are highly AI-controlled, then A^M/D^M is high and direct prescribing raises optimal liability. If the new prescribing margin is dominated by patient nonadherence, missing information, drug-interaction uncertainty, or defensive over-prescribing, then A^M/D^M is low and direct prescribing can lower the optimal pooled liability share.

The result has a simple interpretation. Direct prescribing raises optimal liability when medication decisions are both important and verifiably AI-controlled: prescriptions are logged, contraindication checks are auditable, pharmacy fills are observable, and the AI can reduce drug-specific harm at low cost. Direct prescribing lowers optimal liability when the new margin is dominated by patient-side information and adherence problems, or when liability mainly induces defensive over-prescribing, excessive contraindication warnings, refusal to prescribe, or automatic referral. Thus autonomous prescribing does

not mechanically call for stricter or weaker liability. It adds a block to the liability index, and the sign depends on whether that block has a higher or lower incentive-to-distortion ratio than the existing AI-care block. If medication liability can be unbundled from diagnostic and triage liability, the prescription-specific power is $\left[A^M/D^M\right]_0^1$ rather than the pooled ratio in (4.68). Access benefits from direct prescribing affect whether direct prescribing belongs in the AI service scope; conditional on inclusion, they affect the liability share only to the extent that they vary with liability.

A second margin is the balance between omission and commission liability. Let r denote prescription intensity. Let $p^U(r)$ be untreated-disease risk and $p^M(r)$ be medication-injury risk, with marginal effects $(p^U)'(r) < 0$ and $(p^M)'(r) > 0$ over the relevant range. An omission-liability share that is too high relative to commission liability pushes r upward and can generate over-prescribing. A commission-liability share that is too high relative to omission liability pushes r downward and can generate under-prescribing or excessive referral. Thus direct prescribing separates the liability problem into two prices: the price of failing to treat and the price of treating unsafely. The optimal design equates each price to the verifiable marginal harm on that side of the prescription margin.

4.9 Robust liability under ambiguity as a secondary extension

The baseline analysis treats the relevant primitives as known. A secondary extension allows the planner not to know the true record distribution, cost functions, or defensive-design technology. Let $\vartheta \in \Theta$ index these primitives and let $W(\lambda;\vartheta)$ be reduced welfare under liability power λ. A robust planner evaluates

$$\max_{\lambda\in[0,1]} \inf_{\vartheta\in\Theta} W(\lambda;\vartheta). \tag{4.70}$$

Proposition 17 (Robust liability)**.** *Suppose Θ is compact, $W(\lambda;\vartheta)$ is continuous in (λ,ϑ), and $W(\cdot;\vartheta)$ is differentiable and concave for every $\vartheta \in \Theta$. Let $W^R(\lambda) = \inf_{\vartheta\in\Theta} W(\lambda;\vartheta)$, so a robust maximizer exists on $[0,1]$. Let $\bar{\lambda}$ be the largest maximizer under a benchmark model $\bar{\vartheta}$. If*

$$W_\lambda(\lambda;\vartheta) \le W_\lambda(\lambda;\bar{\vartheta}) \qquad \text{for all } \lambda \in [0,1] \text{ and all } \vartheta \in \Theta, \tag{4.71}$$

then there exists a robust maximizer weakly below $\bar{\lambda}$. Conversely, if the inequality is reversed for all ϑ, then there exists a robust maximizer weakly above the smallest benchmark maximizer. Under rectangular ambiguity over legal scores, and for the pointwise max-min

analogue of Theorem 3, the robust legal score is

$$\bar{S}^R_{\mathcal{G}}(D) = \inf_{\vartheta \in \Theta} \mathbb{E}_\vartheta[S_{0,\vartheta}(D) \mid \mathcal{G}, Y = 1]. \tag{4.72}$$

Thus robust liability loads only on record components whose legal score remains positive in the worst case. If one wants every robust maximizer, rather than some robust maximizer, to lie on the same side of the benchmark, an additional no-flat-region or uniform strict-order condition is required.

This result pushes the model toward robustly informative records. If ambiguity is mainly about whether a signal indicates AI fault or patient nonadherence, high-powered liability is fragile. If ambiguity is mainly about the level of loss but the record robustly separates marginal responsibility, high-powered liability is less fragile. Robustness therefore favors simple safe harbors or comparative-fault partitions only when those partitions are stable across plausible models of medical causation.

5 Legal, Regulatory, and Empirical Background

I use the model as a direct extension of models of classic accident law in the law and economics literature. Calabresi (1970) frames tort as a problem of minimizing accident costs, prevention costs, and administrative costs. Shavell (1980, 2009) show that strict liability and negligence differ in how they affect care levels and activity levels. Landes and Posner (2013) and Polinsky and Shavell (2007) emphasize the role of legal error and administrative cost. The present model preserves that logic but adds two medical-AI features: the patient is also a hidden-action agent, and the record is a technologically chosen object.

The institutional choice between tort, contracts, and regulation is an established literature. Glaeser, Johnson, and Shleifer (2001) emphasize that private ordering and court-based enforcement can fail when transaction costs, information problems, or enforcement constraints are severe. Glaeser and Shleifer (2003) study the rise of regulation as an alternative to litigation in environments where courts are costly or unreliable. Glaeser and Shleifer (2002) is useful background for the broader point that legal institutions differ in how they adapt to new information. Gennaioli, Shleifer, and Vishny (2015) is also useful because AI medical advice is an advice market in which trust, delegation, and accountability affect demand. In my framework, ex ante regulation is valuable not because it is conceptually separate from liability, but because it can refine $\mathcal{G}$, lower proof costs, create

safe-harbor partitions, or change which party can be made residual claimant.

The contract theory connection is the informativeness principle of Holmstrom (1979). A verifiable signal is valuable when it incrementally changes the usable legal score in a responsive region and is worth its proof, loading, and endogenous-record-design costs. A signal that is noisy in levels may still be useful if it adds conditional information about the relevant score, while a precise-looking signal should not be used if it mostly reflects manipulable formal compliance rather than marginal responsibility. The closest modern liability-design paper for the information-acquisition margin is Poggi and Strulovici (2020). They study firms that privately know product risk and can acquire evidence before launching; liability can be represented by tariffs on the evidence acquired if harm occurs, and such tariffs shape due diligence and launch thresholds. I view their paper as the pre-deployment counterpart to the present model. AI medicine also has validation and launch evidence, but the central additional difficulty is that the medically relevant evidence is generated during treatment and is jointly produced by the AI, the clinical enterprise, and the patient. Warranty-signaling models, including Spence (1977) and Grossman (1981), provide a separate reason why guarantees may affect demand when firms have private information about quality.[13] Section 4.3 uses a compact two-type signaling game to microfound the dynamic warranty-adoption channel, while the main static liability index treats the demand-side value of accountability through $R(\lambda)$. The core contribution remains how record-based liability changes incentives, patient moral hazard, learning, and defensive design.

