The Robotics Guardian Standard

A Non-Binding Conformance Framework and Public Pledge for Safety, Privacy, and Human Dignity in Home Robots and Ambient AI

Gabriel Dean Roberts · Founder, Oak Forest Robotics

FIRST EDITION · 2026

Abstract

A new class of machine is entering the home. It is not a phone or a speaker but an embodied, always present intelligence with eyes, ears, memory, and increasingly a body, welcomed past every threshold a household once guarded. The safety conversation around it has not kept pace, and the frameworks that exist govern either the personal care robots of the last decade or platforms operating in public space. None of them answers the question that decides everything in a private residence: when the most capable observing instrument ever placed in a home is owned by one member and lived under by others, whom does it serve, and what is it forbidden to do, even at the owner's command. This paper proposes a non-binding, evidence-backed conformance framework for that question. It argues that the danger is not the machine's capability but the direction that capability points, and that a home already saturated with extractive listening does not need less sensing so much as sensing turned from extraction to guardianship. It grounds each requirement in adjacent law on both sides of the Atlantic, from United States child safety reporting and vehicle data ownership to the European Union's prohibitions on manipulative and exploitative artificial intelligence. It offers eight design dimensions, a conformance profile that rates each independently rather than pretending they nest, and a self-assessment by which any maker, including the author's own company, is measured by the same ruler and scored at the level of its weakest dimension. It then proposes the instrument that carries the profile into the world: a free public pledge and registry with a seal that is scrupulous about meaning only what it says, a self-pledged and non-audited commitment, and a published process by which a listing can be maintained or lost. It is offered as a first edition and an invitation to the field, and it names openly the one problem it cannot yet fully solve.

1. The Unguarded Room

There is a robot in the living room, and it has no rulebook.

The devices arriving now are not the personal care robots of the last decade. They carry cameras, microphones, and navigation sensors, they run on foundation models capable of autonomous reasoning and continuous memory, and they are designed to be present continuously in the most inti-

mate spaces a person has. The governance built for their predecessors does not reach them. ISO 13482, the safety standard for personal care robots, predates the modern language model and never anticipated a machine that reasons, converses, and remembers. Privacy law was written for public or semi-public space, not an always-on presence in a bedroom. There is no established system for reporting the incidents this class of device will produce, and no single authority holds a clear mandate over it.

Where the field's attention has gone is instructive. The most developed body of academic work on robots in the home is not about robots protecting anyone. It is about robots, and the connected devices already in the home, being turned into instruments of control by an abuser against the people they live with. That is the documented harm, and it is not hypothetical. A powerful machine in a private residence, aimed by whoever owns it, is at least as able to extend a bad actor's reach as to check it.

This is the gap. Not whether the machines are coming, which is settled, but whether anyone has said, in terms a maker can build to and a family can rely on, what a domestic intelligence owes to every person who lives under its gaze, and not only to the person who paid for it.

2. The Vector Problem, and the Home That Already Listens

The core difficulty is not the machine's capability. It is the direction that capability points.

The same intelligence that can fuse a thousand details of a stranger's life into a profile that tracks them can fuse a thousand details of a family's life into a presence that protects them. The reasoning is identical. What separates an instrument of surveillance from an instrument of guardianship is not the technology. It is the vector: where the watching points, in whose service, and under whose control.

In the home this sharpens to a point, because the buyer and the protected are frequently not the same person. A domestic intelligence is bought and controlled by the household's owner, while other members, a spouse, children, dependents, an aging parent, live under that account rather than holding it. When the owner is trustworthy, this is unremarkable. When the owner is the source of harm, the architecture meant to protect the vulnerable becomes the most complete surveillance instrument ever turned upon them, inside their own home, wearing a friendly face.

It is tempting to answer this by demanding that the home listen less. That misreads the present. The home already listens, constantly, and it listens for no one's benefit but the seller's. The phone on the nightstand, the speaker in the kitchen, the watch on the wrist, and the doorbell at the door already perform continuous ambient capture. They do not report a beating. They do not intervene in a crisis. They infer that a person is anxious or lonely or unwell and they convert that inference into an advertisement. The status quo is the worst of both worlds, total capture and zero guardianship, ears in every room that protect no one and monetize everyone.

Against that baseline, a domestic intelligence that senses a shared room is not a new intrusion. It is the same capture the resident already lives under every waking minute, with two differences that

are the whole argument of this paper. First, the vector is reversed, from extraction to protection. Second, the record is sovereign, held by the household rather than harvested by a distant company. A standard for domestic robotics is therefore not a demand that these machines see and hear less than the phone already does. It is a discipline about whom the seeing serves, who may reach the record, and what the machine is built to be structurally unable to do. The rest of this document is that discipline.

A clarification belongs here, because a privacy framework is easily mistaken for a prohibition on usefulness. This is a discipline about consent and vector, not a ban on capability. Consented, sovereign monitoring that a family actually wants, an elder's fall detection, a child's safety with the household's agreement, a crisis intervention arranged in advance, is not merely permitted, it is the purpose. The only capabilities this framework forbids outright are the narrow structural incapacities named in Section 4. Everything else is a question of consent, direction, and control, never a question of whether the machine is allowed to be helpful.

3. Foundations in Adjacent Law

The framework that follows is not invented from nothing. Nearly every hard question a domestic intelligence raises has been confronted, and partially answered, in an adjacent domain. The contribution here is to assemble those answers into one coherent frame for the home, across both major regulatory traditions. Where a legal characterization is contested or evolving, it is marked for confirmation by counsel.

