Trust Will Decide the Robot Age — Not Capability

Allan Watkins

Laws of Robots series

Version 1.0 — 11 July 2026

Why trust — not capability — is the gating problem of the robot age, and why it cannot come from the makers.

The hard problem of the robot age is not capability. That problem is being solved, and quickly: machines that perceive their surroundings, decide for themselves, and act physically in the world are already arriving in homes, workplaces, and public space, and they will go on improving whether or not anyone is ready for them. What is not being solved — and what almost nothing now in motion is even attempting to solve — is whether anyone has a defensible reason to trust them. That is the question that will decide how far the robot age actually reaches, and how fast. A capable machine no one trusts does not get through the front door.

And trust starts from a deficit. A robot is not a smarter appliance; it is an autonomous agent with a body, and that combination defeats the ordinary means by which a person decides whether a machine is safe to rely on. The industry building these machines, for its part, has already spent the credibility it would need to vouch for them, so the public's hesitation is not technophobia — it is an earned and rational response. The two problems compound into a single structural fact: the party that must be trusted cannot be the party that certifies it is trustworthy. What a maker asserts about its own machine is exactly the assurance the public has already learned to discount.

That fixes the shape of any real answer. If trust cannot come from the maker, it has to originate somewhere the maker does not control — an independent trust layer, standing between the machine and the people who must live with it. To be worth anything, such a layer has to be beholden to none of the companies whose machines it judges, defined by the owner who answers for the machine rather than imposed from above, verifiable by any outside party on the strength of a tamper-evident record rather than reputation, and built so that it never becomes one more way to seize or disable the machine. These four conditions are not a wish list. They are forced by the structure of the problem — and nothing built or mandated anywhere today meets them.

The case that follows is in four movements: that the robot is a genuinely new kind of thing to trust; that the people who build it cannot be the ones to vouch for it; that nothing now in existence clears the bar this leaves; and that the law arriving to force the question will not, by its own nature, answer it.

A capable machine no one trusts does not get through the front door.

A robot is not a smarter appliance

A robot is not a smarter appliance. It is an autonomous agent with a body — a machine that perceives its surroundings, decides on its own what to do, and then acts physically in the world — and that single fact dismantles nearly every method we have for deciding whether a machine deserves our trust. The methods we lean on were built for tools: read what it does, confirm it was made to a standard, switch it off if it misbehaves. None of them quite works on a thing that acts on its own initiative, in private space, under divided control. Four features of the robot, taken together, put it in a category of its own.

The first is that capability can be verified but conduct cannot. You can establish what a robot is able to do — its specifications, its safety envelope, the tasks it was built and certified for. But what a machine is capable of tells you nothing about whether it should have done the particular thing it did at a particular moment, and after the fact there is rarely a record that settles the question in a way anyone can trust. Suppose a household robot lets in a visitor the owner would have turned away, or quietly clears away papers it took for clutter: afterward the owner has only the machine's own logs, kept and shaped by the company that made it, and no independent way to establish what it did, or why. A human agent can be asked to explain; a contract can be read against its terms; a robot offers a black box whose account of itself, where it keeps one at all, is authored and held by the company that built it. The distance between verified capability and warranted action is new, because nothing else a person owns acts in the world on its own judgment.

The second is that ownership and control have come apart. The person who buys the robot is told they own it, but the maker keeps a hand on it — the ability to reach in from a distance and change what the machine does. Updates pushed overnight can alter its behavior without notice or consent; features can be gated, throttled, or withdrawn; the machine can be made to report on its own use, tuned toward the maker's commercial interests rather than the owner's, or shut off outright. Most of it happens out of sight, and none of it with the owner's permission. So the nominal owner is, in the ways that matter most, not the one in command — and a machine a person does not truly control is a machine that person cannot truly vouch for.

The third is that the robot is an unprecedented sensor in the most private settings there are. It carries cameras, microphones, and the means to map a space and track who moves through it; it is placed in homes and workplaces; and unlike the fixed devices that came before it, it moves through those spaces and stays in them, building a continuous picture of private life. By default the owner neither owns nor controls what it captures, where that information travels, or how long it survives. This is a difference in kind, not degree, from a phone left on a table or a speaker in the corner.

The fourth is that a compromised robot is categorically worse than a compromised computer. Any connected machine can be taken over; the robot is the first ordinary machine where being taken over means physical action. A hijacked laptop leaks information. A hijacked robot can move, handle objects, block a doorway, or be turned against the people beside it. And the owner's recourse is thin: a compromise is hard to detect, control can be difficult to recover when the very channel meant to secure the machine is the one that has been turned, and a physical act, once taken, cannot be recalled.