The medical-malpractice literature motivates the defensive-design term. Kessler and McClellan (1996) provide evidence that malpractice pressure can change treatment intensity, while Sloan and Shadle (2009) and Sloan and Chepke (2008) emphasize mixed deterrence and high administrative costs. Currie and MacLeod (2008) show that tort rules can affect medical outcomes, highlighting that liability changes real clinical behavior rather than merely redistributing losses. In AI medicine, defensive behavior may be designed into the software itself.

The AI medical-liability literature emphasizes doctrinal fragmentation. Price, Gerke, and Cohen (2019) discuss physician liability for using AI. Price, Gerke and Cohen (2024) analyze how liability should be distributed among physicians, developers, and institutions. Sullivan and Schweikart (2019) argue that existing tort doctrines may be poorly

[13]The paper uses "warranty" in the economic sense of a liability-backed guarantee, not as a claim that warranty doctrine is always the right legal hook. In many cases, design defect, warning defect, negligent implementation, malpractice, or enterprise liability will be the relevant doctrinal form. The model asks what record and transfer each form can implement.

suited for AI-caused injury. Price and Cohen (2024) emphasize that medical-AI liability turns on locating responsibility among developers, clinicians, and health-care institutions. Mello et al. (2024) provide the most direct legal motivation for the record-based model. They argue that liability risk from health-care AI depends on sparse case law, whether software is treated as a device or as clinical decision support, whether clinicians can independently evaluate the recommendation, the opacity of the model, the plaintiff's ability to prove defect and causation, and the contracts and indemnities that allocate risk among developers, clinicians, and institutions. In my notation, this article is about the joint determination of the feasible sigma-field, the posterior probability of AI fault, and the identity of the residual claimant. It is also a reminder that "AI liability" is not one rule: the same injury may be analyzed as malpractice, product liability, negligent credentialing, vicarious liability, breach of contract, or insurance exposure depending on who controlled the relevant margin and what the record makes verifiable. Holm et al. (2021) argue that AI-related medical injuries strengthen the case for no-fault compensation. Lior (2022) surveys insurance-based approaches to AI risk. The commonality is that the record is poor and the responsible actor is ambiguous. In the present model, those concerns are summarized by low verifiability, high contamination, high administrative cost, high value of enterprise integration, and a distinction between external patient-facing liability and internal indemnity.

Current regulation affects the model's parameters rather than directly solving tort allocation. In the United States, the 21st Century Cures Act excluded certain software functions, including some clinical decision support software, from the statutory definition of device. FDA's January 2026 Clinical Decision Support Software guidance clarifies the non-device CDS criteria and distinguishes those functions from device software functions (Food and Drug Administration, 2026a). FDA's guidance on predetermined change control plans for AI-enabled device software concerns how manufacturers specify planned modifications, methods for developing and validating them, and impact assessments over the device life cycle (Food and Drug Administration, 2025). FDA also maintains a public list of AI-enabled medical devices authorized for marketing in the United States (Food and Drug Administration, 2026b). In the European Union, the AI Act creates a risk-based framework for AI systems, including obligations for high-risk systems involving risk management, data governance, documentation, logging, information to deployers, human oversight, accuracy, robustness, and cybersecurity (European Parliament and Council, 2024a; European Commission, 2026). The revised Product Liability Directive includes software within the concept of product and modernizes defect liability for digital prod-

ucts (European Parliament and Council, 2024b). In the framework above, these rules alter the record, verifiability, feasible safe harbors, and proof costs.

The adoption evidence is still emerging, but it is central to the model. The key empirical fact motivating the accountability channel is that patients and clinicians appear to value liability clarity beyond actuarial compensation. A liability attribute coefficient, divided by the marginal utility of money, gives a private money-metric value of the liability attribute; a difference in required compensation for AI primary care with and without liability gives the same private object. This is not the social term $R(1) - R(0)$ unless the econometrician has already mapped private demand into social surplus, financing costs, and risk-bearing value. The evidence identifies the private demand shifter in Section 4.4, not the full numerator of the liability index, which also contains deterrence, legal-cost, and learning terms. To evaluate welfare, the econometrician must combine that shifter with the surplus of marginal AI users, the cost of financing compensation, and the precision value of the data they generate. If liability raises trust for high-surplus care, it can increase welfare. However, if it raises trust for unsafe or poorly monitored care, it can lower welfare.

6 Policy Instruments as Restrictions of the Framework

Without ranking policy proposals in the abstract, I map them into four institutional restrictions: restrictions on the transfer, restrictions on the record, restrictions on the liable entity, and restrictions on scope or timing. This keeps the policy discussion tied to the model's primitives rather than to labels such as "product liability" or "malpractice."

Restrictions on the transfer. No AI liability is the restriction $w = 0$. It is optimal only when verifiable AI control, social accountability value, and learning value are too small to justify legal and insurance costs. Strict liability is a high-powered transfer for covered injuries. It is optimal in the framework when it approximates posterior AI-caused harm, not when it indiscriminately compensates all adverse outcomes. Partial warranties, caps, coinsurance, and deductibles are intermediate restrictions on w. They are useful when liability has deterrence or trust value but full outcome liability would over-insure patient-side conduct, invite defensive design, or create excessive loading. No-fault compensation is a different transfer restriction: it compensates eligible injuries without individualized proof of fault. It is attractive when compensation has social value and proof is costly, but deterrence must then be supplied by premiums, audits, experience rating, or regulatory

penalties.

Restrictions on the record. Medical malpractice negligence, safe harbors, burden shifting, comparative fault, and logging mandates all change the information used by the transfer. Negligence is a cutoff on a breach score. A safe harbor sets liability to zero on a certified compliance set and uses a fallback rule outside it. Comparative fault adds patient-input, test-completion, follow-up, and adherence records. Burden shifting and rebuttable presumptions change the effective posterior when plaintiffs lack access to technical evidence. Audit, logging, documentation, and human-oversight regulation refine $\mathcal{G}$ when they produce clinically informative records; they do not help merely by producing more text.

Restrictions on the liable entity. Developer liability, physician malpractice, hospital vicarious liability, common-enterprise liability, mandatory insurance, and AI-corporation proposals differ in who bears T and whether that party controls or can contract over the relevant margin. Enterprise liability is useful when the enterprise improves monitoring, adherence, escalation, and attribution and can transmit incentives internally through indemnity, insurance, procurement, and audit rights. Mandatory insurance changes solvency and pricing, not the optimal transfer itself. AI personhood or a special liability vehicle is useful only if assets, premiums, and governance transmit incentives back to the developer and clinical enterprise.

Restrictions on scope and timing. Regulatory sandboxes, temporary caps, prescribing restrictions, clinician-signoff rules, and limits on autonomous AI doctoring restrict the set of cases, actions, or periods in which high-powered liability applies. They are favored by the framework when high-powered liability would degrade learning by reducing adoption, data quality, or representativeness; when prescribing or high-severity decisions are poorly logged; when pre-deployment validation evidence has not crossed an evidence-based launch threshold; or when no liability rule can simultaneously induce AI safety, preserve patient effort, avoid algorithmic defensive design, and compensate injuries at acceptable cost. Large early learning externalities by themselves do not point toward weaker liability: if liability increases trustworthy use and clinically informative records, the same learning channel can favor stronger early accountability.