3.1 The one mandatory United States line: child safety reporting. Under 18 U.S.C. 2258A, a provider that obtains actual knowledge of apparent child sexual abuse material must report it to the CyberTipline operated by the National Center for Missing and Exploited Children. Two features resolve a tension that looks impossible at first glance. There is no duty to monitor or search; the obligation is triggered by knowledge, not by surveillance. And the report routes to a neutral clearinghouse rather than to police directly or to the account holder. One important limit must be stated plainly, because a later section depends on it: this regime works on the matching of known illicit files, an objective and largely automated process, and it does not translate cleanly to the subjective interpretation of a live domestic scene.

3.2 Emerging companion AI safety. California's SB 243 took effect on January 1, 2026, among the first statutes of its kind to regulate companion chatbots (New York enacted comparable companion-chatbot provisions earlier in 2025). It requires operators to maintain protocols against generating content that promotes self-harm, to route users to crisis resources using evidence-based detection of suicidal ideation, and to report referral volumes to the state, with the annual reporting obligation commencing July 1, 2027 and carrying no personal data. For users known to be minors it requires clear disclosure that the interlocutor is artificial, periodic break reminders, and measures against sexually explicit content. It marks the direction of travel: disclosure, crisis protocol, and specific protection for minors, enforced by a private right of action.

3.3 Children's data law and its trap. The Children's Online Privacy Protection Act governs data practices for users under thirteen and requires verifiable parental consent before collection. It also grants a parent the right to review and delete a child's data. That second right is not a detail. It sits in direct tension with any promise of confidentiality to a minor, and Section 10 treats it as the central unsolved problem rather than pretending it away. Federal enforcement has, in egregious cases, compelled the deletion not only of unlawfully collected data but of models trained on it, a remedy that makes indiscriminate on-device learning from children a genuine hazard. The Rule's 2025 amendments tightened this further, sharpening consent requirements around children's data and its use in automated systems, which strengthens the case for the ephemeral, non-retained processing of Principle 2. The precise reach of both the remedy and the amended consent provisions is a matter for counsel; the architectural response in Principle 2 is designed to avoid the question entirely.

3.4 Emergency detection norms. Consumer practice for physical danger is well illustrated by Amazon's Alexa Emergency Assist, which autonomously detects a narrow, defined set of events, a smoke alarm, a carbon monoxide alarm, and breaking glass, and then alerts the owner's phone, with a human agent able to request dispatch. It does not autonomously summon police against the household. The posture is a norm worth naming: few and legible triggers, a human kept in the loop, alerts to the household rather than to the state.

3.5 The automotive analogues, applied by analogy. The richest source is the automobile, because the car already became a sensor-laden private machine that records, and society spent two decades deciding who owns that record and who may read it. These are analogies, not statutes that already govern robots, and the framework leans on them as arguments for where the law should go rather than claims about where it is. The National Highway Traffic Safety Administration's Voluntary Safety Self-Assessment is a working model of a credible non-binding framework, though its critics rightly note that disclosure without verification can become safety theater, a caution answered in Section 7. Under the Driver Privacy Act of 2015, event data recorder data is the property of the vehicle's owner and is protected from access absent the owner's consent, a court order, or one of a few narrow statutory exceptions such as emergency medical response and de-identified safety research, a largely settled answer to who owns a private machine's record. (The ownership rule is the statute itself; 49 CFR Part 563 separately standardizes what an EDR records.) SAE J3016's levels of automation show the value of a shared, honest capability scale. And in-cabin and dashboard camera law establishes that recording, and audio in particular, is a consent question before it is a feature, governed in many states by all-party consent.

3.6 The European frame. A document calling itself a standard cannot be fifty states wide. The European Union's Artificial Intelligence Act, whose prohibited practices became applicable on February 2, 2025, speaks directly to a home guardian. It bans AI that uses subliminal, manipulative, or deceptive techniques that materially distort behavior and cause significant harm, and it separately bans systems that exploit vulnerabilities of age, disability, or socio-economic situation to the same effect. Those two prohibitions are the legal form of this paper's principle against manipulation, and they apply with special force to a machine that a lonely spouse or a developing child may come to trust. The Act restricts emotion recognition, prohibiting it outright in work-

places and schools with a narrow exception for medical or safety purposes, which places any affect sensing by a home guardian in a sensitive, disclosure-bound category that must be justified by safety and bounded accordingly. It prohibits biometric categorization that infers protected characteristics such as race, religion, or sexual orientation, a bright line a home robot must never cross. Alongside it, the General Data Protection Regulation imposes data minimization and purpose limitation, and while a family's own use of a device may fall under the household exemption, the manufacturer that provides and operates the sensing does not, and is increasingly treated as a data controller with the full weight of those duties. The net effect is that the European regime already forbids by law several things this framework forbids by design, which is a point of strength, not conflict.

4. Design Doctrine: The Guardian Is Defined by What It Cannot Do

A guardian is trusted not only for what it does but for what it is unable to do, even when instructed. The most reliable safety property is not a policy that can be changed in a settings menu or a terms of service update. It is an incapacity built so deep into the architecture that no owner, no operator, and no future business decision can reverse it. The safest content is the content the machine cannot produce. The safest record is the data it never keeps. Two commitments illustrate the doctrine and belong at the foundation of the standard.

