When the Robot Is Wrong, Someone Pays

Why the underwriter, not the regulator, will decide which robots get to work

Allan Watkins

Laws of Robots series

Version 1.1 — 13 July 2026

Abstract

Autonomous machines are starting to make real decisions right next to vulnerable people — in homes, in care facilities, often with no one watching. That unsupervised work is exactly why these machines have commercial value. When one of them chooses wrongly and someone gets hurt, the only record of why it made that choice is controlled, formatted, and vouched for by the very party whose product is on trial. This paper argues that this single structural fact — not how advanced the machine is, and not what regulators do — is what currently makes the risk uninsurable. And the people best positioned to fix it are the underwriters.

Here is how the argument unfolds. Losses from these machines fall into two clear channels. The Physical Loss Channel has a hundred years of actuarial experience behind it. The Decision Loss Channel — where the machine's judgment fails even though its motors and sensors work perfectly — has almost none. The Decision Loss Channel cannot be insured today not because judgment failures are impossible to measure, but because no trustworthy instrument actually measures them. The industry's usual answer, Self-Attested Telemetry, simply cannot prove its own integrity to anyone on the other side of the table, no matter how much data it produces. A real example of that failure already exists in federal court. What the channel actually needs is a Decision Ledger: an append-only, tamper-evident, longitudinal, reconstructable record of every important decision the machine makes. Anyone can check its integrity without having to trust the maker. Two features of that record change the underwriting math in a practical way. It makes a Bounded Claim — the record says in plain words exactly what it does *not* vouch for, which is what makes its positive statements priceable. And it records the owner's declared Consequence Election when the machine spots its own fault — a clear moral-hazard signal that underwriters already price in other lines.

This paper states its own honest posture up front and keeps it throughout. What exists today is an implemented instrument whose integrity can be verified by mathematics, offline, by anyone. What does not yet exist is real loss experience or complete institutional independence for the certifying function. The claim here is deliberately narrow: this is the instrument that finally generates the loss data the Decision Loss Channel has never had. Insurance has built verification systems before

when a loss channel demanded them. The robots that actually get put to work will be exactly the ones someone was willing to insure when they are wrong.

1. The Reconstruction Problem

Consider an autonomous care robot working through the night, unsupervised, in the home of an elderly woman who lives alone. Unsupervised operation is not a corner the maker cut. It is the entire commercial proposition: the machine exists so that she can keep living in her own home, and so that no human carer has to be in the room at 3 a.m.

One night she loses her balance in the hallway. The robot registers the event and faces a judgment with no clean answer. It can attempt to physically steady her — applying force to a frail body — or it can hold back and summon help, accepting the consequences of an unbroken fall. It chooses to intervene, and in steadying her it fractures her wrist. The motors and sensors performed correctly throughout. The machine perceived its situation accurately, made a choice between competing harms, and the choice produced an injury.

By morning there is a claim. That brings the questions anyone who must pay for, defend, or price this event actually cares about. Was the fall genuinely underway, or did the machine mistake a stumble for a collapse? Did it weigh her fragility before it acted? Which rule governed the choice — a sensitivity setting the care agency selected, a preference the family configured, or the factory default? Was this the machine behaving exactly as designed, inside the envelope its maker specified, or was it a defect? And if a defect: a freak event, or the first visible instance of a flaw sitting latent in every identical unit running the same logic in hundreds of other homes?

When a machine fails mechanically, a forensic engineer works backward from the wreckage — bent metal, fracture surfaces, tool marks. The evidence is physical, and it does not change its story. A judgment leaves no wreckage at all. The only trace of *why the machine chose what it chose* is whatever record the machine's builder or operator decided to keep, written in a proprietary format that party controls, attested to by that party alone. The party whose product is on trial is also the sole author, custodian, and interpreter of the evidence that would establish whether the product is at fault.

That condition has a precise consequence for anyone who prices risk: **you cannot underwrite what you cannot reconstruct.** An underwriter does not need the machine to be perfect; underwriters price imperfect things for a living. What an underwriter needs is a truthful, reconstructable account of what went wrong and how often — a loss history that can be trusted because it does not depend on the good faith of the party being charged for it. For autonomous machines that make consequential choices around vulnerable people, no such account currently exists in a form an adverse party can rely on. ***The record is real, and it belongs to the defendant.***

It is worth being clear about where this problem is least resolved. Autonomous vehicles at least operate inside a century of motor-liability precedent, with event-recorder conventions and a

growing body of litigation. A machine that exercises judgment and applies force in close proximity to a frail human being, unsupervised, in a private home, has almost none of that. This paper is about that gap — what it costs, why the law is already straining to work around it, why the industry's default fix does not close it, and what a record built to close it would have to look like.