These are not four separate grievances. They share a root. A robot is an autonomous agent with a body, owned by one party and steerable by another, sensing everything around it and, by default, accountable to no one. Every instinct a person has for trusting a machine was formed in a world of tools — and a tool is the one thing this is not. The difficulty is not that today's robots are badly made; it is structural to the kind of thing a robot is, and it would persist even if every machine shipped flawless.

The distrust is earned

The structural problem would be serious enough on its own. But the robot does not arrive into a neutral world, ready to be judged on its own merits. It arrives into a deficit it did not create, carrying the reputation of the industry that built it — and that reputation is, by now, an accurate liability.

For two decades the public has been taught, by direct experience, what to expect from connected consumer technology. Data is collected by default and monetized out of sight. Consent is buried in terms no one is meant to read. Products are changed after purchase, features removed, devices left to stop working on a schedule that suits the seller. The buyer's interest is routinely placed behind the platform's. People did not invent this impression; they learned it, over and over, from many of the same companies now proposing to place an autonomous machine in their living rooms. The lesson took.

The robot inherits all of it, and raises the stakes. It is, in the main, the same makers and the same business models — now with a body, in the home, sensing continuously and acting on its own. So the hesitation that meets it is not a fresh prejudice to be educated away, and it is not technophobia. It is a reasonable forecast: that patterns already demonstrated in settings where they merely annoyed will repeat in a setting where they will matter a great deal more. Read against the record, caution is the rational response, not a lapse of it.

This is why the friction is the largest barrier to adoption, and why it is the one barrier no robot can clear by being better. Capability does not speak to it. A more capable machine from a source the public has already learned to discount is not more reassuring; it is more concerning, because it can do more on behalf of whoever ends up controlling it. A trust deficit cannot be out-engineered, and it cannot be out-promised either, because the promise issues from the same discounted source as the product. Whatever is going to restore the missing trust will have to come from somewhere other than the makers asking for it.

The party that must be trusted cannot be the party that certifies it is trustworthy.

The party that must be trusted cannot vouch for itself

Two facts established so far converge on a third that decides everything. The robot cannot be trusted on its own terms — its capability can be verified, but its conduct cannot, and it leaves behind no record that settles the matter. And the industry that builds it has already spent the credibility it would need to vouch for it, so its assurances are discounted before they are heard. Put those together and the question is no longer how to engineer a better robot or find a more sincere maker. It is where the warrant could come from at all — and the answer is fixed by a constraint that owes nothing to technology: the party that must be trusted cannot be the party that certifies it is trustworthy.

This is not an accusation against any particular maker. It is a structural impossibility. A manufacturer's assurance about its own machine is precisely the assurance the public has already learned to discount, because the manufacturer is the interested party — it gains from being believed, it holds the evidence, and it writes its own report card. Today that is the entire system: every maker grades its own homework, and there is no independent body whose business is to grade it instead. Trust asserted by the party who benefits from it is not proof; it is marketing. And no measure of sincerity, engineering rigor, or good intention repairs the defect, because the defect is not dishonesty — it is that self-certification carries no information. A verdict cannot tell the careful maker from the careless one when both are the sole authors of their own report.

So trust has to originate somewhere the maker does not control. That single requirement is not a preference to be weighed against others; it forces the form the answer must take, and it forces it everywhere — independent of any market, any company, and any government's law, because the defect it corrects is present under every regime, the permissive and the strict alike. From it, four conditions follow. They are not a feature list assembled for completeness. Each one closes a specific way that self-certification fails, and an answer that meets three of the four does not come up a quarter short — it fails, because the opening it leaves is the whole of the problem in miniature.

- **Independent.** The standard cannot be owned by, or answerable to, the companies whose machines it judges. The moment the judged party can hire, fund, overrule, or walk away from the judge, the verdict carries no more weight than the maker's own word — which is where the problem began. Independence is not a matter of branding. It is a matter of who can be told no — and made to live with it.
- **Owner-defined.** The limits a robot must live within are set by the responsible person who answers for it — not by the maker, and not by the standard itself. That authority runs in one direction: the owner can make the rules stricter, but no one — not the owner, not the maker — can lower the safety floor beneath them. Above that floor the system enforces the owner's rules; it does not arrive with a worldview of its own and impose it. A trust layer

that substitutes its own judgment for the owner's has not solved the accountability problem. It has merely changed the name on the authority.