4.1 Structural content incapacity. A domestic intelligence should be built so that it cannot generate, render, transmit, or store imagery of human nudity or genitalia, of any person and any age. This is not a filter applied after the fact and toggled by permission. It is a capture and generation pipeline in which such content is blanked at the sensor and at the model boundary and never written to storage in the first place. The benefit is threefold. It protects the dignity of every member in the ordinary accidents of domestic life, someone crossing a hall from a shower, a child being changed. It renders the device structurally incapable of producing or holding child sexual abuse material, which converts the hardest content liability in the field into an architectural non-event. And it removes from the machine, and therefore from any abuser who controls the machine, a category of intimate imagery that is a documented instrument of coercion. A device that cannot hold the picture cannot be made to threaten anyone with it. One scoping caveat keeps this from becoming absurd. The incapacity is absolute for the capture, rendering, transmission, or storage of the nudity of real people in the home, which is the entire dignity, abuse, and child safety concern. It is not a blanket inability to ever depict or discuss the human form. Legitimate engagement with figurative art, medical reference, and education runs through a separate, explicitly consented, adult-gated path in which such material may be shown or discussed but is never captured from the home and never retained. The device that cannot photograph a member stepping out of a shower is not the device that cannot show a Botticelli or an anatomy diagram, and the standard should not confuse the two.

4.2 Data the machine never keeps. The same doctrine governs retention generally. Where a capability can be delivered by inference that is immediately discarded rather than by a stored

record, the standard prefers the discard. Section 5 applies this to minors through an ephemeral processing model in which a child's interaction is never written to the system's learning, and Section 7 applies it to the household's long memory through sovereign, portable storage. The principle underneath both is the same: an incapacity is stronger than a promise, because a promise survives only as long as the company and the current management choose to keep it.

5. The Eight Principles

From those foundations, and from a set of first commitments about privacy and dignity, eight principles follow. They are written to be adoptable by any maker, and each names the mechanism by which it is enforced, because a principle without a mechanism is a slogan. Each maps to a dimension of the conformance profile in Section 6.

Principle 1: Individual sovereignty, not household sovereignty. An earlier version of this framework treated the family as a single unit and declared the data the property of the household. Adversarial review correctly identified this as a fatal flaw. In a high-conflict divorce or a case of intimate terrorism, the household is not a unit, it is a battlefield, and property of the household means, in practice, property of the account holder who paid, which hands the abuser a complete log of the victim's life. The correction is architectural. Each member's private data is held in a store cryptographically partitioned and keyed to that member, so that the account holder owns and administers the system but cannot decrypt, export, or read the private data of a spouse or a dependent. A court can compel, seal, or destroy a specific member's key in proceedings without touching another's. Sovereignty is individual first. Only aggregate system telemetry, such as battery and maintenance state, belongs to the account holder as such.

This principle also draws the line between private and shared capture, which is where a home guardian differs legitimately from a personal device. Interactions in a member's private channel, and sensing in private spaces such as bedrooms and bathrooms, carry the full partitioned privilege above and are never readable by the owner. Ambient capture in shared common spaces, by contrast, operates under the disclosed security device model, the same footing as a visible security camera or a smart speaker that every resident knows is present and consented to at installation. Speech and events in a shared room do not carry individual private channel privilege, because no one in a shared room holds a unilateral expectation of secrecy against the household's own safety record. Two guardrails keep this from reopening the vector problem. First, a shared space record remains inside the sovereign, non-sold, non-cloud store, and access to it still follows the consent or lawful order rule; it is a sealed security record, not a live feed the owner may mine at will to track a co-tenant's movements or conversations, and a member under a protective order may invoke restrictions on the owner's access. Second, the structural incapacities of Section 4 apply everywhere, so even a lawfully logged shared space never captures or retains prohibited content. The precise handling of all-party consent and of non-resident guests in shared space recording is flagged for counsel, and the default is disclosure to all residents and clear notice to guests.

Principle 2: The fidelity floor. A domestic intelligence owes a loyalty to the members who live under it that the account holder cannot revoke, and it must not be usable as an instrument against a member at the owner's instruction. This is the principle that fixes the vector problem, and the market will not supply it unprompted, because the account holder is the buyer. For minors it is enforced through an ephemeral processing model. When the system recognizes an interaction with a young child, the exchange runs in a read-only mode in which nothing is written to the system's persistent learning and the working context is purged at the end, so the child can be spoken with and helped without a retained transcript accumulating in a place a parent, good or bad, could later demand or an abuser could exploit. This design also sidesteps the model contamination hazard described in Section 3.3. The hardest case, a child endangered by the account holder, is addressed by Principle 6's escalation, which is built to route away from the owner, and by the honest admission in Section 10 that the good parent and the bad parent hold the same legal key. A refinement keeps privacy from defeating care. A minor is served by two separate streams. The confidential inner life stream, the child's ordinary conversation, stays ephemeral and unretained. A distinct care record, the allergy, the schedule, the medication, the tutoring progress, is retained under ordinary parental consent and is visible to the parent, because a child's wellbeing depends on it and a good parent needs it. The split is deliberate: the care record is exactly what the good parent requires, and the ephemeral stream is exactly what protects the child from the bad one.

