The First Machine Whose Breach Moves

The case for independently verifiable evidence when robots are compromised

Allan Watkins

Laws of Robots series

Version 1.1 — 31 October 2026

Companion to "The Maker Has Always Graded Its Own Homework" and "An Anchor Outside the Industry."

Abstract

The companion papers argue for scrutiny independent of the maker and examine a certification anchor against that requirement. This paper asks what those arrangements must withstand when the machine is compromised. Its premise has a record: malicious code has targeted industrial processes and safety systems, and cybersecurity vulnerabilities have prompted vehicle and medical-device recalls. Those cases support two proposals. The certifying role should report rather than command the machine, and evidence offered in support of its reports should be independently verifiable without requiring its live server. Neither proposal guarantees safety or exposes every breach. The companion paper's offline rehearsal demonstrates a narrower operating property: saved public inputs reproduce the shared bounded verification values at the same verification time. The public check authenticates a certifier-signed checkpoint; it does not independently inspect the private conduct record or establish its completeness. This paper distinguishes that result from the conduct contract it proposes: evidence of governed conduct should survive compromise under stated assumptions, integrity failures should be made visible, and the record should help people investigate and assign responsibility. The historical cases motivate that design. They do not establish universal detection, automatic liability, or the absence of earlier physical harm from autonomous systems.[1,2]

§1 — When a breach becomes physical action

A computer breach can expose, alter, encrypt, or deny information. When software also commands a physical process, the consequences can include movement, interruption of a safety function, or injury. Restoring a backup cannot undo a physical action that has already occurred. Information harms can also be irreversible; the distinction here concerns the additional authority placed in software, rather than a claim that data loss is harmless or recoverable in every case.

Security engineering already recognizes this distinction. NIST's operational-technology guidance addresses systems that interact with the physical environment and their particular safety, reliability, and performance requirements. Its examples include industrial control, building automation, transportation, and physical access control. The present final revision was published in 2023, succeeding a 2015 revision.[3] This paper builds on that established concern. It does not claim that the security discipline has only just discovered physical consequences.

Autonomous machines make the governance question more demanding because software can select an action as well as carry it out. A machine may perceive a situation, interpret an instruction, and move in space shared with people. The assurance question includes what happens when that decision process is manipulated, what evidence is retained, and who can examine it without relying exclusively on the interested parties' account.

The pattern paper's revised historical comparison supports independent examination, the ability to challenge a verdict, and authority to require correction. It does not prove that every self-assessment arrangement must fail or that one institutional reform guarantees safety. [1] The anchor paper then separates publicly checkable properties from the author's statements and unbuilt intentions. Its revised Appendix A supplies the verification exercise on which this paper relies. [2] This paper adds an adversary to that examination and asks which conclusions the evidence actually supports.

The method follows those distinctions. An indictment alleges; a government advisory reports or assesses; a researcher analysis supplies its own evidence and limitations. The proposals drawn from the cases are the author's reasoning. Defensive implementation remains outside the paper's scope, but that exclusion is not a substitute for stating the limits of the claims made here.

§2 — The record of physical consequences

The cases below establish the premise that authority over software can become authority over a physical process. They include documented incidents and regulatory responses to vulnerabilities. Those are different kinds of evidence, and they remain different throughout the argument.

Stuxnet provides the technical case. Ralph Langner's 2013 analysis distinguishes an overpressure attack from a later rotor-speed attack against centrifuge systems. Recording and replaying process inputs belonged to the overpressure attack; it should not be described as the rotor-speed attack's method. The analysis instead explains how suspending legitimate control logic could leave displayed rotor-speed values unchanged. [4] The relevant lesson is the combination of process manipulation and interference with the information available to operators, with the attack variants kept distinct.

The scale of the damage requires separate attribution. David Albright, Paul Brannan, and Christina Walrond estimated from inspection data that about one thousand centrifuges were decommissioned and replaced at Natanz in late 2009 or early 2010. Their December 2010 assessment considered mechanical and operational explanations and treated Stuxnet's contribution as uncertain. Their February 2011 update supplied further support for the centrifuge-targeting interpretation. [5,6] Replacement estimates are not an adjudicated count of machines destroyed by malware. This paper relies on the published technical analyses and makes no attribution of responsibility for the operation.