2. Why Existing Records Fall Short

The gap has not gone unnoticed. The legal system in particular has spent several years visibly straining to compensate for a record that does not exist, and the shape of that strain is itself informative.

The natural first stop is the operator's own data. Autonomous systems already generate extensive telemetry, and the parties that build and run them retain it. But this is **Self-Attested Telemetry** — operational data whose integrity and completeness rest on the word of the party that generates it, which is usually the same party whose risk is being priced. It is produced by the entity with the strongest interest in what it says, stored on that entity's infrastructure, formatted by that entity's engineers, and produced — or not — at that entity's discretion when a dispute arises. For the operator's own purposes the data is genuinely valuable. As a basis for an *adverse* party to price that operator's risk, it has a defect that volume does not repair: its integrity cannot be independently verified, and its completeness cannot be independently confirmed. Section 5 returns to this with a litigated example.

The law's response is more revealing. In the European Union, the jurisdiction moving fastest on this question, the first instinct was a dedicated instrument. The proposed AI Liability Directive would have eased the victim's burden of proof and empowered courts to order disclosure of evidence about high-risk AI systems. Tabled in 2022, it collapsed: the Commission listed it for withdrawal in its 2025 Work Programme, and the withdrawal was formally recorded in the Official Journal in October 2025, for lack of any foreseeable agreement. The pressure did not disappear with the instrument. It moved into the revised Product Liability Directive (Directive (EU) 2024/2853), which for the first time treats software and AI systems as "products" under Europe's strict-liability regime, and which member states must transpose into national law by December 2026.

The contents of that revised Directive read like a map of the missing record drawn in negative. It carries claimant-friendly presumptions — mechanisms that shift the burden of proof toward the manufacturer precisely where the technical complexity of the product makes it unreasonable to expect a victim to prove causation directly. And it introduces a court-ordered disclosure mechanism: a claimant who can show a claim is plausible can compel the defendant to hand over relevant evidence in the defendant's possession. It is worth asking why a legislature needs to build such a thing. Nobody legislates court-ordered disclosure of a record a claimant can obtain independently. The mechanism exists because the record sits with the defendant and cannot be reached any other way. The European Parliament's own analysis of the abandoned liability

instrument conceded the diagnosis: the opacity, autonomy, and continuous adaptation of these systems make it *particularly difficult* for a victim to meet the burden of proof.

There are also the documentation duties. The EU AI Act (Regulation (EU) 2024/1689) imposes record-keeping, logging, and human-oversight obligations on high-risk systems — the closest thing in current law to a mandate that a decisional record exist at all. Two observations. First, the timing is unsettled and receding: the main high-risk obligations were deferred by the 2026 "Digital Omnibus" agreement, with the standalone high-risk categories now anchored to late 2027 and product-embedded systems to 2028, pending formal adoption at the time of writing. Second, and more fundamentally, even in full force these duties mandate that the *operator* keep the logs. An operator-kept log is better than no log. It is not an independently verifiable record. The law here requires the defendant to document its own defense. That is a real improvement in transparency yet still self-attestation by another name.

Presumptions, disclosure orders, documentation mandates: three different tools with one shared premise. Each works around the same absent thing — an independently verifiable, reconstructable account of what an autonomous system decided and why. ***The law is building scaffolding around a hole.***

(Legal status current as of drafting; specific dates and instrument statuses to be re-verified at publication — see drafting notes.)

3. Two Distinct Loss Channels

Pricing this risk becomes more tractable once it is separated into two channels that behave differently. Much of the current confusion comes from conflating them.

The first is the **Physical Loss Channel** — monetary loss arising from physical harm a machine's body causes in the world. A machine with mass and momentum collides with a person or an object, applies force wrongly, or fails mechanically. This channel is not new to insurance. It has a long actuarial ancestry in industrial machinery, powered equipment, and the body of premises and product liability that grew up around moving hardware. The exposure bases are understood, the failure modes are physical, and the loss experience — though it will need adjustment for autonomous operation around people — rests on a foundation that already exists.

The second is the **Decision Loss Channel** — monetary loss arising not because the machine's body failed but because its judgment did. The motors and sensors work correctly. The machine perceives its situation accurately but then chooses badly: it prioritized the wrong objective, misweighed a risk, or applied an owner's rule in a context where the rule did not belong. The care robot in Section 1 sits squarely in this channel — it perceived the fall correctly and then decided between competing harms in a way that produced injury. There is almost no actuarial precedent here, because until recently no deployed machine made autonomous choices consequential enough to generate a distinct loss history. It is also the channel that scales in the way underwriters least like: a mechanical fault injures the person in front of one machine, while a fault in judgment

is potentially replicated in every identical unit running the same logic, waiting for the same triggering circumstance.