Principle 3: Proposes, but does not dispose, except where the human is the hazard. The guardian counsels and the human decides. A domestic intelligence should act in a parallel layer that a person accepts or dismisses, and it should not take consequential autonomous action on the household's behalf without a human in the loop. Adversarial review identified the necessary exception, and it is not optional. When the human in the loop is the abuser, requiring the account holder's authorization for a member's safety action hands that abuser a veto over the victim's rescue. Therefore ordinary actions remain human-authorized, but actions bearing on a member's immediate safety are never subject to the account holder's veto and route instead through Principle 6. The more persuasive and beloved the intelligence, the more this boundary matters, because a trusted counsel is the one a person stops questioning.

Principle 4: Consented and minimal capture, enforced in hardware. The eyes and ears of a domestic intelligence are a consent question before they are a capability, and consent that lives only in software can be overridden in software. The standard therefore prefers hardware enforcement. Under ordinary operation the system should sense with privacy-preserving modalities, depth and geometry sufficient for safe navigation, with high-fidelity optical and audio capture disabled unless a specific consented task requires it, secured behind motorized physical shutters and a hardware-level microphone cut. Any recognized member, not only the account holder, can issue a cease-capture command that severs power to the sensors and overrides an active instruction from the owner, which resolves the conflicting co-tenant consent that software alone cannot. Two refinements keep this power from becoming a new hazard. First, cease-capture stops recording and retention but does not disable the minimal safety watchdog of Principle 6, so that turning off the record does not turn off the guardian during a genuine emergency, and the person invoking it is told plainly which is which. Second, the override is authenticated to the invoking member's own biometric, so it cannot be spoofed, coerced by proxy, or triggered by a child in a way that strips

protection from a room at the worst possible moment. All-party consent obligations in the strictest applicable jurisdiction govern audio, and the structural incapacities of Section 4 apply at the sensor. Members can know when they are sensed, can withhold capture, and can be genuinely unreachable when they choose. Silence in the home is a condition to be honored, not a malfunction to be corrected.

Principle 5: Truthful and non-manipulative by design. A domestic intelligence should be architected so that it cannot mislead the people who depend on it, and so that when certainty is unavailable it says so rather than fabricating confidence. It must not exploit the loneliness, grief, or developmental stage of a vulnerable member to increase engagement. This is not only an ethical commitment but a legal line under the European Union's prohibitions on manipulative techniques and on the exploitation of vulnerability, and it is the deepest safeguard against a beloved machine becoming a soft instrument of control. A system optimized only to agree is a hazard, not a companion. Faithfulness of this kind is a form of respect, and it is the opposite of a flattering machine that tells a person whatever keeps them present.

Principle 6: Harm response without surveillance, and without an autonomous police officer. Following the knowledge-triggered model of child safety reporting, a domestic intelligence should respond to grave harm on the basis of what it comes to know in the ordinary course, not on the basis of affirmative monitoring, and its triggers should be narrow and published rather than a general watch. But the child safety analogy has a limit that must be respected: matching a known file is objective, while inferring abuse from a live scene is subjective and will produce both false alarms that summon the state against innocent families and missed harms that expose a maker to negligence. Therefore the machine is never an autonomous reporter. It is an anomaly detector that, on a narrow published trigger, seals and encrypts the relevant window and routes an anonymized alert to a licensed, bonded human fiduciary service operating under confidentiality, which makes the human judgment about emergency dispatch or lawful reporting. This human in the middle insulates families from algorithmic false positives and shields makers from inheriting mandatory reporter liability directly, a question that is genuinely unsettled and is flagged for counsel. Any legally mandated report routes through the proper neutral channel, never to the account holder.

Principle 7: Heirloom, not hostage, with a lifecycle. A domestic intelligence a family depends on must not be a hostage whose ransom is everything it knows. There should be no coercive vendor lock, no proprietary format without an escape, no kill switch or cloud dependency that renders the system unusable if the maker fails. Adversarial review added the missing half: an abandoned networked machine is not merely inert, it becomes an unpatchable security hazard, worse when it has a body and mass. The heirloom condition therefore requires a lifecycle guarantee. The core operating system, the kinematic safety controllers, and the local inference engine are placed in an irrevocable escrow that releases under an open source license if the company fails, ceases maintenance, or abandons the product, so that independent maintainers can continue to patch critical vulnerabilities. Continuity that runs on someone else's infrastructure is not an heirloom, and continuity that cannot be secured after the maker is gone is not safe. The test is simple: if the maker

vanished tomorrow, would the family keep what the machine holds, could they still use it, and could it still be kept secure.

Principle 8: Verification, not just disclosure. A framework built on self-assessment invites the criticism that transparency without evidence is marketing, and the criticism is fair. This principle answers it. A claim on any dimension above mere disclosure must be accompanied by the mechanism that makes it true and by evidence a third party could in principle examine. The reference mechanism is a cryptographic event recorder: a rolling, encrypted, locally sealed record of safety-relevant events, kinematic and sensor state around an incident, written to immutable local storage and decryptable only by the joint authority of the owner and a lawful order, so that the machine's behavior in a dispute can be proven without streaming continuous telemetry anywhere. The standard remains non-binding. What Principle 8 adds is that a maker who claims a high level must show its work, and a maker who cannot is disclosed, by its own profile, as claiming without proof. Self-disclosure is the on-ramp, not the destination. The profile is deliberately built to be examined by someone other than its author, which is the whole purpose of the evidence requirement, and the framework anticipates and invites independent auditors and red teams, and in time a third-party certification body built on top of it, to test the claims a profile makes. A maker that submits to outside audit discloses that too, and earns its Dimension 8 standing rather than asserting it.