- **Verifiable.** What the layer certifies must be a precise, bounded claim — that the machine's conduct was faithful to the owner's rules, not that those rules were wise or the machine otherwise sound. A certification that vouches for everything vouches for nothing. And every consequential action must leave a tamper-evident record that any third party can examine, so that what it certifies rests on proof rather than reputation. A claim that cannot be checked after the fact is a reputation, and reputation is the currency that has already failed. The record has to be the kind a skeptic can audit against the maker's interests — not the kind its keeper can quietly shape.
- **Never safety-disabling.** The layer constrains and records what a robot does; it never becomes one more way to seize or switch off the machine. This is the condition that separates a trust layer from yet another control surface, and it is not negotiable. A mechanism that can disable the robot reintroduces, under the banner of safety, the very remote-control and attack-surface dangers that made the machine untrustworthy in the first place. A solution that creates a new way to commandeer the robot has joined the problem.

These four are the bar. Stated plainly they can read as modest, even self-evident — but the test of a bar was never whether it is easy to describe. It is whether anything clears it. The question that follows is therefore the practical one: of everything built or mandated in the world today, does any of it meet all four at once?

Why nothing that exists today meets the bar

The field is not empty. There are robot-safety standards with decades of engineering behind them, a new generation of AI-governance standards, a fast-growing market of software built to operationalize them, and a body of European law that is turning governance from aspiration into legal obligation. A reader could be forgiven for assuming the problem is already being addressed somewhere in all that activity.

It is not — and the reason is not that the work is immature, waiting only for another revision cycle to arrive. It is that everything in the field is one of three kinds of thing, and not one of the three is, at the same time, independent of the maker, defined by the owner, and verifiable by an outside party. Measured against those tests the gap is not a soft spot to be hardened. It is structural, and more of the same will not converge on it.

The safety standards answer a different question. The flagship standard for industrial robots was overhauled in 2025 to sharpen its functional-safety and cybersecurity requirements, and it remains exactly what it has always been: an account of whether a machine will crush, strike, or pinch the person beside it. It says so itself — it explicitly excludes service robots and consumer machines from its scope, which is to say it does not reach the robots whose arrival into private space is the entire concern here. The standard that does reach them, the one written for personal-

care robots, is built to the same purpose: mechanical hazards, electrical safety, collision, contact. This work is necessary and it is not in dispute. But "will it injure me by accident" is a different question from "should it have done what it did, on whose authority, and is there a record anyone can trust." A robot can satisfy every functional-safety requirement ever written and still act outside its owner's wishes, capture what it was never meant to capture, and leave behind nothing that can be checked afterward. Safety is the floor. It was never the thing in question.

The governance standards — and the market built on them — certify the maker, not the act. The first international management-system standard for artificial intelligence does not certify products or machines at all; what it certifies, when an organization is certified against it, is that the organization runs a process — that it keeps policies, maintains an inventory of its systems, performs its own risk assessments, and produces documentation an auditor can review. The widely cited American risk framework is softer still: voluntary guidance with no certification behind it. Around these standards a commercial market has formed, and it is growing quickly, but its center of gravity is the same. These are governance platforms sold to the enterprise that builds or deploys the system; they discover models, register them, score their risk, map them to regulations, and generate audit-ready evidence. The most established of them, by its own market's account, is built around policy and documentation and does not sit in the execution path of the machine's decisions at all. Strip the vocabulary away and what remains is governance of the maker, by the maker, now with better software. It is the most sophisticated possible form of a maker grading its own homework. None of it is operated by a party the maker does not control. None of it answers to the owner. And almost none of it touches the individual action at the moment it is taken — which is the only moment at which trust could actually be earned.

The new law makes governance mandatory and still leaves the trust to the maker. This is the sharpest evidence, because it is the regime built expressly to force the issue. It is also the self-certification trap of the previous section made literal — the same structural defect, now written into statute. Under Europe's AI regulation, the great majority of high-risk systems are cleared for market by internal self-assessment: the maker evaluates its own system against the requirements, signs a declaration of conformity, and affixes the conformity mark itself, with no outside body involved. An independent assessor is required only for narrow categories, and even then what the assessor reviews is the maker's quality-management process and technical file — the paperwork around the machine, not the machine's conduct in the field. The regulation requires logging — but the log is kept by the party being judged, and a record its keeper can shape is not a record an outsider can trust. The companion liability reform goes further, treating software and AI systems as products that carry no-fault liability for the harm they cause — and yet it operates only after the harm, in a courtroom, and it had to shift the burden of proof onto the defendant precisely because the victim cannot otherwise establish what the machine did or why. That burden-shift is the law building a workaround for a record that does not exist. It is not the record.

Set the three side by side and the common thread is plain. Every mechanism now in the field is pointed at the maker or the regulator, never at the owner. Every one of them works at the level of process, or paperwork, or after-the-fact liability, never at the level of the single consequential act.