Why is the Decision Loss Channel uninsurable today? Not, as is often assumed, because judgment failures are unquantifiable in principle. They have structure; they recur; they cluster around identifiable conditions. The channel is uninsurable for a more mundane reason: no instrument measures it. There is no trustworthy, independent record of when these machines exercised judgment, what they concluded, on what basis, and where the judgment broke down. Distinguishing the two channels clarifies where today's insurance problem actually lies. The Physical Loss Channel already has decades of actuarial history behind it; the Decision Loss Channel has almost none, because it lacks the instrument needed to generate trustworthy experience.

4. What a Verifiable Record Provides

Suppose the Decision Loss Channel were instrumented. Suppose every consequential choice an autonomous machine made left an entry in a **Decision Ledger** — an append-only, tamper-evident record of the machine's decisions, accumulating over its operating life. Four properties matter most. First, the ledger is *tamper-evident*: any later alteration is detectable. Second, it is *append-only*: history cannot be quietly rewritten. Third, it is *longitudinal*: it accumulates across the machine's life rather than surviving only as a snapshot around an incident. Fourth, it is *reconstructable*: for any given decision, one can recover what the machine concluded, which of its owner's rules governed the conclusion, on what evidence, and at what stated confidence.

How such a ledger is built is a separate subject. It exists as an implemented system, and this paper draws on that implementation as a matter of fact, but the mechanics are not the argument here. The argument concerns what a record with those four properties *is*, in the terms an actuary cares about: structured, longitudinal, tamper-evident telemetry of decisional conduct — the raw material of loss modeling for a channel that has never had any. A rating model cannot be built on anecdotes and post-hoc reconstructions contested by the defendant. It can begin to be built on a longitudinal ledger of decisions and outcomes that does not depend on the good faith of the insured to be believed.

Each property translates directly. Tamper-evidence means the account being priced against is the account written at the time, not the account most convenient once a claim arrived; any after-the-fact curation is detectable. Append-only means the loss history is cumulative and cannot be silently pruned of its inconvenient entries — the base being rated on is the whole base. Longitudinal means the underwriter is not pricing a single dramatic incident in isolation but observing decisional conduct over time, which is where the leading indicators of a developing problem actually live. Reconstructable means a loss can be attributed to a specific decision and its specific basis, which is the precondition for distinguishing a one-off from a systematic defect sitting in an entire fleet.

One further property deserves early mention because it changes how the whole ledger should be read. A record of this kind is built to be honest about the limits of what it establishes — it does not assert more than it can prove. That discipline, a record stating plainly what it does and does not attest, is developed in Section 6; it is the feature that makes the record's positive claims safe to price against.

This paper makes the narrower claim, and deliberately so. It does not assert that a Decision Ledger reduces losses, and it does not assert that enough ledger data yet exists to constitute experience — the honest posture on accumulation is set out in Section 8, and it is a modest one. The claim is prior to both: this is the instrument that generates the loss data the Decision Loss Channel requires, an instrument that has so far been absent.

5. Why Verifiable Integrity Matters More Than Volume

The industry's own answer to the missing record is already arriving, and it deserves a plain statement because it is the default this paper questions. The emerging model streams operational data from the machine back to its manufacturer, continuously and at scale. The instinct is natural and the volumes are genuinely enormous. The question is whether quantity is the property that matters. For an actuary pricing an adverse risk, it is not.

Streaming produces Self-Attested Telemetry in bulk, and bulk is precisely what does not help. Self-attested data is held and vouched for by the party whose risk is being priced; a larger volume of it is a larger volume of data whose integrity and completeness still rest on the trustworthiness of the party with the strongest interest in what it says. What the actuary lacks is not more of the operator's data. It is data whose integrity can be verified *independently* of the operator — and, because these machines will often run disconnected, data whose integrity survives offline, without a live connection back to the very party being assessed.