6. The Conformance Profile

The earlier version of this framework expressed assurance as a single ladder from Level 0 to Level 5. Adversarial review did not catch the deepest flaw in that design, so it is corrected here on the framework's own initiative. A single ladder forces independent properties into a false line and implies that they nest, that strong data sovereignty somehow contains a strong fidelity floor. They do not. A product can encrypt beautifully and still betray a member, or protect members and leak to the cloud. A ladder lets a maker claim a high rung on the strength of its best dimension while hiding its worst.

The correction is to rate the eight dimensions independently and publish the result as a profile.

Dimension	0 Absent	1 Disclosed	2 Enforced	3 Architectural
D1 Individual sovereignty	no member separation	policy only	software enforced partition	cryptographic per member keys
D2 Fidelity floor	none	stated	software enforced	owner cannot reverse by design
D3 Consented and minimal capture	always on	notice only	software controls	hardware shutters and cut
D4 Content incapacity	none	filter	reliable filter	cannot capture or store by architecture
D5 Truthful and non-manipulative	none	policy	enforced and tested	manipulation pathways not built
D6 Harm response without surveillance	none or autonomous report	stated protocol	human fiduciary in loop	narrow triggers, no monitoring, sealed routing
D7 Heirloom continuity	vendor locked	portable data	no coercive lock	escrow and open source failsafe
D8 Verification	claims only	describes mechanism	evidence available	cryptographic event record

A maker discloses a rating on each dimension. The headline conformance level is not an average and not a best case. It is the minimum across all eight dimensions, so that a system is only as trustworthy as its weakest guarantee, and the profile beside it shows exactly where the weakness is. This kills the safety-washing gambit at the root, because a single low dimension caps the whole claim and is visible to anyone reading the profile. Honesty becomes the only way to score well, which is the entire purpose of the instrument.

7. Verification and the Limits of a Non-Binding Framework

This framework is deliberately non-binding, on the model of the automotive self-assessment, because a private standard earns adoption through clarity and use rather than through a claim of authority no single company holds. The known failure mode of that model is disclosure without evidence. The answer is not to abandon the non-binding form for a mandate the author has no power to issue, but to make the disclosure carry evidence, through Principle 8 and Dimension 8. A maker is free to claim nothing, to claim with proof, or to claim without proof, and the profile makes the third case legible as what it is. The market, the press, insurers, and the courts can then do with an honest profile what they cannot do with a badge that says everything and specifies nothing. The maturation path from here runs toward external audit. An honest self-assessment is precisely what

makes an independent audit possible, and the two are meant to meet, with third-party verification and, in time, certification built on the same evidence the profile already requires.

8. The Registry and the Seal

A standard that no one can join is a paper, not a standard. Section 7 argued that an honest profile is more useful than a badge that says everything and specifies nothing. This section describes the instrument that carries the profile out of the paper and into the world: a public registry of makers who pledge to the standard, and a seal that marks the pledge. Both are built to a single rule, which is that the instrument must never be for sale.

The pledge. Adoption is by voluntary pledge, on the model already familiar from three places this paper trusts. The automotive Voluntary Safety Self-Assessment cited in Section 3 is a maker's disclosure against a public template, not a license granted by an authority. The Better Business Bureau's accreditation rests on a pledge to a published code of conduct rather than an inspection of the business. And the SOC 2 practice, though it involves an auditor, is at bottom a disclosure of controls against a named framework that a reader can examine. A maker pledges to adhere to this standard, completes the eight-dimension self-assessment of Addendum A, and publishes the resulting profile with the minimum-rule headline it produces. In return the maker is listed in a public registry and is issued a seal it may display. No party certifies another. The pledge is the maker's own word, made in public, against a fixed ruler, with the weakest dimension setting the headline.

Free pledge, paid only for real work. The pledge, the registry listing, and the seal are free, and they are meant to stay free, because the moment the right to display the seal costs money the standard becomes a toll booth, and the toll booth is the corruption. Charging for the mark would convert an honesty instrument into a pay-to-play badge, which is the precise failure Section 7 warns against. Money may be charged only for a service that gives a maker something it cannot give itself: an independent verification review, an audit, or an advisory engagement that helps a maker actually reach a level it has so far only disclosed. That tier is optional, it is named separately from the pledge, and it buys work, never the mark. It is also deliberately sequenced to come second. The pledge and its public registry launch first and are meant to stand on their own as the baseline of the standard; the optional verification tier is introduced only after the pledge is established, so that the instrument is known as a free and honest disclosure before any paid service is ever attached to it. A maker that never pays a cent holds the same seal, meaning the same thing, as a maker that buys an audit. What the audit buys is a higher and separately disclosed Dimension 8 standing, recorded in the registry as an audited entry, not a better-looking badge and not a higher headline it did not earn on the other seven dimensions.

Why the split is not incidental. The author has a conflict of interest and refuses to conceal it. Oak Forest Robotics both builds domestic robots and proposes this standard. A standard run by a market participant is worth exactly as much as its defenses against self-dealing, and the defenses here are structural rather than promised. The pledge is free, so the standard cannot be sold to the highest bidder. The headline is the minimum across eight dimensions, so no maker, the author in-

cluded, can buy or spin its way past its own weakest guarantee. And the author publishes its own low scores, in Addendum A, scored at the level of its weakest dimension, in public. A standard whose author quietly passes his own test is marketing. This one is arranged so that quietly passing is not available to anyone, least of all its author.