A policy instrument is not good because of its doctrinal name. It is good when it changes the record, transfer, liable entity, or scope in the direction required by the sufficient statistics above.

7 Implications for Optimal (Market) Design

I use the framework to derive conditional design statements rather than categorical policy prescriptions. The decision can be read as a map from sufficient statistics to legal form. The relevant statistics are: verifiability of AI-controllable harm (σ); verifiability and salience of patient adherence (η, ξ); legal and insurance cost (K); the social value of compensation; productivity of algorithmic shielding (ϕ); learning and adoption value (Δ); cross-condition scope (Ω); enterprise productivity in improving records and follow-up; and market-power or access distortions.

High AI verifiability and low patient contamination. If AI-controllable fault is highly verifiable, patient nonadherence is observable or unimportant, defensive legal shielding is difficult, and administrative costs are moderate, then high-powered AI-fault liability is optimal. The model-implied rule is not a guarantee of cure. In the separable-cause benchmark it is compensation for the posterior AI-caused component of harm; in joint-causation settings it is the closest feasible approximation to the marginal-responsibility score.

High patient effort and poor adherence observability. If patient effort is central and hard to verify, then full outcome warranty is generally inefficient. The optimal rule is conditional: it pays more when required inputs were supplied and advice was followed, and pays less when nonadherence or use outside the covered context is visible. This is the case for partial warranty, comparative fault, deductibles, and adherence conditions.

Low causation verifiability and high compensation value. If proof of causation is very noisy and litigation costs are high, and compensation has social value because of risk aversion, liquidity constraints, distributional weights, or lower administrative cost, then compensation and deterrence can be separated. A no-fault pool can pay eligible injuries while premiums, audits, and experience rating provide incentives. This is optimal only if financing responds to risk; otherwise no-fault becomes pure insurance against preventable design errors.

High algorithmic shielding productivity. If liability mainly induces over-referral, over-testing, over-prescribing, or refusal of ambiguous patients, then the legal-shielding term is large and the optimal liability share falls. Depending on which instrument changes the primitive, the optimal design shifts toward lower liability power, more informative

records, safe harbors based on clinically meaningful protocols, or enterprise integration that makes defensive behavior observable.

Large learning and adoption externalities. If encounters create large learning spillovers and liability increases both adoption and data quality, then the dynamic term raises optimal liability. If liability reduces adoption by increasing price or exclusion, or reduces data quality or representativeness through defensive selection, then the dynamic term lowers optimal liability. Dynamic learning does not mechanically favor immunity or strict liability; its sign is empirical.

Learning traps and record regulation. If liability-backed adoption produces clinically informative encounters, AI quality, adoption, and liability can reinforce one another. If liability instead induces exclusion or record jamming, the system can remain in a low-quality, low-adoption regime. Record regulation and audit logs are valuable when they preserve the learning content of adoption by separating true clinical performance from formal legal shielding.

Large scope complementarities. If an AI doctor derives value from broad scope, condition-by-condition warranties can destroy usefulness. The optimal covered set depends on the standalone liability value of each condition and the scope complementarity from a coherent medical service. The framework therefore allows narrow warranties for validated tasks, broad triage warranties, or enterprise-wide liability, depending on the magnitude of scope externalities.

High enterprise-record productivity. If clinical integration sharply improves verification, adherence, escalation, and attribution, then enterprise liability is optimal. The enterprise can face the patient and allocate responsibility internally through indemnity, insurance, monitoring, and procurement. If the AI task is narrow, logged, autonomous, and low in patient-effort moral hazard, standalone developer liability may be sufficient. If integration mainly creates market power or excludes complex patients, the enterprise assignment is less attractive.

There are clear ties between empirics and policy in these statements. Evidence of high AI-error verifiability and low patient-fault contamination favors high-powered AI-fault liability. Evidence that causation is noisy but compensation is valuable favors no-fault compensation combined with experience-rated premiums or audits. Evidence that clinical workflow greatly improves verification and follow-up favors enterprise liability. Evidence

that liability mainly induces over-referral, exclusion of hard cases, or record jamming favors caps, safe harbors, or regulation of the record technology. Large positive learning spillovers favor higher liability when liability increases trustworthy use and data quality, but lower liability when it selects away the cases from which the model needs to learn.

7.1 Coverage in AI primary care

Primary care illustrates why the optimal object is a layered liability contract rather than a single product warranty. A primary care patient usually presents with a symptom, not a known condition. In the model, cross-condition scope means that $\Omega(J)$ is large: a narrow warranty for a short list of diagnoses may be actuarially clean but clinically confusing, because the patient may not know whether the true condition is in the covered set. Primary care also has workflow complementarities–intake, differential diagnosis, red-flag triage, medication reconciliation, testing, prescribing, and follow-up–which are better represented by the workflow set R and complementarity $\Psi(J, R)$ described in Section 4.5, or by the record and enterprise technology that determines the incremental coverage value Λ_k. When workflow records mainly create future training data, the same facts also enter the dynamic learning term Δ_k.

The first implied coverage layer is intake, red-flag, and escalation coverage. This layer is broad because the AI controls the questions asked, the urgency classification, the warning language, and the instruction to seek in-person or emergency care. The model-implied transfer is high-powered when the patient supplied the requested information and the record shows that the AI failed to ask a necessary question, ignored a red flag, falsely reassured the patient, or failed to escalate. The transfer is low-powered when the record shows missing inputs, use outside the service boundary, or refusal to follow clear urgent-care instructions.

The second layer is diagnostic and test-ordering coverage for validated presentations. This layer is stronger when the relevant vitals, labs, images, medication lists, allergies, or history elements are verified and logged. It is partial when the diagnosis depends mainly on unverified patient report or when the patient must complete tests before the AI can safely classify the case. In the notation above, verified tests and logs raise σ_k and lower η_k; missing tests and unverified histories do the opposite.

The third layer is treatment and medication-management coverage. For chronic-care settings such as hypertension follow-up, diabetes titration, asthma plans, depression monitoring, anticoagulation reminders, and medication refills, repeated measurements and pharmacy records raise verifiability. But patient effort is also central. The model-implied

contract therefore places weight on logged measurements, lab completion, prescription fills, symptom check-ins, and escalation compliance. If the AI directly prescribes, Proposition 16 governs how the medication block changes the pooled optimal liability share. The separate omission-versus-commission margin then balances liability for failure to intensify therapy against liability for unsafe medication changes, interactions, contraindications, and monitoring failures.