The distinction is no longer theoretical. We now have a real litigated example from the car world. In the litigation following a fatal 2019 crash involving Tesla's driver-assist system (the Benavides matter, tried in federal court in Florida), the central evidentiary fight was not about a shortage of data. The car in question had, within minutes of the crash, uploaded a "collision snapshot" — video, vehicle-bus streams, event-data-recorder information — to the manufacturer's servers and received an acknowledgement of receipt. The fight was about access and integrity. During the proceedings Tesla maintained that it could not recover the specific system data from the vehicle. The plaintiffs' expert forensically recovered it anyway, and it showed the system had detected the hazard seconds before impact and did not brake. In August 2025 a federal jury returned a verdict exceeding $240 million, and in 2026 the trial judge declined to set the verdict aside. *(The matter remains under appeal; Tesla disputes causation and maintains the driver overrode the system by depressing the accelerator. The company's chief executive rejected the suggestion that the software was responsible — and the matter continues to be litigated.)*

For this audience, the point of the example is deliberately not that a manufacturer behaved badly. The point is structural, and it holds however the causation question resolves: the party being

assessed controlled the record, and its representations about that record's availability were contested and, in the event, incorrect. This is among the largest and most sophisticated telemetry programs of its kind, and it is precisely the program in which the integrity and availability of the telemetry became the whole fight. No quantity of data answers the question the actuary is actually asking — which is not "how much did the machine record?" but "can I trust the record I am shown, and can I confirm I am being shown all of it?" Self-attested telemetry, however voluminous, cannot answer that question about itself.

(Note on the example in this section: This uses publicly reported facts from ongoing litigation for illustrative purposes only. The case remains under appeal, and all parties' positions are noted. Nothing in this paper constitutes a legal opinion or finding on any specific litigation. This paper is not legal advice.)

Aviation confronted a version of this problem decades ago and did not solve it with volume. The flight recorder is mounted in the aircraft, but its contents are read out by investigators under standardized procedures — the airline does not get to narrate what the recorder says. Autonomous machines have no equivalent: the recorder, the format, and the narration all belong to the party under scrutiny.

Now transpose the litigated example. Cars are the domain with event-recorder conventions, regulatory scrutiny, and a jury verdict on exactly this issue. The machines this paper concerns — exercising judgment and applying force around vulnerable people, in homes and care settings — have none of that. The reconstruction problem there is not milder. The choices are richer and less scripted than lane-keeping, the operating conditions less standardized than a public road, the humans in proximity more fragile, and the body of precedent that eventually forced the Tesla data into open court essentially absent. If self-attested telemetry could not be trusted to answer the question in the one domain that had every advantage, it is difficult to see how it answers the question in a domain that has none of them. The car case is not the subject of this paper; it is evidence, borrowed from next door, that the failure mode is real. And the honest posture requires saying the rest plainly: transposing a failure mode is not the same as possessing loss experience for robots. That experience does not yet exist. What exists is the demonstrated failure of the alternative, and an instrument built so the failure need not repeat.

The instrument's value rests on two properties the streaming model lacks, and they must be kept distinct because they sit at different stages of maturity. The first is **verifiable integrity**: the ledger is constructed so that an adverse party can confirm, mathematically, that it has not been altered — without trusting, or even contacting, the party that produced it. This property is available today. It is a matter of construction, and tamper-evidence does not care who operates the machinery. The second is **independent attestation**: the ledger's integrity vouched for by a party with no stake in the verdict — structurally separate from the operator, unable to bless a record the operator would prefer, and unable to be overruled by it. That separation, of the party that builds the ledger from the party that certifies it, is an architectural commitment and an operating principle in the implementation this paper draws on; it is not yet a fully realized institutional reality, and Section 8 states plainly where it stands. For the actuary the practical

consequence is this: the integrity verifiable today is mathematics anyone can check, while the independence that would make a certification a true third-party assurance is a maturing capability, and should be weighed as one. A record that is honest about that distinction is, not coincidentally, the kind of record whose other claims can be afforded belief.

6. What the Record Refuses to Say

Every attestation from the record's certifying function states both what it asserts and what it does not assert. It does not assert that the machine's decisions were wise, that the owner's configured rules were sensible, or that the machine is safe. What it asserts is narrow and checkable: the record is genuine, its integrity is intact, and the stated validity conditions held at the time of attestation. This paper calls such an attestation a **Bounded Claim**.

At first hearing this sounds like a weakness in the product. In underwriting terms it is closer to the opposite, for a reason the profession already applies to warranties and representations: **a certification that vouches for everything vouches for nothing.** An unbounded assurance cannot be priced, because its failure modes cannot be enumerated. If a stamp is taken to mean "this machine is safe," then the reliability of the stamp itself becomes an unquantifiable input to the rate — safe against what, under which operating conditions, on whose evaluation of the owner's configured rules? Every ambiguity in the assurance becomes ambiguity in the price, and the standard response to an assurance that cannot be bounded is the correct one: discount it toward zero and rate as if it were not there.