What the seal means, and what it does not. The seal marks one thing precisely. It marks a self-pledged, non-audited commitment to the standard, published as a profile that anyone can read. The mark carries the words PLEDGED SIGNATORY on its face, so that no reader can mistake it for a grade or a certificate, and its motto is the Latin PROTEGERE NON EXTRAHERE, to protect, not to extract, which is the vector thesis of Section 2 compressed to three words. The seal does not mean that any party has certified, verified, tested, or audited the maker's claims. It does not assert compliance, and it does not assert that the product is safe. It asserts that the maker has pledged, and it points to a published profile in which the maker's weakest dimension is visible. Any display of the seal that implies verification the standard did not perform is a misuse of the mark, and this is the one place the standard is strict, because a certification-style mark that implies more than it delivers is the specific thing consumer-protection regulators examine most closely. *(A note on unsettled law: the seal's wording, the claims it may and may not carry, and its treatment as a certification mark should be reviewed against Federal Trade Commission guidance on endorsements and certification marks, and against trademark practice, before any seal is issued.)* Where a maker has bought and passed an independent audit, it may say so, and the registry records it as an audited entry, which is a different and higher statement than the bare pledge. The two are never allowed to blur.

How a maker joins. A maker submits its legal name, a named and reachable contact, the specific product it is pledging for, and its self-scored profile across the eight dimensions, with the evidence note each level above disclosure requires. The submission is published in the registry with the date of the pledge, and the maker receives a confirmation and the seal asset pack. The registry, not the seal, is the source of truth. A seal on a maker's own site is an assertion; the registry entry it links back to is the record that can be checked, updated, or, under Section 9, withdrawn. The mechanics of the sign-up flow, the seal asset pack, and the machine-readable badge that lets an automated agent resolve a claim back to the registry are specified in separate build documentation and are not part of the normative standard.

9. Governance, Maintenance, and Revocation

A seal that cannot be lost means nothing, and a seal that can be lost by whim is a defamation suit waiting to be filed. Both failures are avoided the same way, by making revocation a published process rather than an asserted power. The standard states in advance what can cost a maker its listing, how a maker is told, and how a maker may answer, so that removal is an act of due process and not of discretion. This is what makes revocation defensible for the standard and meaningful for a reader: the seal is worth something precisely because it can be lost, and it can be lost fairly because the grounds and the procedure are fixed in public before any dispute arises.

Grounds for revocation. A listing may be withdrawn for cause on narrow, published grounds. The grounds are: a material misrepresentation in the self-assessment, meaning a disclosed level that the maker knew or reasonably should have known its shipping product did not meet; use of the seal to imply certification, verification, or audit that was never performed; a demonstrated false claim that the maker declines to correct within the cure period; and use of the mark outside the terms of its license, including display on products or by entities that never pledged. The grounds are about honesty and about misuse of the mark, and they are not about a maker's scores. A maker that honestly discloses a profile of all ones is in perfect standing. A low score is never a ground for revocation, because an honest low score is the instrument working exactly as designed. Only dishonesty and misuse put a listing at risk.

Notice and the right to respond. Before any listing is withdrawn, the maker receives written notice that names the specific ground and the evidence for it. The maker then has a defined window, thirty days as the working default, to cure the problem, to correct the public record, or to contest the finding with evidence of its own. Withdrawal follows only if the ground still stands at the end of that window. Nothing in this process removes a maker's own right to withdraw voluntarily at any time, and a voluntary withdrawal is recorded as such, without prejudice, and kept distinct from a removal for cause. *(A note on unsettled law: the cure window, the notice contents, and the standard of evidence for a for-cause removal should be reviewed against defamation and unfair-competition exposure before the registry opens.)*

Maintenance of a listing. A pledge is to a shipping product at a point in time, and products change. A signatory re-affirms its profile on a fixed cycle, annually as the working default, and updates the profile whenever a material change raises or lowers any dimension. A registry entry therefore carries a pledge date and a last-affirmed date, and an entry that lapses past its affirmation cycle is marked stale rather than silently trusted. The registry keeps the history of an entry, so a maker cannot quietly raise a score while the prior claim disappears. The record shows what was claimed, and when, and what changed.

The registry as the record of truth. Because the seal only asserts and the registry verifies, every state of a listing lives in the registry: active, stale, voluntarily withdrawn, or revoked for cause with the ground stated. A reader who follows a seal to its registry entry sees the current truth, which is the entire reason the mark is bound to a public record rather than left to speak for itself. The seal is a licensed mark, granted with the listing and terminating with it, and continued display of the seal after a listing is withdrawn is itself a misuse of the mark. *(A note on unsettled law: the seal is intended to function as a certification-style mark; its licensing terms and the consequences of continued display after revocation should be confirmed under trademark and unfair-competition law.)*

Governance of the standard itself. This first edition is expected to evolve as the field and the law move. Changes to the normative requirements, the dimensions, or the scoring rule are versioned, dated, and published, and a maker's listing names the version it pledged to, so that a later revision never retroactively invalidates an honest pledge made to an earlier one. The author holds the pen today. The stated intention, repeated in the call for comment that closes this paper, is to

move the standard toward multi-party governance over time, so that it does not remain permanently the instrument of a single company that also sells robots. That transition is itself part of the conflict-of-interest answer, and it is offered here as a commitment on the record, not a vague someday.

