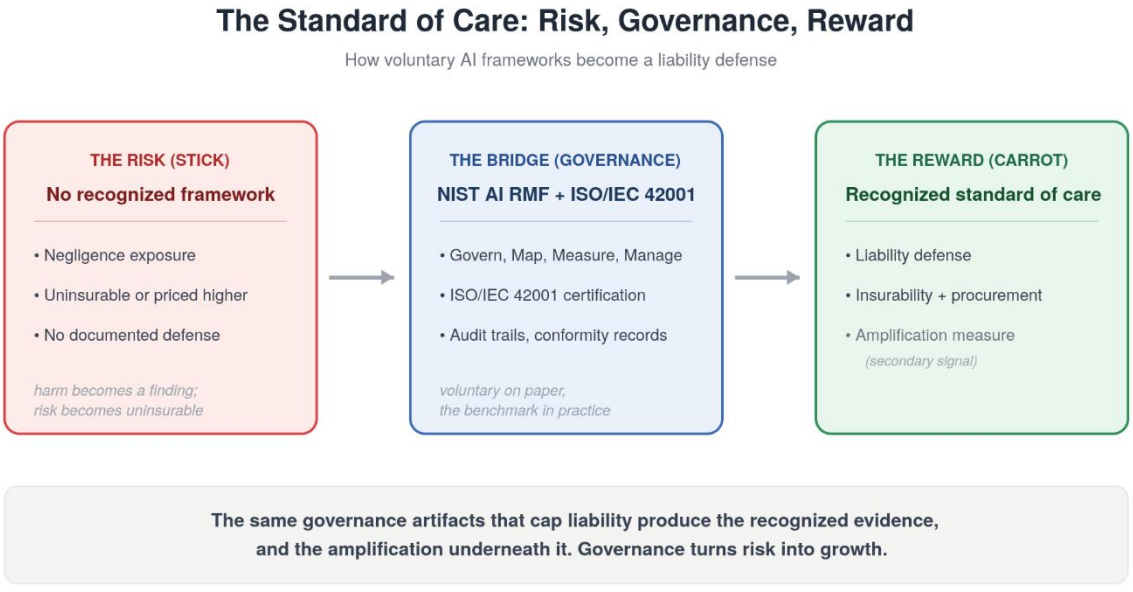
The Standard of Care: How NIST and ISO Are Turning Voluntary AI Governance Into a Liability Defense

Basil C. Puglisi, MPA
A Human-AI Collaboration

Most organizations treat two of the best-known AI governance standards, the NIST AI Risk Management Framework and ISO/IEC 42001, as "nice to have." Both are voluntary, after all. In the United States, no general AI statute compels adoption of either framework directly, no regulator issues a fine for skipping them, and a busy executive can be forgiven for treating them as paperwork that competes with shipping product. That treatment is a mistake, and the people most likely to expose it are not regulators. They are plaintiffs' attorneys and insurance underwriters.

The reason rests on a concept that predates artificial intelligence by more than a century: the standard of care. In ordinary negligence law, an organization is judged not against perfection but against what a reasonable organization in its position would have done. Once a recognized practice becomes common enough, courts and insurers increasingly treat it as persuasive evidence of what reasonable care looks like. This is not automatic, and it cuts both ways. Compliance does not guarantee a defense, and falling short is not negligence by itself. But a practice widely adopted is hard to ignore when a court or an underwriter asks whether an organization behaved reasonably. For AI systems, that benchmark is taking shape now, and these two standards are among the most visible forms it is taking. Each deserves a plain introduction before the stakes make sense.



What the NIST framework actually is

Begin with the agency behind it. The National Institute of Standards and Technology (NIST) is a United States federal agency inside the Department of Commerce, and it develops the measurements, guidance, and technical standards much of the economy quietly relies on, from cybersecurity to manufacturing tolerances to weights and measures. It is not a regulator, and it issues no fines. Its role is to define what good practice looks like; markets, procurement systems, expert practice, and courts often give that work its practical force later.

Its AI Risk Management Framework, released in January 2023, offers a voluntary, risk-based way to manage AI across its lifecycle. The architecture is simple enough to explain in a sentence and deep enough to organize an entire program. It rests on four functions: Govern, which sets the policies and accountability; Map, which identifies context and risks; Measure, which evaluates systems for problems such as bias, privacy leakage, and unreliable output; and Manage, which treats those risks with safeguards, monitoring, and a plan for retirement.

In July 2024, NIST extended the framework with a companion document, the Generative AI Profile, catalogued as NIST AI 600-1. It defines a set of risks specific to or worsened by generative systems, including fabricated output, data leakage, harmful content, and information-integrity failures, and it ties suggested actions back to the same four functions. The base framework supplies the operating model. The profile supplies the generative-specific risk taxonomy. Organizations deploying or integrating foundation models, the large general-purpose models that power most generative AI, can apply both, adjusted to their role in the AI lifecycle.