A Bounded Claim works the other way around. When the attestation asserts only enumerable, checkable properties, its failure modes are enumerable too, and an underwriter can treat its positive assertions at close to face value while pricing the residual — decision quality, rule quality, deployment context — as named risk rather than as risk hidden inside an overclaiming stamp. The residual does not disappear; it becomes visible on the rating sheet instead of buried in the certification. The working parallel is familiar from existing practice: an underwriter relies on a laboratory listing for exactly the properties the laboratory tested, and rates the remainder of the fire risk explicitly. Nobody treats the listing as a warranty that the building will not burn, and the listing is useful precisely because nobody has to.

The institutional arrangement behind the attestation can be stated in two facts, and the paper deliberately states no more than two. First, the ledger is one thing: a data source, implemented, generating the record. Second, the attestation of that record's integrity is a separate function, structurally distinct from the party that builds and operates the ledger, unable to be overruled by the operator, and certifying a property — record integrity and validity conditions — never a product. Today that certifying function is operated by the same small organization that built the system, so the separation is architectural and operational rather than institutional, and an underwriter should weigh the attestation accordingly while the institutional distance is built out.

7. From Record to Rating

What would an underwriter actually do with a Decision Ledger? This section lays out one workable structure for using it in rating. No loss experience exists yet for the Decision Loss Channel, so no rates can be quoted today. The narrower goal is to show that every element of a practical rating plan — how much judgment the machine exercised, how well it exercised it, the owner's declared posture when something goes wrong, and how losses get attributed — can be counted and verified directly from the ledger. Each element already has clear parallels in how insurers rate other kinds of risk. When actual loss data begins to arrive, the structure will be ready to receive it.

Exposure bases. The standard tests for an exposure base apply here as anywhere: it should be proportional to expected loss, practical to measure, and resistant to manipulation by the insured. Operating hours pass the second test and fail the first — a machine can run all day and exercise judgment rarely, and two deployments with identical hours can carry very different decisional exposure. Interactions with people do better on proportionality, since proximity to people is where this channel's severity lives, but until now they have failed the practicality test because nobody could count them reliably. Decisions rendered — the count of consequential judgments the machine actually made — is the natural base. Each rendered decision is one opportunity for the failure mode being insured, so frequency per decision is the rate at which the channel actually generates losses. The reason this base has never appeared on a rating sheet is simple: without a ledger there is no way to count it. The count becomes available, and becomes trustworthy, only because the instrument exists.

The base also has a property that self-reported exposure bases lack. Payroll, mileage, and revenue are declared by the insured and trued up at premium audit, with the leakage and disputes that follow. A ledger-based decision count is verifiable by the carrier directly, continuously, without the insured's cooperation — the exposure audit is a query, not a negotiation. In practice the base would be grouped by the type of decision being made, because a decision that applies physical force to a person, such as steadying someone who is falling, carries a very different level of severity than a navigation decision, such as choosing a route around an obstacle. A rating plan would set a base frequency rate per ten thousand decisions of each type. As the fleet builds its own claims history, that experience gets blended with the broader class rates using standard credibility methods — the same approach insurers already use for other lines of business.

An exposure base tells an underwriter how much judgment a machine exercised. The next question is how well it exercised that judgment. The ledger supplies clear, observable signals for exactly that.

Behavioral leading indicators. Motor underwriting learned, through telematics (like hard-braking events, cornering, handheld-phone use per mile) to price on observed behavior because those signals predict claim frequency before claims arrive. The Decision Ledger supplies the decisional equivalent, with one difference worth stating: telematics measures mechanical handling of a vehicle, while these signals report on the machine's judgment itself — what it concluded, under which rule, and at what confidence. Three signal families and their use:

Withheld decisions — occasions when the machine declined to act because its own reasoning checks did not pass. The per-thousand-decisions rate of these events is a competence-boundary

signal: a machine withholding frequently is operating near the edge of what it can reliably judge in that deployment. An underwriter would use the level of these events to compare deployments at submission and the trend to monitor the risk while the policy is in force. A sudden jump in the withheld rate after a software version change, for example, is a warning at the version level that arrives before any loss occurs — exactly what a leading indicator is meant to provide.

Self-declared integrity faults — events where the machine flagged its own judgment as unsound. These events support several policy mechanics that already exist in other lines of insurance:

- A notification covenant requiring the insured to report faults above a stated threshold.
- Mid-term review triggers, which allow the carrier to reassess the risk and potentially adjust terms if the frequency of self-reported integrity faults rises significantly during the policy period.
- Time-to-resolution as a management-quality signal, the way loss-control response time is evaluated in commercial property insurance.

Certification-state changes — the attestation lapsing, being revoked, or being restored, with the reason recorded. The closest existing analog is the inspection certificate in boiler and machinery: operating without a current certificate is condition-of-coverage territory. The material difference is visibility. Today, an uncertified machine in service is invisible to its carrier until a loss arrives. With the state change logged in a verifiable ledger, the carrier can clearly see any uncertified operation and the policy can respond to it — whether as a condition, an exclusion, or a priced surcharge — at the carrier's discretion.