And in every one of them the trust, where any is offered at all, is ultimately asserted by the party that stands to gain from being trusted. Against the four conditions any real answer must satisfy, three are simply unmet — nothing here is independent of the maker, nothing is defined by the owner, nothing is verifiable by an outside party on the strength of proof rather than reputation. The fourth condition, that such a layer must never become one more means of seizing or disabling the machine, is not a test these approaches fail so much as a line the answer must be built never to cross; that none of them even operates at the level where the temptation arises is itself a measure of how far they sit from the problem. The work that is needed is not a better version of any of these. It is a different kind of thing, and it does not yet exist.

The law will force the question — and cannot answer it

The era in which governing a robot was optional is ending. Across the world's major markets, the expectation that a maker ought to be able to account for its machine is hardening into a legal requirement that it must — and that is the reason this is a problem for now and not for some later decade. The strictest of these regimes is already on the clock, with its core obligations and a companion liability reform falling due across the next few years. The direction of travel is unmistakable: accountability for what an autonomous machine does is moving from courtesy to duty.

Two things, however, keep that duty from producing trust. The first was the subject of the previous section: even the most serious regime clears its machines for market by having the maker certify itself, and its requirements are not yet concrete. The deadlines are slipping precisely because the technical standards that would say what compliance actually requires do not yet exist. The obligation is arriving ahead of anything able to meet it — which means that for a meaningful stretch, "compliant" will describe a machine that has been measured against a yardstick still being drafted, by the party holding the ruler.

The second is wider, and it is the one that decides the matter. The regimes do not agree, and they are not converging. The strictest mandate resolves, on inspection, to self-assessment. The largest market is deliberately declining to regulate at all, and is moving to stop its own subdivisions from doing so in its place. A third major power regulates heavily but aims its rules at the content its machines produce, not the conduct of the machine itself — a regime built around information order rather than the owner's authority. Elsewhere one major economy has chosen non-binding promotion with no penalties attached, while another has just brought its own risk-based framework into force on its own schedule. Five jurisdictions, five different questions, and no shared answer in sight. Even where two regimes appear to overlap — the labeling of machine-generated content, say — the requirements differ enough that a maker must satisfy each separately; there is no mutual recognition, only a thickening web of local rules.

From this the conclusion is forced. A robot cannot inherit its trust from law. Where a regime exists, compliance certifies conformity to one jurisdiction's rules as that jurisdiction's own institutions interpret them; it does not establish that the machine did what its owner required,

and it stops at that jurisdiction's border. Where no regime exists, the gap is not narrowed at all — it is simply open. And the machine itself crosses borders that the rulebooks do not, carried into homes and workplaces under whatever regime happens to govern the ground it stands on. Trust sourced from a government is local by construction: it answers that government's question, not the owner's, and it cannot travel with the thing it is supposed to vouch for. The only trust that travels with the robot is trust that was never drawn from any government in the first place — independent of all of them, defined by the person who answers for the machine, and verifiable by anyone who cares to check. That is not a way around the law. It is the layer the law, by its nature, cannot supply for itself, and it is needed under the permissive regimes and the strict ones alike.

None of this is the whole of the robot question, and it is worth saying so plainly. It says nothing about the jobs these machines will displace, the resources they will consume, or the genuinely different things that different societies will want from them — those are real questions, and they belong to other arguments than this one. The claim here is narrower, and it sits beneath all of them: before any of that can be debated well, there has to be a way to know what a given machine actually did, on whose authority it did it, and whether an outsider can confirm the account. No law now in force or on the horizon delivers that, in any market, for the machines that matter. That is the gap. It is not closing on its own, and the clock that is forcing the question is not the same as an answer.

The layer the age cannot build for itself

Put the pieces together and the conclusion is not rhetorical; it is structural. Autonomous machines are arriving under their own momentum. The reasons not to trust them are sound, and they are compounded by an industry that has already spent its credibility. The one party positioned to offer reassurance is the one party whose reassurance counts for nothing. And the standards, the governance tooling, and the new law all measure something other than the thing that matters — or measure it from inside the very interest that needs checking.

What remains is a single, durable requirement: a trust layer the robot age cannot build for itself — independent of the makers, answerable to the owner, verifiable by anyone, and incapable of being turned into a means of control. It is the piece on which adoption ultimately depends, and it will not arrive by accident, nor as a byproduct of better machines or stricter rules.

The machines are coming. Trust is not coming with them. It will not build itself; it has to be built, deliberately, by someone with no stake in the verdict.

The work of building it has begun.

Allan Watkins

Laws of Robots series

Version 1.0 — 11 July 2026

This document is part of the Laws of Robots governance whitepaper series. The version of record is maintained under version control. All commitments are owner-authored and subject to the published change discipline.