The fourth layer concerns high-severity or ambiguous primary care presentations. Chest pain, neurologic deficits, severe abdominal pain, pregnancy-related symptoms, pediatrics, frailty, polypharmacy, and immunosuppression are examples in which loss severity is high and the cost of missed escalation can dominate. If standalone records are weak and patient effort is hard to verify, the framework implies weaker autonomous-treatment coverage, enterprise integration, or pooled compensation. If a clinical enterprise verifies inputs and controls escalation, higher-powered enterprise liability becomes efficient.

Finally, some primary care interactions are closer to information services: education, wellness advice, administrative navigation, and low-stakes self-care guidance. When the AI does not control a major medical risk, the patient can cheaply verify the advice, or expected loss is small, the framework implies low-powered medical liability or no medical-injury liability. Thus, primary care coverage is broad at the triage boundary, strong for verified and validated treatment contexts, conditional for adherence-sensitive management, and weak for low-risk information. This is not a direct policy proposal; it is the coverage pattern generated by the model's comparative statics.

8 Conclusion

I find that optimal medical liability for AI is a problem of contract design under tort constraints. The optimal liability rule is not chosen only to compensate injuries after they occur. It shapes model design, patient effort, clinical workflow, adoption, insurance markets, and learning. The mathematical object is a transfer written on the verifiable medical record.

Liability should be high-powered for AI-controllable harm when the record can distinguish that harm from patient nonadherence and natural disease progression. When the record is contaminated, partial liability is not a compromise between politics and economics; it is the second-best implication of bilateral moral hazard. The general model adds that liability can induce algorithmic defensive design, can create or destroy learning externalities, and must distinguish recommendation from direct prescribing. The dy-

namic result is that AI medical liability can generate complementarity: liability raises adoption and adoption creates learning, but only when the resulting encounters are clinically informative. Because liability can also induce exclusion and record jamming, the same technology can support low-quality/low-adoption and high-quality/high-adoption regimes. Existing policy proposals are therefore best understood as different ways of tweaking the same primitives: verifiability, contamination, legal cost, insurance capacity, learning value, prescribing authority, scope, and enterprise control. The same logic applies to other high-stakes advice markets (e.g., finance, education, law, employment, and public benefits) whenever the user supplies inputs, the advice provider controls the interface, and accountability depends on a record that can be designed.

A Proofs

Proof of Theorem 1. The social first-order conditions for an interior first best are (2.3) and (2.4). The AI's private first-order condition is $C'(e) + T_e(e, a) = 0$. For this to coincide with (2.3) at (e^*, a^*), it is necessary that $T_e(e^*, a^*) = Lp_e^A(e^*)$. The patient's private first-order condition is $G'(a) + Lp_a^P(a) - T_a(e, a) = 0$. For this to coincide with (2.4), it is necessary that $T_a(e^*, a^*) = 0$. If the private objective functions are convex, these first-order conditions are sufficient for private optimality. □

Proof of Lemma 1. For any bounded transfer, differentiation under the integral gives

$$\frac{\partial^2}{\partial e^2}\mathbb{E}[w(D) \mid e, a] = \int w(d)\frac{\partial^2 f(d \mid e, a)}{\partial e^2}\,\mathrm{d}d.$$

Because $0 \leq w \leq L$, the absolute value of this expression is at most LM_e. Hence the developer's objective has second derivative at least $C''(e) - LM_e \geq \varepsilon$. The patient objective has second derivative

$$G''(a) + Lp_{aa}(e, a) - \frac{\partial^2}{\partial a^2}\mathbb{E}[w(D) \mid e, a],$$

which is at least $G''(a) + Lp_{aa}(e, a) - LM_a \geq \varepsilon$. Thus both unilateral objectives are strictly convex, and their first-order conditions are sufficient for global optimality. □

Proof of Corollary 1. By the stable-attribution condition (2.11), the expected transfer generated by w^* is

$$T(e, a) = \mathbb{E}[Lq_A(Z, Y)Y \mid e, a] = L\mathbb{P}(C = A, Y = 1 \mid e, a)$$

for all (e, a) in a neighborhood of the first best. Under separability, $\partial \mathbb{P}(C = A, Y = 1)/\partial e = p_e^A(e)$ and $\partial \mathbb{P}(C = A, Y = 1)/\partial a = 0$. Therefore $T_e = Lp_e^A$ and $T_a = 0$, which are the implementability conditions in Theorem 1. Convexity of the private problems gives first-best implementation. □

Proof of Proposition 1. The score identity gives $T_j = \partial \mathbb{E}[w]/\partial x_j = \mathbb{E}[ws_j]$ for $j \in \{e, a\}$. The first-best and private first-order conditions imply $T_e = Lp_e$ and $T_a = 0$, giving (2.14). For the minimum-second-moment statement, solve

$$\min_w \frac{1}{2}\mathbb{E}[w^2] \quad \text{s.t.} \quad \mathbb{E}[ws] = \tau.$$

The Lagrangian is $\frac{1}{2}\mathbb{E}[w^2] - \lambda'(\mathbb{E}[ws] - \tau)$. The first-order condition is $w = s'\lambda$. Substituting into the constraint gives $G_s\lambda = \tau$, hence $\lambda = G_s^{-1}\tau$ and (2.15). Since scores have mean zero, adding a constant preserves the score-moment constraints; it changes the second moment and leaves the variance of $w^J + c$ equal to the variance of w^J. Thus (2.15) is the unique minimum-L^2-norm solution, or the minimum-variance solution after imposing the normalization $\mathbb{E}[w] = 0$. Posterior AI-fault liability is a feasible rule in this class only if its induced score moments equal τ. □

Proof of Theorem 2. The necessity of (2.18) follows directly from Proposition 1: any bounded record-based rule that implements the first best must generate the target score moments. The set $\mathcal{M}(\mathcal{G}, L)$ is convex because $\mathcal{W}_{\mathcal{G},L}$ is convex and the moment map is linear. It is bounded because $0 \leq w \leq L$ and the scores are square-integrable. The admissible set $\mathcal{W}_{\mathcal{G},L}$ is closed, convex, and bounded in L^2, hence weakly compact because L^2 is reflexive. The moment map is weakly continuous, so $\mathcal{M}(\mathcal{G}, L)$ is compact in finite-dimensional Euclidean space. If welfare is exactly a positive definite quadratic loss in the moment vector, projection onto this nonempty compact convex set exists and is unique. If welfare also contains direct terms in w, the relevant optimization is over the transfer rule itself, so the projection statement is no longer a complete characterization. For the collinearity claim, if w is $\mathcal{G}$-measurable, then

$$\mathbb{E}[ws_a] = \mathbb{E}[w\mathbb{E}(s_a \mid \mathcal{G})] = \rho\mathbb{E}[w\mathbb{E}(s_e \mid \mathcal{G})] = \rho\mathbb{E}[ws_e].$$

Thus every feasible moment is on the line $m_a = \rho m_e$. A target with nonzero AI moment and zero patient moment is not on that line when $\rho \neq 0$. □