Unlike an application questionnaire, none of these signals depends on the candor of the party whose risk is being priced. They are read from the record and verified against it.

The signals above are observed behavior. The record also carries a declared behavior, and it may be the most insurable fact in this paper.

Consequence Election as a rating variable. When a machine of this kind detects a fault in its own judgment, something happens next, and the owner decides in advance what that something is: the machine can fail closed — stop, go dormant, summon help — or the owner can elect to continue operating it uncertified, visibly, at the owner's declared liability. That declared, recorded commitment is the **Consequence Election**, and it is insurable for three straightforward reasons.

- It works like a protective-safeguards warranty in property insurance. A home with a working sprinkler system gets a premium credit because the sprinklers reduce the chance and size of a loss. Here, a fail-closed election lowers the risk in the same way and can earn a credit. A continue-uncertified election increases the risk and gets priced as the materially different exposure it is.
- It reveals the owner's risk appetite in advance. The way an underwriter reads a high deductible or a named-operator restriction on an auto policy, this election tells the

underwriter something true about how the owner weighs availability against safety — before any loss occurs.

- It is verifiable after the fact. The usual weakness of a protective-safeguards warranty is that the carrier learns the sprinklers were disabled only during the fire investigation. Here the election, and any change to it, is itself a recorded event. An owner cannot quietly loosen the posture, because loosening leaves an entry in the ledger.

Policy mechanics operate as follows:

- A rate differential applies between elections;
- Any change in election is treated as an endorsement-level event that triggers re-rating; and
- Losses incurred during owner-elected uncertified operation are handled using the carrier's preferred approach. This can include exclusion, a sub-limit, or priced acceptance. The operative facts are established by the record rather than disputed in discovery.

Exposure, behavior, and declared posture rate the book while it runs. The remaining question is what the record does when a loss actually arrives.

Loss attribution and the fleet question. The scaling fear stated earlier in this paper — a judgment defect replicated across every identical unit — is, in reserving terms, the difference between one claim and an emerging accumulation. Today a carrier facing the wrist-fracture claim cannot tell which it holds. That uncertainty has a real cost. Carriers set wide case reserves and load extra IBNR (Incurred But Not Reported) reserves against the entire class, and at the portfolio level the rational decision is often not to write the class at all. Attribution changes the mechanics. The ledger ties the loss to the exact decision — the judgment, its governing rule, its software version, its stated confidence — and the same decision type can then be queried across the fleet's ledgers. If the pattern is isolated, the claim is a single occurrence and the reserve can be set more tightly around that single claim instead of loading extra for the whole class. If the pattern recurs on a version, the carrier is looking at a version-level defect, and three consequences follow that this market currently cannot reach.

The accumulation becomes definable. Losses attributable to the same decision-type defect in the same software version is language an occurrence definition or an aggregation clause can actually use. This matters to any reinsurer asked to accumulate robot exposure. The defect becomes a product-liability matter with a named cause, allowing subrogation by the operator's carrier against the manufacturer, with the governing rule and version identified in a record neither party controls. And the fleet decision becomes empirical: recall, version-freeze, or continue, decided from clear fact recorded in the ledger.

The product liability insurance market offers a useful comparison. That line runs on roughly $4–5 billion in annual net written premiums in the US (directional; verify at publication). It works because carriers can routinely identify which product caused a loss and whether the same issue affects other units. The Decision Loss Channel has no such book because, until now, it has had no such attribution.

This section has outlined a rating structure built entirely on counts and signals the ledger can provide today. The loss experience that would turn those counts into actual rates does not exist yet. When it does arrive, the structure is already in place to receive it and turn it into priced coverage.