The Ukraine grid incidents carry an official investigative and charging record. The industrial-control emergency-response alert about December 23, 2015 reported remote intrusions at three regional distributors and outages affecting approximately 225,000 customers. It described remote breaker operations, wiped systems, overwritten communications-device firmware, and

telephone denial of service. Its qualification matters: the interagency team based its assessment on interviews and corroborating documentary findings and could not independently review the technical evidence.[7] Disrupting restoration and communications is evidence of a broader attack on operations; it is not proof that every such act was intended to falsify a conduct record.

The October 2020 indictment alleges that six Russian military-intelligence officers participated in attacks including the 2015 incident and a December 17, 2016 intrusion that interrupted electricity in Kyiv for approximately one hour. It describes purpose-built grid malware and a disk-wiping component in the latter attack.[8] These are criminal allegations, not findings of guilt. The paper uses the dated charging document without asserting a current trial status or claiming that it independently proves every technical detail.

The vehicle example concerns a vulnerability and a safety recall. In July 2015, FCA initiated a recall covering approximately 1.4 million vehicles over radio-software vulnerabilities that could permit unauthorized remote modification and control of vehicle systems. The report describes a potential injury risk. It also expressly says FCA had not determined that a statutory defect existed and was conducting a safety recall to protect customers.[9] The supported proposition is that cybersecurity risk prompted use of safety-recall machinery. A voluntary safety recall should not be presented as a formal adjudication of a statutory defect.

The medical examples bring that risk close to the body. In 2017, Abbott issued a pacemaker firmware correction to mitigate unauthorized radio-frequency access; its contemporaneous announcement reported no unauthorized access to a patient's implanted device and did not recommend replacement of implanted devices.[10] In June 2019, the FDA warned that a nearby unauthorized party could potentially change certain MiniMed insulin-pump settings and affect delivery. Its announcement recommended switching affected pumps and stated that it knew of no confirmed reports of related patient harm. The FDA recall record identifies the action initiated on June 27.[11] Those dated statements concern known reports and potential harm, not proof that exploitation could never have occurred.

Together, the cases support a consequential premise: cybersecurity can be a physical-safety matter, including before exploitation has been observed. They do not establish one uniform attack method, a universal concealment pattern, or the proposition that all earlier examples occurred in protected industrial settings. Vehicles and personal medical devices already place software-controlled physical functions near people.

§3 — Actuation replication and presence

An **Actuated Breach** is a compromise whose consequences include physical action by the affected system. The term names the consequence class this paper examines. It does not require autonomous judgment: a remotely opened breaker qualifies. For an autonomous machine, the additional question is how an adversary can influence the process that selects the action.

Three properties explain the concern. First, actuation connects compromise to what the machine can physically do. Exposure depends on its capabilities, operating environment, safeguards, and the people nearby. Existing safety and incident-response practice remains relevant; the challenge is to carry that practice into the assurance arrangements for autonomous conduct.

Second, common software can create shared exposure. A vulnerability in a common component may affect many units, but the affected population depends on versions, configuration, access, and mitigations. One compromised unit does not establish that every unit can be compromised, or that failures will occur simultaneously. The pattern paper's concern about replicated judgment supports examining fleet exposure without substituting a universal forecast for evidence.[1]

Third, proximity changes who bears the risk. Some robots operate in controlled industrial settings; others operate in homes, care environments, workplaces, or public space. The paper does not assume that every consumer machine lacks safeguards or that every machine is continuously reachable. It asks what evidence and authority the arrangement provides where close contact and autonomous action do occur.

The possibility of manipulating robotic judgment has also been demonstrated in research. RoboPAIR, published by its authors in 2024, reports controlled experiments eliciting harmful physical actions from LLM-controlled robotic systems, including a commercial robot dog.[12] These experiments are not evidence of an uncontrolled malicious incident in public use. They nevertheless belong in the record and remove any basis for treating the technical possibility as wholly untested.