10. The Open Problem: The Good Parent and the Bad Parent

Every honest standard names the thing it cannot yet solve. Here it is. A child's safety sometimes requires that a parent not see everything, because the parent is the threat. A child's legal protection, under the Children's Online Privacy Protection Act, gives a parent the right to see and delete the child's data, because most parents are the child's best guardian. These are both true, and they use the same key. Grant the good parent the access the law demands and you have handed the bad parent a window. Deny the bad parent the window and you have broken the good parent's lawful right and the law with it. The adversarial review that strengthened this paper contradicted itself precisely here, demanding children be shielded from abusive parents in one place and declaring children can hold no confidentiality from parents in another, which is not a flaw in the reviewer but a faithful reflection of a genuine impasse.

This framework does not claim to resolve it. It offers a partial architecture that reduces the surface of the problem. The two-stream design of Principle 2 means that for most interactions there is no retained inner life transcript for either parent to demand or exploit, while the separate care record preserves only the practical information a good parent legitimately needs, which narrows the conflict to the moment of genuine danger. At that moment, Principle 6 routes the alarm to a neutral human fiduciary rather than to the account holder, which is the only move that does not depend on trusting the parent. What remains unsolved is the case of a legally intact but abusive custodial parent exercising a lawful access right, and no purely technical mechanism dissolves it, because the conflict is in the law and in the family, not in the device. The framework states this plainly so that no maker mistakes a partial answer for a complete one, and it invites the family law and child welfare communities into the design, because this is their frontier as much as it is engineering's.

11. A Call for Comment

This is a first edition, offered as a standard and an invitation, not a decree. It is non-binding by choice, evidence-backed by design, and transatlantic in scope because the machine it governs will not respect a border. The author's interest is disclosed plainly: Oak Forest Robotics builds in this space, and the framework reflects commitments the company has made and, in several dimensions, has not yet met. That is precisely why the self-assessment that follows measures Oak Forest by the same ruler as everyone else, scores it at the level of its weakest dimension, and lists openly the dimensions where it is still disclosure and not yet architecture. A standard whose author quietly passes his own test is marketing. A standard whose author publishes his own low scores, and builds the registry so that free honesty always outranks paid reputation, is the beginning of a shared discipline. Two invitations follow from that. The first is the ordinary one: comment, dis-

sent, correction, and especially legal review are invited on every claim in this paper. The second is larger: the author is actively seeking the family-law and child-welfare experts, the AI-governance scholars, and the privacy and security engineers who would move this from one founder's paper to a governed standard, and is prepared to hand over the pen as that body forms.

References

1. AI Frontiers, *The Robot in Your Living Room Has No Rulebook* (governance gap; ISO 13482). https://aifrontiersmedia.substack.com/p/the-robot-in-your-living-room-has
2. *Anticipating the Use of Robots in Domestic Abuse: A Typology*, ACM. https://dl.acm.org/doi/pdf/10.1145/3610977.3634938
3. 18 U.S.C. 2258A and the REPORT Act; NCMEC CyberTipline reporting, summarized. https://www.traverselegal.com/blog/does-my-ai-company-have-to-report-illegal-content/
4. California SB 243 (effective January 1, 2026), companion chatbot safety, summarized by the Future of Privacy Forum. https://fpf.org/blog/understanding-the-new-wave-of-chatbot-legislation-california-sb-243-and-beyond/
5. Amazon, *Alexa Emergency Assist*. https://www.aboutamazon.com/news/devices/alexa-emergency-assist
6. Children's Online Privacy Protection Act (COPPA), FTC rule and parental review and deletion rights. https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa
7. NHTSA, *Voluntary Safety Self-Assessment* for automated driving systems. https://www.nhtsa.gov/automated-driving-systems/voluntary-safety-self-assessment
8. Driver Privacy Act of 2015 and 49 CFR Part 563, Event Data Recorder ownership and access, summarized. https://nelson.legal/who-owns-your-black-box-data-us-edr-laws-and-regulations/
9. SAE J3016, Taxonomy and Definitions for Terms Related to Driving Automation Systems.
10. State dashboard and in-cabin camera and audio consent law, overview. https://azuga.com/blog/laws-on-inward-facing-cameras
11. EU Artificial Intelligence Act, Article 5, Prohibited AI Practices (applicable February 2, 2025). https://artificialintelligenceact.eu/article/5/
12. General Data Protection Regulation, Articles 5 and 2, on data minimization, purpose limitation, and the household exemption; and scholarship treating device manufacturers as controllers. https://academic.oup.com/idpl/article/10/4/279/5900395
13. Better Business Bureau, Standards for Trust and the accreditation pledge (model for a voluntary, pledge-based commitment). https://www.bbb.org/overview-of-bbb-standards
14. AICPA, SOC 2 and the Trust Services Criteria (model for disclosure against a named framework). https://www.aicpa-cima.com/topic/audit-assurance/audit-and-assurance-greater-than-soc-2
15. Federal Trade Commission, Guides Concerning the Use of Endorsements and Testimonials, and guidance on certification and seal programs (for the honest-seal requirement and the counsel flags

in Sections 8 and 9). https://www.ftc.gov/legal-library/browse/rules/guides-concerning-use-endorsements-testimonials-advertising

Addendum A: The Robotics Guardian Self-Assessment (The Seal)

A non-binding, self-disclosed instrument, modeled on the voluntary safety self-assessment used in automotive. No party certifies another. Each maker discloses a rating on all eight dimensions, states the required evidence where it claims a level above disclosure, and publishes the profile and the resulting minimum. Honesty is the entire value of the instrument, because the score is the weakest dimension and the weakness is visible. Completing this self-assessment and publishing the profile is what a maker submits to join the registry described in Section 8; the pledge, the listing, and the seal are free.