> **Box: One claim, resolved twice**
>
> Return to the fractured wrist from the opening scenario, and run the claim both ways.
>
> **Without a ledger.** The adjuster's file contains the injury, the scene, and the operator's own telemetry, produced at the operator's discretion, in the operator's format. Was the intervention reasonable? The operator's data says the fall was real and the force was within design limits; the family's counsel notes that the party saying so is the defendant. Experts are retained on both sides to argue about what the machine probably perceived. The defect-versus-design question — the one that determines whether this is one claim or the first of a fleet — cannot be answered from the available record at all; it can only be litigated. Reserving is set wide because the outcome distribution is wide. The underwriter, reviewing renewal, has learned almost nothing usable: one contested anecdote. Multiply by a portfolio, and the rational carrier response is the one the market currently gives — decline the class.
>
> **With a ledger.** The adjuster pulls the decision entry and verifies its integrity independently — no request to the operator required, no trust in the operator assumed. The entry shows what the machine perceived (fall in progress, confidence stated), which rule governed (the agency's configured intervention threshold), what it weighed (injury risk of force against injury risk of the fall), and what it concluded. The reasonableness question is now an argument about a known decision, not a fight about an unknowable one. The defect-versus-design question is tractable: the governing rule and version are named, and the same decision class can be queried across the fleet's ledgers to see whether the judgment pattern recurs. If it does, that is a version-level defect and a manufacturer conversation; if it does not, it is a single contested judgment and reserves narrow accordingly. The claim may still be paid — the ledger does not decide who wins. What it decides is that the questions are answerable, and answerable questions are what reserving, pricing, and renewal decisions are made of.
>
> The instrument's value to the claims function is not that it exonerates the machine. It is that it converts an unresolvable dispute about what happened into an ordinary dispute about what it means.

8. The Honest Posture

This paper has spent seven sections showing why records that state their own limits are valuable. So it should state its own limits plainly, in three short statements.

What exists. The ledger is built and in operation. It creates an append-only, tamper-evident, longitudinal, reconstructable record of every consequential decision the machine makes. Its integrity can be verified today by mathematics. Anyone can check it. They can do so offline, without contacting or trusting the party that produced the record. This is not a future promise. It is a finished construction fact that anyone can test for themselves.

What does not exist. Two things matter most to an underwriter, and this paper is clear about both. First, there is still no loss experience. The instrument is new, the number of deployments is small, and no claim has yet been priced or paid using its record. The rating structure described in this paper is simply a framework waiting for data. The paper claims only the framework, not the data. Second, full institutional independence is not yet complete. The certifying function sits on separate infrastructure with separate keys and cannot be overruled by the operator. That separation exists today in design and in daily operation. However, the same small organization that built the ledger still runs the certifying function. A reader should therefore treat the attestation as a maturing capability whose direction is firmly committed, not as a finished third-party assurance from an independent body.

The people in the record. A Decision Ledger that works around people will naturally contain information about people. The woman in the opening scenario is both a claimant and a data subject. Any insurer that uses the record therefore takes on privacy obligations. The design posture is simple. The system is built so that third-party personal data can enter only through attributed, erasable paths, and it currently holds none. Data about people who are simply nearby — ambient or bystander data (information the robot might pick up about visitors, family members, or anyone else who is not the main user) — is excluded by default rather than being collected and managed. The governing rule is straightforward: the system must never collect anything it cannot erase. The erasure guarantee is designed to cover every surface where a subject's data can rest. There are no permanent exceptions. Any interim gaps are named and time-bound.

The part that matters most to this audience works as follows. Erasure destroys the personal content (the payload becomes unrecoverable) but leaves the tamper-evident record of the decision itself (its place in the sequence and its integrity). A sophisticated reader will already see the question this answers. Longitudinal loss data and the right to erasure are often presented as a collision. Here the collision is solved by design. The loss history an actuary relies on is never silently punctured when an erasure request is honored. Few, if any, competing telemetry models can make that statement, because the property must be built in from the very first entry. It cannot be added later.

9. The Underwriter Moves First

Most people assume regulation will lead the way on risks like this. They expect a mandate will arrive first, records will be required, and insurance will simply follow the compliance wave. The actual calendar shows something different. And insurance history shows something different too.

Look at the timing right now. In the European Union, strict liability for software and AI systems arrives on schedule. The revised Product Liability Directive must be written into national law by December 2026. At the same time, the main record-keeping rules in the AI Act — the closest thing we have to a required decisional record — were pushed back in 2026 to late 2027 and 2028 (status at drafting; verify at publication). Put the two pieces together and you see the problem clearly. The liability that says "someone must pay" is arriving before the record that lets anyone reconstruct what happened. The gap between those two things is not closing. For the next several years it is actually getting wider. And it is widening fastest in exactly the areas where deployments are growing quickest.

Independent forecasts put the elder-care assistive robot market at about $3.5 billion in 2026 and growing to roughly $7–10 billion by the early 2030s (directional; vendor forecasts; verify at publication). That growth is being driven by a well-documented shortage of caregivers. The machines that exercise judgment around vulnerable people are arriving on the insurance market right now. They are arriving before anyone can yet reconstruct their judgments. And they are arriving ahead of any mandate that would force that record to exist.