Autonomous systems also have a history of physical harm without a demonstrated cyberattack. The NTSB investigated the fatal 2018 collision involving a developmental automated-driving system in Tempe.[13] That case does not establish an actuated cyber breach. It does establish why "no catastrophe yet" is an unsuitable description of autonomous systems generally. The question here is how a governed machine's first serious compromise will be documented and answered, rather than when physical harm or robotic attack research will first appear.

§4 — The safety layer as a target

The 2017 Triton case makes a further point: a protective system can itself be attacked. A safety instrumented system monitors a process and can act to bring it to a safe condition. Its protective role gives its integrity particular importance. The joint CISA, FBI, and Department of Energy advisory describes Triton as targeting Schneider Electric Tricon safety controllers at a foreign refinery and explains the potential physical consequences of disabling or modifying their safety functions.[14]

The Department of Justice's March 2022 account of the unsealed indictment alleges that the conspirators installed malware designed to prevent the refinery's safety systems from functioning while appearing normal. It further alleges that faults during deployment caused two automatic emergency shutdowns in 2017, followed by attempts against a U.S. company's systems in 2018.[15] The allegation framing applies throughout. Neither source establishes that the intruders controlled every other plant system, that the safety system was the hardest available target, or that it is always the most valuable target in any facility.

The author's architectural conclusion is narrower. A role given authority to command a machine creates a consequential capability that an attacker may seek to obtain. Access to one component is not automatically access to every authority behind it; the boundaries between components

matter. Nevertheless, reducing the authority available to a compromised certifying role can reduce what that particular compromise enables.

For the governance arrangement proposed in this series, **the layer that vouches for the machine must be incapable of commanding it.** Its job is to issue and report a bounded attestation, with any response taken through the owner and other responsible people. That is a proposed division of authority. It is not an argument for removing local safety controls whose protective purpose requires them to act.

The companion paper's published verification formats carry results rather than machine-control commands.[2,16] A reader can inspect those interfaces and confirm that limited property. The formats alone cannot prove the absence of every other control path in an entire deployment, or establish the trustworthiness of its unseen infrastructure. Those broader questions require separate evidence.

Report-only design also leaves reporting risks. An attacker might suppress or replay information, and compromise of a valid signing key could permit false material that passes signature verification. Outside verification checks the evidence against its anchors and rules; it does not guarantee that the issuer's assertions are true. Limiting command authority and authenticating reports address different risks, and neither completes the safety case.

§5 — Call-home dependency and its failure modes

A machine may exchange diagnostics, updates, instructions, or records with its maker. Those functions need separate examination. A connection that is useful for maintenance does not necessarily authorize remote commands, and evidence delivered online does not necessarily depend on an editable server copy for its authenticity.

The liability paper calls attention to Self-Attested Telemetry: evidence held and presented by an interested party requires scrutiny beyond the amount of data collected.[17] This paper adds a security question: which compromises could alter, suppress, or withhold the evidence, and which checks remain outside those compromised components? Custody by a maker does not by itself prove one undivided security domain. Separate keys, access controls, and retained copies can create meaningful boundaries. Their existence and effectiveness must be established rather than assumed.

A **Call-Home Dependency** is a design in which continued function, or access to evidence needed for verification, requires one particular party's live infrastructure. A required command or update channel can be an attractive target; a required service can also fail without an intruder. The relevant questions are what authority the channel carries and which functions become unavailable when the service stops.

Revolv supplies an official record of the latter risk. In 2016, FTC staff investigated Nest's termination of the services needed to operate the Revolv hub and app. The closing letter records customer-unusability concerns and refunds, and explains the decision not to recommend enforcement action. It expressly disclaims a determination that a violation did or did not occur.[18] Insteon's June 9, 2022 statement supplies a different outcome: new owners announced that hubs were coming back online, with some functions still being restored.[19] These examples

illustrate service dependence, including restoration; they do not establish that every attached device became permanently inert.

For a governance arrangement, independent retention and verification of evidence can reduce dependence on the issuer's service. That is an integrity and availability benefit with a defined boundary. Signatures can authenticate material received over an untrusted network. A live connection does not imply that the network can silently edit a signed record. Conversely, an authentic saved record does not establish that no later event or revocation exists. Authenticity, availability, completeness, and currency are distinct questions.