NIST does not enforce any of this. Yet the framework's voluntary status does little to reduce its practical weight, because it has become the common vocabulary for AI risk in the United States, increasingly referenced in public-sector guidance, procurement, and enterprise governance programs. When buyers, regulators, and expert reviewers reach for a shared reference, this is often the one already in their hands.

What ISO/IEC 42001 adds

If NIST is the American reference, ISO and IEC are the international one. The International Organization for Standardization and the International Electrotechnical Commission are independent global bodies that publish the voluntary standards businesses worldwide are measured against, the same family that produced ISO 9001 for quality management and ISO 27001 for information security. Published in December 2023, ISO/IEC 42001 is the world's first international AI management system standard. It governs how an organization runs its AI, not whether any single model is accurate or fair, a distinction worth holding onto.

Where the NIST framework describes what good risk management looks like, ISO/IEC 42001 establishes an auditable structure that an accredited certification body can certify against, the same way organizations certify against ISO 27001 for information security. The two fit together rather than compete. Organizations can use the NIST framework as the risk-management

engine inside an ISO/IEC 42001 management system, combining the framework's substance with the standard's certifiable form. The certification ecosystem is itself new, with accreditation for certifiers rolling out through 2024 and 2025, which is part of why an early certificate may still carry weight.

That certificate matters for a reason that has nothing to do with vanity. A certification from an accredited certification body converts an internal claim into external evidence. "We govern our AI responsibly" is an assertion. A current ISO/IEC 42001 certificate, backed by the records behind it, is third-party evidence of conformity within the certificate's defined scope.

These standards are voluntary, but they do not sit in a vacuum. In the European Union, the AI Act will make risk management mandatory for high-risk systems as its obligations phase in. It relies on harmonized standards, the technical standards that grant a presumption of conformity once the EU formally lists them, to help providers comply, a regime this series takes up on its own. The argument here is narrower and, for most organizations, more immediate. Even where no statute yet applies, the voluntary standard is already doing legal and commercial work.

The risk: voluntary on paper, load-bearing in practice

Here is where the standard of care turns from theory into exposure. When an AI system causes harm, whether through a biased hiring screen, a discriminatory loan denial, or a defamatory fabricated output, the question that follows is whether the organization behaved reasonably. The legal theory varies: negligence for some harms, anti-discrimination statutes like the Equal Credit Opportunity Act or Title VII for others, and an organization's exposure also turns on its role, whether it built the model or merely deployed someone else's. Across many of them, a recognized standard can become a yardstick, evidence of whether the organization took the precautions a careful peer would have taken.

This is terrain Joy Buolamwini has mapped for years. Her 2018 Gender Shades study, with Timnit Gebru, documented commercial systems failing darker-skinned women at rates as high as roughly 35 percent, against near-perfect accuracy on lighter-skinned men, the kind of algorithmic audit a regulator understands.

Her 2022 MIT doctoral thesis formalized a second form of evidence, the evocative audit, which pairs what a system produces with lived experience to show harm people can feel rather than only measure. Its mechanism is the counter-demo: a system's own output turned into evidence against its claims. Her 2018 visual audit "AI, Ain't I a Woman?", which the thesis takes up as a case study, is where those examples originate: commercial systems unsure whether Michelle Obama's hair was a toupee, and reading iconic Black women, Serena Williams among them, as men. In the author's reading, the point is that numbers move institutions while the counter-demo moves people, and that accountability needs both. This article extends Buolamwini's audit logic into legal-governance terrain, and that extension belongs to the author, not necessarily to her. Her work reads here as an early naming of the kind of harm a court could later weigh when assessing reasonableness, and an insistence that it be witnessed, not only counted.

A plaintiff's attorney could point to the NIST framework and ISO/IEC 42001 and ask a straightforward question: these recognized practices existed, they were widely adopted, and the defendant ignored them, so on what basis was the conduct reasonable? Voluntary industry or consensus standards have repeatedly become persuasive evidence of the standard of care, admitted to show what a careful organization would have done. They are evidence, though, not a safe harbor. As Judge Learned Hand wrote in *The T.J. Hooper*, 60 F.2d 737 (2d Cir. 1932), a whole industry can lag in adopting an available precaution, and a court, not the industry, ultimately decides what prudence required. So adopting one of these standards is no guarantee and ignoring one is not automatic negligence, but the gap between common practice and a defendant's choices is exactly where a negligence argument lives.

There is an honest counterargument, and it cuts toward adoption rather than against it. A defendant could argue these standards are too new to be settled custom, so declining to adopt them is not yet unreasonable. That very newness is the reason to move now: an organization that adopts while the floor is still being set assembles the record before the standard hardens, not after.