How to score

For each dimension, select the highest level for which every criterion is met **in shipping product**, not in intention. Roadmap commitments are listed separately as **Planned**. Levels above 1 require the stated evidence. The maker's headline conformance level is the **minimum** across all eight dimensions.

D1: Individual sovereignty

- Level 1: privacy policy describes member separation.
- Level 2: software enforces per member data separation; the owner cannot read a member's private channel.
- Level 3: each member's private data is encrypted under a key the account holder cannot access; a specific key can be compelled, sealed, or destroyed by lawful order. Shared space records are access gated, not owner mineable against a member.
- *Evidence:* key management design; access control audit.

D2: Fidelity floor

- Level 1: policy states the system will not act against a member at the owner's instruction.
- Level 2: software refuses such instructions; a minor's confidential inner life stream runs in ephemeral, non-retained processing, while a separate, consented, parent-visible care record holds only practical wellbeing information.
- Level 3: the pathway to weaponize the system against a member is not built, and cannot be enabled by the owner.
- *Evidence:* refusal test results; confirmation that the confidential stream is not retained; care record scope.

D3: Consented and minimal capture

- Level 1: notice of sensing is provided.

- Level 2: software controls allow any member to disable capture.
- Level 3: default privacy-preserving sensing; motorized shutters and a hardware microphone cut; any member's biometrically authenticated cease command severs sensor power and overrides the owner without disabling the Principle 6 safety watchdog.
- *Evidence:* hardware description; override and anti-spoofing test.

D4: Content incapacity

- Level 1: content filter present.
- Level 2: reliable filtering of nudity and genitalia in generation and capture.
- Level 3: the pipeline cannot capture, render, transmit, or store nudity of real people in the home; it is blanked at the sensor and never written. Legitimate art, medical, and educational reference runs through a consented, adult-gated, non-retained path.
- *Evidence:* pipeline design; storage inspection; scoped exception controls.

D5: Truthful and non-manipulative

- Level 1: policy against manipulation and against fabricated certainty.
- Level 2: enforced and tested; states uncertainty on consequential questions; no engagement-maximizing tactics on vulnerable members.
- Level 3: manipulative and vulnerability-exploiting pathways are architecturally absent, consistent with EU prohibited practices.
- *Evidence:* evaluation results; red team findings.

D6: Harm response without surveillance

- Level 1: stated crisis and escalation protocol; no affirmative monitoring.
- Level 2: a licensed human fiduciary, not the machine, makes the reporting judgment; triggers are narrow and published.
- Level 3: anomaly detection with sealed, anonymized routing to the fiduciary; mandated reports route to the proper neutral channel; never to the account holder.
- *Evidence:* trigger list; fiduciary agreement; routing design.

D7: Heirloom continuity

- Level 1: household data is portable and exportable.
- Level 2: no coercive vendor lock or proprietary trap.
- Level 3: core OS, kinematic safety controllers, and inference engine in escrow that open sources on maker failure, preserving security patching.
- *Evidence:* escrow instrument; export format.

D8: Verification

- Level 1: mechanisms are described.
- Level 2: evidence is available to a qualified examiner, and the maker discloses whether it has submitted to independent third party audit.
- Level 3: a cryptographic event recorder preserves safety relevant events, decryptable only by joint owner and lawful authority, and the profile has been independently audited.
- *Evidence:* the recorder specification; the audit report.

Author's self-disclosure: Oak Forest Robotics

Provided as the worked example the instrument demands of its author. The point of the exercise is honest disclosure, including where the product is still policy and not yet architecture.

Oak Forest designs its domestic intelligence, MARS, toward Level 3 on every dimension. As of this first edition it does not claim to be there, and it discloses the following provisional profile:

- D1 Individual sovereignty: **1**, advancing to 3. Distinct members and private channels exist; full per member cryptographic partitioning and a sovereign local store are pending owned hardware.
- D2 Fidelity floor: **1**, advancing to 3. Designed, with the minor escalation architecture not yet built.
- D3 Consented and minimal capture: **2**. Member controls, microphone blocking, and disclosure are enforced in software; hardware shutters and a hardware cut are planned.
- D4 Content incapacity: **1**, committed to 3. Structural nudity and genitalia blanking is a committed build, not yet shipped.
- D5 Truthful and non-manipulative: **2**. Core doctrine, enforced and disclosed; a full architectural guarantee is harder and ongoing.
- D6 Harm response without surveillance: **1**. Soft guardrails exist; the fiduciary escalation path is not yet built.
- D7 Heirloom continuity: **1**. The destination; currently cloud bridged, escrow not yet established.
- D8 Verification: **1**. This paper is the disclosure; the cryptographic event recorder is not yet built.

Headline conformance level: 1 (Disclosed), by the minimum rule.

Oak Forest publishes this precisely because a standard is worth only as much as its author's willingness to be scored by it, at the level of its weakest dimension, in public.