§6 — What the offline demonstration establishes

The defensible property is specific: a relying party should be able to verify supplied evidence from saved files and independently trusted anchors, without needing the certifier's server for that computation. The companion paper's revised Appendix A documents an operating example and gives the tested preparation and comparison instructions. [2]

In the October 31, 2026 rehearsal, the existing reference verifier used saved public inputs, the matching public key, and the same verification time. Its shared bounded result values matched the hosted result after presentation normalization. A modified grant signature and a verification time after grant expiry were rejected. Verification succeeded in a network-restricted execution environment while a control HTTP request failed. The workstation's physical network adapter remained connected. [2]

The public bundle contained an empty chain segment and no robot audit public key. Its integrity result therefore authenticated a certifier-signed checkpoint. The authorized full-record tier described in the specification is a different check: it supplies the chain segment and robot audit key needed for an independent chain walk. The rehearsal did not exercise that private tier. [2,16] Neither a checkpoint signature nor a successful chain walk, by itself, proves that every real-world action was recorded or that every entry is truthful.

Time is another boundary. The tested checkpoint and anchor/revocation statement were signed as of June 21, 2026. The October endpoint check time did not turn those inputs into newly signed currency evidence. An offline verifier can assess the material and time supplied to it. It cannot learn a later revocation from files saved before that revocation. A relying party must judge the age and coverage of the evidence and seek updated material when its decision requires it. [2]

Trust in the anchor also remains necessary. A claim and a public key downloaded together demonstrate consistency under that key; they do not independently establish whose key it is. Offline reproduction removes the need to trust a live server's computation. It does not remove the need for a justified trust anchor or make an issuer's false assertion true.

The companion appendix remains the home of the walkthrough. The public specification describes the reference verifier package as forthcoming; a reader currently needs an available build or an implementation of the specification. [16] This paper does not represent a public ready-to-run download as already available. Its demonstrated claim is reproducibility of the shared bounded values from the tested saved inputs, with the limitations preserved alongside the result.

§7 — A conduct contract and its limits

Assume that a machine can be compromised. The governance question then concerns what evidence remains available, what failures can be made visible, and which conclusions a person can reasonably draw. The historical cases show reasons to protect both operations and the information used to understand them. They do not establish that attacking the record is always a separate second job, or that concealment is uniformly cheap.

The design target is still demanding: falsifying or suppressing evidence should require crossing a boundary beyond the compromised machine, under stated assumptions about custody, keys, and the recording path. A contract with a machine that may be lying must be checkable by someone the machine cannot deceive. That sentence states the requirement. The public checkpoint demonstration alone does not establish that an entire machine satisfies it.

The proposed conduct contract has three clauses. They are requirements and design reasoning, rather than additional operating properties inferred from the offline test.

First: governed conduct should produce evidence that can be checked outside the compromised component. Signed, linked records and independently retained checkpoints can make certain alterations detectable under their verification rules. The assurance case must separately establish which conduct is captured, whether an attacker can bypass capture, what defines the expected end of the record, and what prevents undisclosed branches or suppression. Verification of a supplied prefix cannot establish that it is the whole history. The public tier authenticates the checkpoint assertion; the authorized tier can examine the chain supplied to it.[16]

Second: detected integrity failures should produce an externally available change of status. The bounded certification claim described by the companion paper concerns reasoning-engine integrity faults that would cause self-de-certification.[2] Repeating that claim does not establish that every corruption is detected or that publication survives every compromise or outage. Detection, status issuance, delivery, and currency each need evidence. Harmful conduct can also occur while software and its record remain cryptographically intact, including through adversarial inputs. An intact status is therefore not a general assurance that the machine is uncompromised, correct, or safe.

Third: evidence protection should make concealment a separate problem for the adversary. Independent custody and keys outside the compromised component can increase the work needed to forge a consistent account. That is the author's design reasoning, conditional on those boundaries holding. No comparative cost study is offered, and the historical cases do not establish a universal economic result. A stolen signing key, an unrecorded action, or a stale saved checkpoint can defeat assumptions that an abstract appeal to an append-only chain leaves unstated.