The insurance industry faces governance pressure of its own. The National Association of Insurance Commissioners, through a model bulletin many states have adopted, asks insurers to maintain a written AI program and governance documentation for their own use of AI systems. That pressure is one reason documented AI governance is increasingly familiar across the industry.

Underwriters, the people who decide whether to cover a risk and at what price, work from what they can assess: recognized controls, attestations, and loss history. An applicant with an ISO/IEC 42001 certificate and a documented NIST-aligned program presents a measurable, familiar risk. An applicant with neither presents an opaque one. The first may be easier to underwrite. The second may draw more questions, higher pricing, tighter terms, or less appetite, depending on the carrier and the risk. As AI-specific exposures move into the language of cyber and professional-liability policies, the absence of a recognized standard can read, to an underwriter, as an unpriced hazard.

This is how AI risk actually behaves. It does not wait for an AI statute to become real. It arrives through the doctrines and the markets that already exist, negligence and, where applicable, product-liability theories on one side, underwriting and reinsurance on the other, and it tends to land first on the organization that cannot show what a reasonable one would have shown.

The reward: defense that compounds into advantage

The case for adoption usually stops at fear, and that is a mistake, because the same work produces a genuine upside. The first and most important reward is a set of recognized artifacts. Those artifacts include an ISO/IEC 42001 certificate, a NIST-aligned risk register (the documented list of a system's risks and how each is handled), conformity records, model and data documentation, test results, approval logs, and an audit trail. These are the kinds of

materials that help defend against a negligence claim, satisfy an underwriter, and clear a procurement questionnaire. They are the currency that lawyers, insurers, and enterprise buyers increasingly ask for or recognize, which is why the reward leads with them rather than with anything proprietary. The organization that builds them early is not performing compliance theater. It is assembling the record a careful field now treats as standard, while the record is still affordable to assemble.

The second reward is larger than the defensive framing suggests, and it is the part the fear leaves out. The comparison that matters now runs between organizations, the ones that govern the collaboration well and the ones that do not, more than between humans and machines. By the available third-party data, the firms pulling ahead are not the ones that emptied their payrolls. PwC's 2025 Global AI Jobs Barometer, built from close to a billion job postings and thousands of company financial reports, found that productivity growth in the industries most exposed to AI has nearly quadrupled since 2022, from about 7 percent to about 27 percent, while revenue per employee in those industries grew roughly three times faster than in the least exposed, and job numbers kept rising even in highly automatable roles. The organizations capturing those gains restructured how human judgment and machine execution fit together, keeping the human in charge of the machine's work rather than removing the human from it.

Whether that advantage lasts depends on the same choice, and the clearest evidence comes from a setting far from the boardroom. A 2025 field experiment with nearly a thousand high school mathematics students, published in the Proceedings of the National Academy of Sciences, gave students two versions of the same AI: a standard chatbot, and one constrained to offer teacher-style hints rather than answers. Both lifted performance while the tool was in hand, the standard version by 48 percent and the constrained version by 127 percent. The difference surfaced once the tool was taken away. Students who had used the unconstrained chatbot then scored 17 percent worse on an unaided exam than students who had never had access, while the constrained version eliminated that loss. The same technology, governed two different ways, either degraded the human capability underneath or compounded it. The method was the variable, not the tool.

That distinction, augmentation rather than displacement, also runs through the work of Daron Acemoglu. The MIT economist and 2024 economics Nobel laureate was recognized for work on institutions and prosperity, and his separate writing on technology argues that innovation is not destiny. As the author reads him, a general-purpose technology can raise productivity broadly or concentrate its gains and displace the people it was meant to help, and which outcome arrives is decided by incentives and institutions rather than by the technology itself. The author may be drawing the line to governance more sharply than Acemoglu would, but on that reading governance is part of the institutional choice that decides whether the reward is shared or captured.

This is why the defensive case and the growth case are the same case. The discipline that produces the records, deciding who is accountable, logging what was reviewed, and recording

where a human exercised judgment, is the discipline that keeps a workforce sharpening rather than hollowing out. Governance is the competitive asset. The tool is available to every competitor; the discipline to govern it is not. An internal measure of that effect is not yet a recognized underwriting standard, so the recognized artifact still carries the liability case, but the organization that governs the collaboration well is not choosing between defending itself and growing. It is doing both with the same work.

The bridge: governance is the common mechanism

The reason these rewards arrive together is the argument the series turns on, and two things are worth separating cleanly. NIST and ISO are recognized governance baselines in standards, procurement, regulatory, and insurance conversations. What follows is the author's proposed method for producing the evidence that baseline calls for, offered as one approach rather than a standard.