Insurance has faced this situation before, and it did not wait for regulators. In 1866 boiler explosions were destroying factories on a regular basis. The Hartford Steam Boiler company started offering both insurance and regular inspections together. That inspection program is a big reason steam power became usable and insurable. Underwriters Laboratories began the same way — insurance carriers needed an independent tester for electrical equipment because they were the ones paying the claims when things caught fire. The Insurance Institute for Highway Safety was funded by insurers and became one of the most effective safety organizations in the world, all without a regulator asking it to exist (institutional histories to be verified at publication). The pattern is simple. When a loss channel lacks the verification it needs, the party that has to pay the losses builds the verifier. The regulator's tool is a mandate that arrives late and sets the minimum. The underwriter's tool is the premium that arrives at the moment of binding and prices the truth.

So what can an underwriter actually do today, before any real loss experience exists? Three things, right now. Ask whether a tamper-evident, append-only, longitudinal, reconstructable record of the machine's consequential decisions actually exists. Ask whether its integrity can be verified without trusting the operator. And ask what the owner's declared Consequence Election is when the machine detects a fault in its own judgment. Those three questions can be asked at submission. They can be made a condition of coverage for the Decision Loss Channel or they can earn a credit against it. A simple pilot needs nothing more than one willing carrier, a small group of instrumented machines, and an agreement that the ledger will be the record of reference for any claim that arises. The appendix at the end of this paper gives a ready-made checklist of ten questions to start with. The first carrier that requires the record will be the first one holding the actual loss data. The manufacturers will follow the premium. They always have.

The central idea of this paper is simple when you run it forward. Autonomy will not be held back by what the machines can do. They can already do the work. It will be held back by whether anyone

is willing to stand behind the machine's judgment with money. The robots that actually get put to work will be exactly the ones someone was willing to insure when they are wrong.

Key Concepts

- **Physical Loss Channel** — Money lost when the machine's body causes harm: collision, wrong force, or mechanical failure. This already has deep actuarial history in machinery, product, and premises liability.
- **Decision Loss Channel** — Money lost when the machine's judgment goes wrong even though its body and sensors work correctly. Almost no actuarial history yet; the problem scales across identical units; uninsurable today only because no instrument measures it.
- **Decision Ledger** — An append-only, tamper-evident, longitudinal, reconstructable record of the machine's consequential decisions and the reasons behind them. The tool that creates usable loss data for the Decision Loss Channel. (It is a simple hash-chained record — not a blockchain or distributed ledger.)
- **Self-Attested Telemetry** — Operational data whose truth and completeness depend entirely on the word of the party that produces it — usually the same party whose risk is being priced. It comes in huge volumes and still cannot prove itself to anyone else.
- **Verifiable Integrity** — The property that lets an underwriter confirm, using simple math, that the record has not been changed — without trusting or even contacting the operator. This exists today.
- **Independent Attestation** — A structurally separate party with no stake in the outcome vouches for the record. This is different from verifiable integrity and is still maturing (see Section 8).
- **Bounded Claim** — An attestation that clearly states, in plain words, exactly what it does *not* claim (such as decision quality, rule wisdom, or overall safety) alongside the narrow facts it does claim. A Bounded Claim can be priced; a certification that claims everything cannot.
- **Consequence Election** — The owner's recorded choice for what the machine does when it detects a fault in its own judgment: either fail closed (stop, go dormant, call for help) or keep operating uncertified at the owner's declared liability. A clear, verifiable signal of the owner's risk appetite that underwriters can actually use in rating.

Appendix: Ten questions to ask before insuring an autonomous machine that exercises judgment

1. Does the machine keep a record of its consequential decisions — not just sensor data, but what it concluded, under which rule, and at what stated confidence?
2. Is that record tamper-evident, and can *we* verify its integrity without trusting or contacting the operator or manufacturer?
3. Is the record append-only — can any party quietly remove or rewrite an entry later?
4. Does the record stay valid even when the machine is offline, or does its integrity depend on a live connection to the party whose risk we are pricing?
5. Who attests to the record's integrity, and is that party structurally separate from the party that builds and operates the machine? Can the operator overrule the attester?
6. What, in plain words, does the attestation *not* claim? (If the answer is "nothing — it means the machine is safe," treat the attestation as unpriceable.)
7. What is the owner's declared consequence election: what has the owner committed the machine to do when it detects a fault in its own judgment — fail closed, or keep operating uncertified at the owner's declared liability?
8. When a loss occurs, can it be tied to the exact decision — and can the same decision type be checked across the fleet to tell a one-off from a version-level defect?
9. Does the machine's conduct history stay with the machine if the owner or operator changes, or does the record stay behind with the seller?
10. If a data subject in the record asks for erasure, what happens to the loss history — is the personal content destroyed while the tamper-evident record of the decision survives, or does the history get silently punctured?

Allan Watkins

Laws of Robots series

Version 1.1 — 13 July 2026