The scope boundary follows. This contract is not a general breach detector. Compromises can affect privacy, sensors, actuators, communications, or the decision process without producing the particular integrity failure that a verifier tests. A snapshot saved before a breach may remain valid evidence about its earlier coverage while saying nothing about the breach. No paragraph here claims that every compromise which changes physical conduct must announce itself through this arrangement.

What the contract supplies is a standard against which evidence can be demanded. Which paths are covered? Which keys and witnesses remain independent? Which faults have been tested? How old is the status? What remains unknowable from the supplied material? Those questions can be answered at the level of properties and test evidence without publishing defensive implementation. The difference between a requirement and a demonstrated result is part of the contract, not a qualification to be removed from it.

§8 — How evidence can support responsibility

The certifying role reports. A revoked, failed, or self-de-certified status changes the information available to people; it does not seize, disable, or command the machine. The owner and other responsible parties act through their own authority. That division preserves the paper's central proposal while keeping the certifier's reporting function separate from the machine's physical control.

Consider a status change supported by evidence whose provenance and dates can be checked. An owner can investigate and take appropriate action. A counterparty can consider the evidence before admitting a machine to a facility. An insurer can assess whether it is relevant to a policy, and an investigator can compare records with physical and other digital evidence. These are possible uses of the arrangement. This paper does not report adoption by an insurer or acceptance by a court.

For a court or adjuster, trustworthy records may improve reconstruction and help test a party's account. They do not automatically establish causation, a duty, a breach of duty, or liability. Authorization, evidence completeness, admissibility, and the applicable law remain material. The liability paper supplies the broader argument for an evidence-bearing governance arrangement; this paper makes no claim that a signed record alone converts every unassignable loss into an assignable one.[17]

The sentence "it was hacked" should be an investigable proposition. Existing incident investigation already tests such claims; outside-verifiable records can contribute an additional source of evidence. An intact chain may preserve evidence of harmful conduct without disproving compromise. A failed check may indicate alteration, damaged material, missing inputs, or another verification problem without proving an intrusion. Investigators must establish which explanation fits. The point is to improve the evidence available to challenge an account, not to turn either verification outcome into a verdict. An unfalsifiable claim is where responsibility goes to disappear.

Actionable responsibility requires more than information integrity. People need lawful access to relevant evidence, competence to interpret it, and authority to require correction or allocate loss. The record can support those institutions; it cannot substitute for them. This aligns the security argument with the revised pattern paper's institutional lesson.[1]

The proposal is therefore practical and bounded. Keep the certifying role from commanding the machine. Preserve independently checkable evidence beyond a single live service. State the assumptions and residual risks. Then show, through testing and investigation, how much of the proposed conduct contract an actual implementation satisfies. Consequence can move through

people when the evidence supports it; the arrangement must not promise the conclusion in advance.

§9 — What this paper asks of its readers

The first exercise is to inspect the published verification formats and the limitation statement that travels with their results. Confirm what those interfaces carry. Treat the absence of machine-control commands in them as an interface property, rather than proof about every component of a deployment.[16]

The second is to follow the companion paper's revised Appendix A with its evidence-age and exposure-tier qualifications intact. Compare the shared bounded values for the same inputs and verification time. Distinguish a reader's physical-disconnection exercise from the network-restricted rehearsal actually documented, and a public checkpoint check from an authorized chain walk.[2] A reproduced historical result is useful precisely because its scope is explicit.

The third is to test the published Independent Verification Specification's claim of sufficiency by implementing it independently. The local reference-tool rehearsal did not establish that a stranger could build a verifier without consulting that implementation. A successful independent build, or a documented specification defect, would add evidence on that separate question.[16]

The author's conflict remains part of the account. The companion paper disclosed the author's role in operating both the certified system and the certifier and distinguished architectural properties from institutional independence.[2] This paper does not upgrade those disclosures into an independent assessment of present governance, legal separation, or key custody. A website, a published format, and a successful signature check do not establish those institutional facts. Nor does this paper present a commissioned external security audit or tested adoption by insurers or courts.

Defensive implementation is withheld by the author's disclosure policy. That policy does not prove the defenses effective. Claims about public properties must remain open to examination, and claims about unseen controls require appropriately authorized evidence. Publishing a threat model is useful only if its assumptions and exclusions are candid.