Proof of Proposition 2. With risk neutrality, fixed adoption, actuarially fair financing, and no administrative cost, a transfer from the liable party to the patient is a resource-

neutral redistribution. It cancels from aggregate surplus except through its effect on actions. With concave utility, a transfer that is high in injury states and financed by a smoother premium reduces consumption risk and has positive marginal value up to the point at which insurance loading offsets risk sharing. With endogenous adoption, differentiating welfare $W(\lambda) = \mathbb{E}[x_i(\lambda)S_i(\lambda)]$ gives $W'(\lambda) = \mathbb{E}[x_i'(\lambda)S_i(\lambda) + x_i(\lambda)S_{i\lambda}(\lambda)]$, plus any risk-bearing term. This is (2.20). □

Proof of Proposition 3. Since $\mathcal{H}' \supseteq \mathcal{H}$, every $\mathcal{H}$-measurable transfer is also $\mathcal{H}'$-measurable. Thus the feasible incentive set $g(\mathcal{W}(\mathcal{H}))$ is contained in $g(\mathcal{W}(\mathcal{H}'))$, so the minimum projection loss under $\mathcal{H}'$ is weakly lower. The gain is strict exactly when the projection of τ onto the larger set differs from the projection onto the smaller set. □

Proof of Theorem 3. Fix a multiplier vector ψ. For a $\mathcal{G}$-measurable perturbation h of the injury-state payment $\widetilde{w}$, the derivative of the Lagrangian corresponding to (3.3) is

$$\mathbb{E}\left[Y\{S_0(D) - \Gamma'(\widetilde{w}(D))\}h(D)\right].$$

By iterated expectations this equals

$$\mathbb{E}\left[\mathbb{P}(Y = 1 \mid \mathcal{G})\{\bar{S}_{\mathcal{G}}(D) - \Gamma'(\widetilde{w}(D))\}h(D)\right],$$

where $\bar{S}_{\mathcal{G}}(D) = \mathbb{E}[S_0(D) \mid \mathcal{G}, Y = 1]$. On cells with positive injury probability, the nonnegative factor $\mathbb{P}(Y = 1 \mid \mathcal{G})$ does not affect the pointwise maximizer; on zero-probability injury cells the injury-state transfer is payoff-irrelevant. Thus the problem reduces pointwise to maximizing $sx - \Gamma(x)$ over $x \in [0, L]$, with $s = \bar{S}_{\mathcal{G}}(D)$. Strict convexity of Γ makes this objective strictly concave, so the maximizer is unique and is the extended inverse $g_\Gamma(s)$. If s lies below the right derivative at 0, the maximizer is 0; if it lies in the interior range of Γ', the interior first-order condition gives $x = (\Gamma')^{-1}(s)$; if it lies above the left derivative at L, the maximizer is L. The global solution satisfies the same pointwise formula at the consistent multiplier $\psi^* = Q\{\tau - \mathbb{E}[Y\widetilde{w}^* s]\}$. If $\Gamma(w) = \gamma w$, the pointwise objective is linear in $\widetilde{w}$ with slope $\bar{S}_{\mathcal{G}} - \gamma$, so the optimum is L when the coefficient is positive, 0 when it is negative, and any value in $[0, L]$ when it is zero. If $\bar{S}_{\mathcal{G}}(z)$ is nondecreasing in a scalar record, the set on which it exceeds γ is an upper set and hence a threshold set. □

Proof of Proposition 4. Fix S_0 and the multiplier vector. If $\mathcal{G}_2 \supseteq \mathcal{G}_1$, every $\mathcal{G}_1$-measurable transfer is feasible under $\mathcal{G}_2$, so value cannot fall. A sharper strict-gain statement follows

from the pointwise value function

$$v_\Gamma(s) = \max_{x\in[0,L]} \{sx - \Gamma(x)\}.$$

Under the coarse field, the relevant score on a cell is $\mathbb{E}[S_0 \mid \mathcal{G}_1, Y = 1]$ and the pointwise value is $v_\Gamma(\mathbb{E}[S_0 \mid \mathcal{G}_1, Y = 1])$. Under the finer field, the expected value on the same coarse cell is $\mathbb{E}[v_\Gamma(\mathbb{E}[S_0 \mid \mathcal{G}_2, Y = 1]) \mid \mathcal{G}_1, Y = 1]$. Since the coarse decision can be replicated by the fine partition, this conditional value is weakly larger. It is strictly larger exactly when the displayed inequality in the proposition holds on a set with positive injury probability. This formulation includes affine-cost kink gains, because $v_\Gamma(s) = L\max\{s-\gamma, 0\}$ is convex and can be strictly improved by splitting an indifference cell into above- and below-threshold subcells. If the multiplier vector is reoptimized across regimes, weak improvement follows from set inclusion, while strict improvement is a statement about optimized values. □

Proof of Proposition 5. If experiment q' Blackwell-dominates experiment q, every decision rule feasible after observing R_q can be replicated after observing $R_{q'}$, so gross value weakly rises before costs. Comparing net values yields (3.12). In a differentiable parametric family, the envelope derivative of the gross value with respect to the induced posterior-score distribution gives $\mathcal{V}'(q)$, and the interior condition is $\mathcal{V}'(q) = H'(q)$. □

Proof of Proposition 6. The first statement is Topkis-Milgrom-Shannon monotone comparative statics: single crossing of the objective in (λ, θ) implies that the extremal selections of the argmax correspondence are monotone in the parameter order. For the differentiable strictly concave case with an interior optimum, the first-order condition is $W_\lambda(\lambda^*; \theta) = 0$. Differentiating gives $W_{\lambda\lambda}\partial\lambda^*/\partial\theta_j + W_{\lambda\theta_j} = 0$, and strict concavity implies $W_{\lambda\lambda} < 0$. Equation (3.15) follows. If the optimum is at 0 or 1, the first-order equality is replaced by the relevant Kuhn-Tucker inequality, so the displayed derivative formula need not apply. □

Proof of Proposition 7. Given (3.18), the AI chooses $e = \alpha\lambda L\sigma$. The patient's private problem uses perceived net compensation ξT^λ, so only the fraction ξ of gross expected liability enters the patient's adherence incentive. The patient chooses $a = \beta(1 - \xi\lambda\eta)L$. Relative to no liability, the AI-effort contribution to welfare is $L\alpha\lambda L\sigma - \frac{1}{2\alpha}(\alpha\lambda L\sigma)^2 = \alpha L^2\sigma\lambda - \frac{1}{2}\alpha L^2\sigma^2\lambda^2$. The loss from reducing patient effort is $-\frac{1}{2}\xi^2\beta L^2\eta^2\lambda^2$. Adding $R(\lambda) - K(\lambda)$ gives a concave quadratic with linear coefficient $L^2\alpha\sigma + r_1 - k_1$ and quadratic coefficient $L^2(\alpha\sigma^2 + \xi^2\beta\eta^2) + r_2 + k_2$. Maximizing over $[0, 1]$ gives (3.22). □