That method, which the author calls Checkpoint-Based Governance, would place a named human at the points where accountability, authorization, or legally consequential use occurs. A checkpoint is a defined moment in the workflow: a required human sign-off before an output ships or an automated decision takes effect. Around those checkpoints, the method would generate the records the Measure and Manage functions expect: role records showing who did what; source custody showing where each input came from; and a dissent log showing where a reviewer disagreed and why. Together they would help answer the question a court or an underwriter may ask, namely whether a person held binding authority over the output.

This is a question Stuart Russell has pressed on the field. In Human Compatible he argues, on the author's reading, that the danger of a capable system is not malfunction but competent pursuit of a mis-specified goal, and that the safer path is to build systems that stay uncertain about human aims and therefore defer to human correction, a related alignment property often called corrigibility, the willingness of a system to be corrected or shut down by a person.

A named human at a defined checkpoint would be the institutional echo of that idea. Russell may frame the goal differently than a governance practitioner would, but the throughline the author sees is the same: the human stays the authority.

The supporting record matters as much as the checkpoint. A tamper-evident audit trail turns a contested claim into a verifiable record, and none of these pieces is exotic. An organization that assembles them is building the record before it needs it.

When those practices run, an organization is not choosing between defense and growth. The records that satisfy ISO/IEC 42001 are also the records that help answer a plaintiff, and the discipline that answers the plaintiff is the discipline that shows whether the work was real. Governance is the mechanism that turns the liability an organization must avoid into the growth it wants to claim.

None of this is a guarantee

A standard adopted is not a standard honored. A certificate can harden into a checkbox, a documented process can be performed badly, and a record built carelessly can be discovered later and used against the organization that made it. Claiming adherence to a framework and then ignoring it in practice can be worse than never claiming it, because it hands a plaintiff the argument that the organization knew the risk and proceeded anyway. The standard is not the paperwork. It is whether the work behind the paperwork was real.

The floor is rising

The standard of care for AI is not static, and it is not waiting. Each new certification, each regulator that references the NIST framework, and each insurer that asks for an AI governance program raises the floor a little higher. Organizations that adopt these standards now build the documented record that helps protect them, and they build it before they need it, which is the only time the record is affordable to build. Those that wait will not escape the standard of care. They will meet it for the first time in a deposition or a claim file, which is the most expensive place to learn what reasonable looks like.

The Other AI: Audio Briefings on Augmented Intelligence and AI Governance

Spotify; Apple Podcasts; Amazon Music; YouTube Playlist;

#AIassisted using HAIA Ecosystem

Sources and currency

- National Institute of Standards and Technology. *AI Risk Management Framework (AI RMF 1.0, NIST AI 100-1).* Released January 2023. https://www.nist.gov/itl/ai-risk-management-framework
- National Institute of Standards and Technology. *Generative Artificial Intelligence Profile (NIST AI 600-1).* Released July 26, 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
- International Organization for Standardization. *ISO/IEC 42001:2023, Artificial intelligence, Management system.* Published December 2023. https://www.iso.org/standard/42001
- National Association of Insurance Commissioners. *Model Bulletin: Use of Artificial Intelligence Systems by Insurers*. Adopted December 4, 2023. State adoption checked June 2026.
- Regulation (EU) 2024/1689 (the EU AI Act). High-risk obligations phasing in, with timing subject to the Digital Omnibus amendment under consideration as of mid-2026.
- *The T.J. Hooper*, 60 F.2d 737 (2d Cir. 1932).
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. *Generative AI Without Guardrails Can Harm Learning: Evidence from High School Mathematics*. Proceedings of the National Academy of Sciences 122(26), 2025. https://doi.org/10.1073/pnas.2422633122
- Buolamwini, J. *Facing the Coded Gaze with Evocative Audits and Algorithmic Audits*. Doctoral dissertation, MIT, February 2022.
- Buolamwini, J. *AI, Ain't I a Woman?* Spoken-word visual audit, 2018.
- Buolamwini, J. *Unmasking AI: My Mission to Protect What Is Human in a World of Machines*. Random House, 2023.
- Buolamwini, J., and Gebru, T. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. 2018.
- Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*. 2019.
- Acemoglu, D., and Johnson, S. *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. 2023. See also Acemoglu, D., and Restrepo, P. *Automation and New Tasks* (2019).
- PwC. *2025 Global AI Jobs Barometer*. June 2025. https://www.pwc.com/gx/en/services/ai/ai-jobs-barometer.html
- Puglisi, B. C. *Checkpoint-Based Governance*. Author-proposed operating model, basilpuglisi.com.

Regulatory, standards, and insurance landscape verified as of June 2026. The standards cited are stable. The legal, insurance, EU AI Act implementation, and ISO/IEC 42001 certification-market claims are the time-sensitive parts and should be re-verified before reuse after roughly 60 days.