The series' argument holds at that size. The pattern paper makes the case for independent scrutiny and effective correction. The anchor paper offers a bounded verification arrangement with its conflicts and limits disclosed. This paper asks what evidence and authority should survive when the machine is compromised. Its title looks toward the incident report for a machine governed under those requirements. The test will be what the retained evidence permits people to establish, challenge, and act upon. That is the conduct contract worth building before the incident occurs.

Appendix A — Physical incidents vulnerabilities and research

Incident accounts, allegations, regulatory actions, and controlled research are identified separately. The table makes no claim that an autonomous-machine category has an empty harm history.

Year	System	Evidence and physical consequence	Source and boundary
2009–10	Centrifuge systems at Natanz	Distinct overpressure and rotor-speed attacks; concealment methods differ. About 1,000 units replaced in the independent estimate.	Researcher technical analysis; replacement estimate and malware causation distinguished. [4,5,6]
2015	Ukrainian electricity distributors	Remote breaker operations and outages affecting approximately 225,000 customers; disruption of systems and communications.	Interview-based government assessment with stated forensic limit; later criminal allegations. [7,8]
2016	Ukrainian transmission system	Indictment alleges purpose-built malware, disk wiping, and an approximately one-hour Kyiv outage.	Dated criminal allegations; no finding of guilt asserted. [8]
2017	Refinery safety controllers	Safety functions targeted; DOJ account alleges two automatic emergency shutdowns during deployment.	Joint government advisory and account of an indictment; attribution and intent retain source status. [14,15]
2015	Vehicle systems	Approximately 1.4 million vehicles subject to safety recall for software-security risk.	Manufacturer recall report expressly withholds a statutory-defect determination. [9]
2017	Implanted pacemakers	Firmware correction to mitigate unauthorized access.	Recall record and contemporaneous manufacturer statement; no reported unauthorized patient-device access in that statement. [10]
2019	Insulin pumps	Nearby unauthorized access could alter delivery settings; recall and replacement recommendation.	FDA records; no confirmed related patient-harm reports known to FDA at its announcement. [11]
2018	Developmental automated-driving system	Fatal pedestrian collision investigated by NTSB.	Physical-harm history, not evidence of a cyber breach. [13]
2024	LLM-controlled robotic systems	Controlled research elicited harmful physical actions.	Author research report; not an uncontrolled malicious incident in public use. [12]

Appendix B — Threat-model scope statement

In scope. Compromise affecting a machine's governed conduct; alteration, suppression, or replay of records and status material; attacks on the reporting and verification roles; service dependence; and shared exposure across affected units. These are classes to examine, not assurances that the demonstrated verifier detects them all.

Assumptions requiring evidence. Which actions enter the record; protection of the recording path and signing keys; independently justified anchors; retained checkpoints and expected record boundaries; detection and publication of particular faults; and sufficient evidence currency for the relying party's decision. A check of supplied material does not establish these assumptions by itself.

Residual risks. Validly signed false assertions following issuer or key compromise; omitted or bypassed recording; undisclosed branches or truncated histories without an independently known boundary; stale evidence; harmful conduct under intact software; and attacks on sensors, actuators, or private data that leave the tested integrity properties unchanged. Public checkpoint authentication and authorized full-record inspection have different coverage.

Excluded implementation detail. Defensive and response mechanisms are outside this paper's disclosure scope. General owner-network security, physical attacks, and hardware supply-chain assurance are not independently evaluated. Where they can defeat an assumption above, the exclusion does not turn them into a guarantee.

References

[1] Allan Watkins, *The Maker Has Always Graded Its Own Homework*, version 1.2 (31 October 2026), Abstract, §§7–8, and Appendix. Companion paper in this series.

[2] Allan Watkins, *An Anchor Outside the Industry*, version 1.1 (31 October 2026), §§4.1–4.2, 4.4–4.5, and revised Appendix A, including the dated local verification rehearsal. Companion paper in this series.

[3] NIST, *Guide to Operational Technology (OT) Security*, SP 800-82 Rev. 3 (September 2023), official abstract and publication history. Revision 4 was an initial public draft when checked on 31 October 2026. Publication record.