Proof of Corollary 2. The unconstrained maximizer of the concave quadratic in the proof of Proposition 7 is A/D. Projection onto $[0,1]$ gives no liability when $A \leq 0$, full liability when $A \geq D$, and partial liability when $0 < A < D$. □

Proof of Theorem 4. The first-best first-order conditions are $C_e + Lp_e = 0$, $B_b + Lp_b = 0$, $M_m + Lp_m = 0$, and $G_a + Lp_a = 0$. The developer's private first-order conditions are $C_e + T_e^D = 0$ and $B_b + T_b^D = 0$. The clinical enterprise's condition is $M_m + T_m^H = 0$. The patient's condition is $G_a + Lp_a - T_a^P = 0$. Equating private and social conditions gives (4.4)–(4.7). If the developer bears no transfer, premium, or internal contract, then $T_e^D = T_b^D = 0$, so its private first-order conditions cannot coincide with the first best unless the corresponding marginal harms are zero. Convexity gives sufficiency. □

Proof of Proposition 8. Under the audit-log regime, the developer minimizes

$$\frac{e^2}{2\alpha} + \frac{b^2}{2\chi} + L(\bar{p} - e).$$

The first-order conditions are $e/\alpha - L = 0$ and $b/\chi = 0$, so $e^A = \alpha L$ and $b^A = 0$. This is also the maximizer of social welfare

$$W(e,b) = Le - \frac{e^2}{2\alpha} - \frac{b^2}{2\chi},$$

so the audit-log regime implements the first best.

Under the self-generated-record regime, the developer minimizes

$$\frac{e^2}{2\alpha} + \frac{b^2}{2\chi} + L\{\bar{p} - (1 - \rho b)e - \phi b\}.$$

The Hessian of this private objective is

$$\begin{pmatrix} 1/\alpha & L\rho \\ L\rho & 1/\chi \end{pmatrix},$$

which is positive definite under (4.14). Hence the private problem is strictly convex. At $b = 0$, the safety first-order condition gives $e = \alpha L$. The derivative of the private objective with respect to b at $(e,b) = (\alpha L, 0)$ is

$$L\rho(\alpha L) - L\phi = L(\rho\alpha L - \phi).$$

If $\phi > \rho\alpha L$, this derivative is negative, so the optimum under the self-generated-record

regime has $b^S > 0$. The safety first-order condition at an interior optimum is

$$\frac{e^S}{\alpha} - L(1 - \rho b^S) = 0,$$

so

$$e^S = \alpha L(1 - \rho b^S) < \alpha L = e^A.$$

Because $W(e, b)$ is strictly concave and uniquely maximized at $(e^A, b^A) = (\alpha L, 0)$, any point with $b^S > 0$ and $e^S < e^A$ has strictly lower social welfare. Thus $W(e^S, b^S) < W(e^A, b^A)$. Finally, under the audit-log regime,

$$T_e^A = -L = Lp_e, \qquad T_b^A = 0 = Lp_b$$

when $p_b = 0$, so the audit log restores the separation between clinical safety and formal record design. □

Proof of Proposition 9. In the pure-shielding approximation, the AI chooses $e = \alpha\lambda L\sigma$, the patient chooses $a = \beta(1 - \xi\lambda\eta)L$, and the AI chooses shielding $b = \chi\lambda L\phi$. The first two terms give the same welfare contributions as in Proposition 7. Shielding has no clinical benefit and costs $b^2/(2\chi)$, so it contributes $-\frac{1}{2}\chi L^2\phi^2\lambda^2$. Adding $R(\lambda) - K(\lambda)$ and maximizing gives (4.20). If b changes σ, η, or ν, relinearize expected compensable claims around the reference allocation (e_0, a_0, b_0). The relevant local derivative is (4.21), and the coefficients multiplying true safety and patient effort are the local values $\sigma(b_0)$ and $\eta(b_0)$. Thus the same quadratic index applies only as a local approximation with all numerator and curvature terms evaluated at that reference record technology, and with any direct clinical harm from p_b added to the marginal welfare cost of liability-induced record design. □

Proof of Proposition 10. Under normal-normal updating, posterior precision equals prior precision plus the sum of signal precisions. A mass $\mu_k x_k$ of independent condition-k encounters each contributes precision $\Xi_{jk}\tau_k\varpi_k$ about θ_j. Adding these contributions across k and allowing depreciation of the old effective learning stock gives (4.23). If expected harm is increasing in posterior variance, then an increase in precision lowers posterior variance and reduces expected harm. Equivalently, if continuation value is increasing in precision, the value of an additional verified encounter is the shadow value $V_j(n)$ times its precision contribution. Hence higher $\Xi_{jk}\tau_k\varpi_k$ raises the value of an additional verified encounter. □

Proof of Theorem 5. Differentiate the Bellman objective (4.24) with respect to an interior λ_k. The current-period derivative is $\mu_k[x_{k\lambda}\mathcal{S}_k + x_k\mathcal{S}_{k\lambda}]$. The derivative of continuation value is $\delta\sum_j V_j(n_{t+1})\partial n_{j,t+1}/\partial\lambda_k$. Differentiating (4.23) gives $\partial n_{j,t+1}/\partial\lambda_k = \Xi_{jk}\mu_k[x_{k\lambda}\tau_k\varpi_k + x_k\tau_{k\lambda}\varpi_k + x_k\tau_k\varpi_{k\lambda}]$. Setting the sum equal to zero gives (4.25). □

Proof of Corollary 3. A local quadratic approximation adds the marginal dynamic term Δ_k to the linear coefficient of the static objective in Proposition 9. Projection of the resulting unconstrained maximizer onto $[0,1]$ gives (4.26). □

Proof of Lemma 2. For a finite chain, increasing differences of Π are equivalent to the claim that, for every adjacent pair $\ell' \succ \ell$, the difference

$$\Delta_{\ell'\ell}\Pi(n) = \Pi(n,\ell') - \Pi(n,\ell)$$

is nondecreasing in n. Under (4.32), this difference is

$$\begin{aligned}\Delta_{\ell'\ell}\Pi(n) = S(n)\{X(n,\ell') - X(n,\ell)\} + \{R(n,\ell') - R(n,\ell)\}\\ - \{K(n,\ell') - K(n,\ell)\} + \delta\{V(\Phi(n,\ell')) - V(\Phi(n,\ell))\}.\end{aligned}$$

The first term is nondecreasing because $S(n)$ and $X(n,\ell') - X(n,\ell)$ are nonnegative and nondecreasing. The second term is nondecreasing by assumption, and the negative of the third term is nondecreasing because the incremental cost difference is nonincreasing. The continuation term is nondecreasing by the direct continuation-difference assumption (4.34). Hence $\Delta_{\ell'\ell}\Pi(n)$ is nondecreasing for every adjacent pair, so Π has increasing differences.