[4] Ralph Langner, *To Kill a Centrifuge: A Technical Analysis of What Stuxnet's Creators Tried to Achieve* (November 2013), pp. 5–14, especially pp. 13–14 on the difference between replay and rotor-speed concealment. Original report on the author's company site.

[5] David Albright, Paul Brannan, and Christina Walrond, *Did Stuxnet Take Out 1,000 Centrifuges at the Natanz Enrichment Plant? Preliminary Assessment* (22 December 2010), opening assessment and Traditional Reasons for Centrifuge Problems. Original analysis.

[6] David Albright, Paul Brannan, and Christina Walrond, *Stuxnet Malware and Natanz: Update of ISIS December 22, 2010 Report* (16 February 2011), Overview and revised assessment. Original update.

[7] ICS-CERT/CISA, *Cyber-Attack Against Ukrainian Critical Infrastructure*, IR-ALERT-H-16-056-01 (2016; revised 20 July 2021), Summary and Details. The original alert is reproduced in Exhibit 1, pp. 1–3, to the Kellie Cloud declaration filed 11 March 2022 in Oregon PUC docket UM 2164; PDF pp. 17–19. Government-filed copy of the alert.

[8] *United States v. Andrienko et al.*, indictment returned 15 October 2020 and unsealed 19 October 2020, W.D. Pennsylvania, Count One, paras. 20–23, pp. 9–11. Original indictment.

[9] FCA US LLC, Part 573 Safety Recall Report, NHTSA 15V-461, submitted 23 July 2015, pp. 1, 5–6, Description of Defect, Safety Risk, and Remedy. Original manufacturer report filed with NHTSA.

[10] FDA, Recall Z-0035-2018, event 78093, initiated 28 August 2017, Manufacturer Reason and Action; Abbott, *Abbott Issues New Updates for Implanted Cardiac Devices* (29 August 2017), unauthorized-access report and firmware-update advice. FDA recall record; manufacturer announcement.

[11] FDA, *FDA Warns Patients and Health Care Providers About Potential Cybersecurity Concerns with Certain Medtronic Insulin Pumps* (27 June 2019), opening paragraphs, potential delivery effects, and replacement recommendation; FDA Recall Z-1581-2020, event 83433, initiated 27 June 2019, Manufacturer Reason and Action. Agency-issued announcement; FDA recall record.

[12] Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J. Pappas, *Jailbreaking LLM-Controlled Robots*, arXiv:2410.13691v2 (9 November 2024; first submitted 17 October 2024), abstract and experimental settings. Original author paper.

[13] NTSB, *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018*, HAR-19/03 (2019), What Happened and What We Found. Investigation record.

[14] CISA, FBI, and DOE, *Tactics, Techniques, and Procedures of Indicted State-Sponsored Russian Cyber Actors Targeting the Energy Sector*, AA22-083A (24 March 2022), pp. 2, 4. Original advisory retained in a House hearing record.

[15] DOJ, *Four Russian Government Employees Charged in Two Historical Hacking Campaigns Targeting Critical Infrastructure Worldwide* (24 March 2022), section on *United States v. Evgeny Viktorovich Gladkikh*. Describes the June 2021 indictment and alleged 2017 shutdowns and 2018 attempts. DOJ account.

[16] *Independent Verification Specification*, v1, §§0, 3–7 and the packaged-verifier notice, checked 31 October 2026. Distinguishes public signed checkpoints from authorized chain verification. Published specification.

[17] Allan Watkins, *When the Robot Is Wrong, Someone Pays*, version 1.1, §5 and its evidence-and-liability argument. Companion paper in this series.

[18] Mary K. Engle, FTC Division of Advertising Practices, closing letter to Richard J. Lutton Jr., Nest Labs, FTC File 162-3119 (7 July 2016), pp. 1–2. Original letter.

[19] Ken Fairbanks, *A New Day for Insteon!* (9 June 2022), acquisition, hubs returning online, and restoration still in progress. Original company statement.

Colophon

Allan Watkins

Laws of Robots series

Version 1.1 — 31 October 2026

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