For the monotonicity claim, (4.35) implies that $\Phi(n,\ell)$ is nondecreasing in n for each fixed regime. Consider $n' < n''$. If $\ell^*(n') = \ell^*(n'')$, fixed-regime monotonicity gives $\Phi^*(n'') \geq \Phi^*(n')$. If the selection switches upward from ℓ to ℓ', the no-downward-jump condition (4.36) implies that at the switch state the transition under ℓ' is at least as large as the transition under ℓ. Since each fixed-regime map is nondecreasing, and since $\mathcal{Q}(n,\ell') - \mathcal{Q}(n,\ell)$ is nondecreasing after the switch, the map cannot jump downward at an upward regime switch. Iterating over the finite number of adjacent switches between $\ell^*(n')$ and $\ell^*(n'')$ gives $\Phi^*(n'') \geq \Phi^*(n')$. Thus Φ^* is nondecreasing. □

Proof of Proposition 11. By increasing differences of $\Pi(n,\ell)$ and because $\mathcal{L}$ is a finite chain, the largest optimal regime selection $\ell^*(n)$ is nondecreasing in n by monotone comparative statics. Define $\Phi^*(n) = \Phi(n,\ell^*(n))$. By assumption, Φ^* is nondecreasing.

The first two inequalities in (4.38) imply that Φ^* maps $[n_1, n_2]$ into itself. For any $n \in [n_1, n_2]$, monotonicity gives

$$\Phi^*(n) \geq \Phi^*(n_1) \geq n_1 \quad \text{and} \quad \Phi^*(n) \leq \Phi^*(n_2) \leq n_2.$$

Thus $[n_1, n_2]$ is forward invariant. Since this interval is a complete lattice and Φ^* is nondecreasing, Tarski's fixed-point theorem gives a fixed point in $[n_1, n_2]$. The same argument applied to $[n_3, n_4]$ gives a second fixed point in $[n_3, n_4]$, and the two fixed points are distinct because $n_2 < n_3$. Forward invariance of $[n_3, n_4]$ follows by the same inequalities.

If Φ^* is a contraction on one of the invariant intervals, Banach's fixed-point theorem implies that the fixed point in that interval is unique and that every path starting in the interval converges to it. Applying the same argument to each trap interval gives the stated convergence refinement.

Let $n^L \in [n_1, n_2]$ and $n^H \in [n_3, n_4]$ be the two fixed points. Since $n^H > n^L$ and $\ell^*(n)$ is nondecreasing, $\ell^*(n^H)$ is weakly above $\ell^*(n^L)$ in the liability order. If λ_ℓ is increasing in that order, liability power is weakly higher at the high-quality fixed point. If aggregate adoption $X(n, \ell)$ is increasing in both arguments, adoption is also weakly higher at the high-quality fixed point. Strictness follows whenever the selected regime or adoption response differs strictly across the two fixed points. □

Proof of Corollary 4. For each regulation level r, the nondecreasing map Φ_r^* has a nonempty complete lattice of fixed points by Tarski's fixed-point theorem. If $\Phi_{r'}^*(n) \geq \Phi_r^*(n)$ for all n, then the standard monotone fixed-point comparative static implies that the least and greatest fixed points of the largest-selection map are weakly increasing in r. If $\Phi_{r'}^*(n) > n$ throughout a lower interval, no fixed point of that largest-selection map can lie in the interval. To rule out fixed points generated by non-largest also-optimal selections, the stronger condition stated in the corollary requires every optimal selection to move the state upward on that interval. □

Proof of Proposition 12. The planner's discrete objective over covered sets is exactly (4.50). Comparing a candidate set J to $J \cup \{h\}$ gives the incremental value (4.51); adding h is optimal iff that incremental value is nonnegative. □

Proof of Proposition 13. Agent i chooses x_i to minimize $C_i(x_i) + \rho_i T(x_D, x_H, a)$, taking the other agent's action as given. At an interior Nash equilibrium without internal contracts, $C_i'(x_i) + \rho_i T_{x_i} = 0$. The first-best condition is $C_i'(x_i) + L p_{x_i} = 0$. Equating the two

conditions yields (4.52). If the common claim and share constraint cannot satisfy these equations for both agents, external liability alone cannot match both first-order incentives. With an internal contract t_i, agent i's first-order condition becomes

$$C_i'(x_i) + \rho_i T_{x_i} + \frac{\partial}{\partial x_i}\mathbb{E}[t_i(R_i) \mid x_i, x_{-i}^*] = 0$$

at the target allocation. Under unrestricted signed transfers, the set of internal marginal incentives that can be generated is exactly $\mathcal{I}_i(\mathcal{G}_i)$. Hence first-order implementation is possible exactly when the missing incentives in (4.54) belong to $\mathcal{I}_D(\mathcal{G}_D) \times \mathcal{I}_H(\mathcal{G}_H)$. Convexity or a local second-order sufficient condition then turns these first-order equalities into local Nash implementation. If the internal contracts are bounded or subject to limited liability or budget constraints, the same argument applies after replacing each linear span by the feasible set generated by the restricted contract class. □

Proof of Proposition 14. The total value of standalone assignment is $\sum_k \mu_k v_k^S - \sum_k F_k^S + \Omega^S$. The total value of enterprise assignment is $\sum_k \mu_k v_k^E - \sum_k F_k^E + \Omega^E$. Enterprise integration is efficient iff the latter weakly exceeds the former, which is (4.59). The amortized-cost expression follows by substituting $f_k^q = F_k^q/\mu_k$. □

Proof of Proposition 15. Equation (4.62) follows by substituting (4.60) into the conduct condition (4.61). Differentiate (4.63) using Leibniz's rule. The derivative of the integrand with respect to q is $v_{kq} - c_{kq} + h_{kq} + \ell_{kq}$. The boundary term equals $x_{kq}\{v_k(q) - B_k x_k - c_k(q) + h_k(q) + \ell_k(q)\}$. By the conduct condition, $v_k(q) - B_k x_k - c_k(q) = \kappa_k B_k x_k$. Combining the integrand derivative, boundary term, and fixed-cost derivative gives (4.64). □

Proof of Proposition 16. Before direct prescribing the local quadratic objective has linear coefficient A and curvature D, so the interior optimum is A/D. The medication block adds linear coefficient A^M and curvature D^M. The new optimum is therefore $(A + A^M)/(D + D^M)$. This exceeds A/D iff $D(A + A^M) > A(D + D^M)$, which reduces to $A^M/D^M > A/D$ when $D, D^M > 0$. □

Proof of Proposition 17. Continuity on the compact set $[0, 1] \times \Theta$ implies that $W^R(\lambda) = \inf_{\vartheta \in \Theta} W(\lambda; \vartheta)$ is upper semicontinuous, and hence a robust maximizer exists on $[0, 1]$. Let $\bar{\lambda}$ be the largest maximizer of $W(\cdot; \bar{\vartheta})$. By concavity, $W_\lambda(\lambda; \bar{\vartheta}) \leq 0$ for all $\lambda > \bar{\lambda}$. Under (4.71), $W_\lambda(\lambda; \vartheta) \leq 0$ for all $\vartheta \in \Theta$ and all $\lambda > \bar{\lambda}$, so each $W(\cdot; \vartheta)$ is weakly decreasing on $[\bar{\lambda}, 1]$. Therefore W^R is also weakly decreasing on $[\bar{\lambda}, 1]$. Any robust maximizer above $\bar{\lambda}$ can be moved down to $\bar{\lambda}$ without lowering robust value, so there exists a robust maximizer weakly below $\bar{\lambda}$. The reversed inequality gives the symmetric existence result relative to

the smallest benchmark maximizer. Under rectangular ambiguity over legal-score cells, the adversary can choose the worst conditional score cell by cell. The same pointwise argument as in Theorem 3 then applies after replacing the conditional score by the worst-case conditional score (4.72). □

References

Arrow, Kenneth J. 1963. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review* 53(5): 941–973.

Calabresi, Guido. 1970. *The Costs of Accidents: A Legal and Economic Analysis*. New Haven: Yale University Press.

Carroll, Gabriel. 2015. "Robustness and Linear Contracts." *American Economic Review* 105(2): 536–563.

Chan, Alex, Yetong Xu, and David M. Cutler. 2026. "Patient Preferences for Artificial Intelligence in Primary Care: An Adaptive Price-List Experiment." Unpublished manuscript.

Currie, Janet, and W. Bentley MacLeod. 2008. "First Do No Harm? Tort Reform and Birth Outcomes." *Quarterly Journal of Economics* 123(2): 795–830.

Eliason, Paul J., Benjamin Heebsh, Ryan C. McDevitt, and James W. Roberts. 2020. "How Acquisitions Affect Firm Behavior and Performance: Evidence from the Dialysis Industry." *Quarterly Journal of Economics* 135(1): 221–267.

European Commission. 2026. "AI Act." Shaping Europe's Digital Future policy page. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

European Parliament and Council of the European Union. 2024a. "Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence." Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

European Parliament and Council of the European Union. 2024b. "Directive (EU) 2024/2853 on Liability for Defective Products." Official Journal of the European Union. https://eur-lex.europa.eu/eli/dir/2024/2853/oj.

Food and Drug Administration, U.S. 2025. "Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions." Final guidance. https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles.

Food and Drug Administration, U.S. 2026a. "Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff." Final guidance. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software.

Food and Drug Administration, U.S. 2026b. "Artificial Intelligence-Enabled Medical Devices." FDA web resource and device list. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.

Frakes, M., and Gruber, J.. 2019. "Defensive medicine: evidence from military immunity." *American Economic Journal: Economic Policy* 11(3), 197-231.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny. 2015. "Money Doctors." *Journal of Finance* 70(1): 91–114.

Gerke, Sara, Timo Minssen, and I. Glenn Cohen. 2020. "Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare." In *Artificial Intelligence in Healthcare*, 295–336. Academic Press.

Glaeser, Edward L., Simon Johnson, and Andrei Shleifer. 2001. "Coase versus the Coasians." *Quarterly Journal of Economics* 116(3): 853–899.

Glaeser, Edward L., and Andrei Shleifer. 2002. "Legal Origins." *Quarterly Journal of Economics* 117(4): 1193–1229.

Glaeser, Edward L., and Andrei Shleifer. 2003. "The Rise of the Regulatory State." *Journal of Economic Literature* 41(2): 401–425.

Grossman, Sanford J. 1981. "The Informational Role of Warranties and Private Disclosure about Product Quality." *Journal of Law and Economics* 24(3): 461–483.

Hacker, Philipp, Andreas Engel, and Marco Mauer. 2023. "Regulating ChatGPT and Other Large Generative AI Models." *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency.*

Holm, Sune, Catherine Stanton, and Oliver Bartlett. 2021. "A new argument for no-fault compensation in health care: the introduction of artificial intelligence systems." *Health Care Analysis* 29(3), 171-188.

Holmstrom, Bengt. 1979. "Moral Hazard and Observability." *Bell Journal of Economics* 10(1): 74–91.

Jewitt, Ian. 1988. "Justifying the First-Order Approach to Principal-Agent Problems." *Econometrica* 56(5): 1177–1190.

Kessler, Daniel, and Mark McClellan. 1996. "Do Doctors Practice Defensive Medicine?" *Quarterly Journal of Economics* 111(2): 353–390.

Landes, William M., and Richard A. Posner. 2013. *The Economic Structure of Tort Law.* Cambridge, MA: Harvard University Press.

Lior, Anat. 2022. "Insuring AI: The Role of Insurance in Artificial Intelligence Regulation." *Harvard Journal of Law and Technology* 35(2).

Mello, Michelle M., et al. 2024. "Understanding Liability Risk from Using Health Care Artificial Intelligence Tools." *New England Journal of Medicine.*

Milgrom, Paul, and Chris Shannon. 1994. "Monotone Comparative Statics." *Econometrica* 62(1): 157–180.

Poggi, Francisco, and Bruno Strulovici. 2020. "Liability Design with Information Acquisition." arXiv:2012.05066.

Polinsky, A. Mitchell, and Steven Shavell. 2007. "The Theory of Public Enforcement of Law." In *Handbook of Law and Economics*, vol. 1, 403–454. Elsevier.

Price, W. Nicholson, Gerke, Sara, and I. Glenn Cohen. 2024. "Liability for use of artificial intelligence in medicine" *Research handbook on health, AI and the law* 150-166.

Price, W. Nicholson, Sara Gerke, and I. Glenn Cohen. 2019. "Potential Liability for Physicians Using Artificial Intelligence." *JAMA* 322(18): 1765–1766.

Price II, W. Nicholson, and I. Glenn Cohen. 2024. "Locating Liability for Medical AI." *DePaul Law Review* 73, no. 2.

Rogerson, William P. 1985. "The First-Order Approach to Principal-Agent Problems." *Econometrica* 53(6): 1357–1367.

Shavell, Steven. 1980. "Strict Liability versus Negligence." *Journal of Legal Studies* 9(1): 1–25.

Shavell, Steven. 2009. *Economic Analysis of Accident Law*. Cambridge, MA: Harvard University Press.

Sloan, Frank A., and Lindsey M. Chepke. 2008. *Medical Malpractice*. Cambridge, MA: MIT Press.

Sloan, Frank A., and John H. Shadle. 2009. "Is There Empirical Evidence for Defensive Medicine? A Reassessment." *Journal of Health Economics* 28(2): 481–491.

Spence, A. Michael. 1977. "Consumer Misperceptions, Product Failure and Producer Liability." *Review of Economic Studies* 44(3): 561–572.

Studdert, D. M., Mello, M. M., Sage, W. M., DesRoches, C. M., Peugh, J., Zapert, K., and Brennan, T. A.. 2005. "Defensive medicine among high-risk specialist physicians in a volatile malpractice environment." *JAMA* 293(21), 2609-2617.

Sullivan, Hannah R., and Scott J. Schweikart. 2019. "Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI?" *AMA Journal of Ethics* 21(2): 160-166.