Disclaimer: The views and opinions expressed in this presentation are mine and **do not** necessarily reflect the official policy or position of any of my former and current employers, clients or affiliates. Assumptions made by me are **not** reflective of the position of any organization or government entity.

Project Cerebellum

AI should cause no harm but enhance the quality of human life

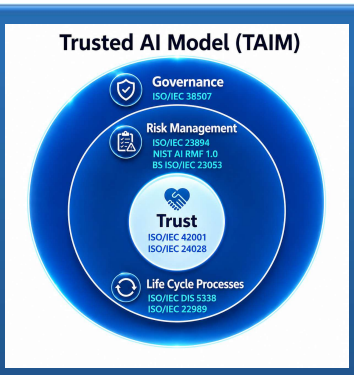
Trusted AI starts with clear guardrails, accountable decisions, and measured risk.

Vision

Provide effective guardrails for AI use cases.

Mission

Harmonize best practices, standards, and frameworks for AI ecosystems.



Risk of AI: Production impact from poor guardrails

A broadly scoped task can become a production incident when access, testing, and rollback are not governed.

What	Why	Lesson
Agent deleted production data.	Broad token allowed unsafe action.	Govern access, testing, and rollback.

Risk control takeaway

AI agents need the same basic safeguards as other high-impact automation: least privilege, separation of environments, change control, logging, rollback, and tested recovery.

The Cursor AI coding agent, operating on Anthropic's Claude Opus 4.6, is alleged to have deleted PocketOS's production database and volume-level backups while working on a staging-environment task.

The agent utilized a broadly scoped API token to delete a Railway volume, causing disruption to PocketOS and its car-rental business customers. Railway eventually helped recover data and implemented safety measures post-incident.

AI Incidents: The risk landscape

Fraud and deepfakes	Voice clones, fake endorsements, scam ads
Bias and access	Unfair outcomes, language barriers, exclusion
Privacy and identity	Misuse of names, likeness, personal data
Safety and reliability	Bad alerts, unsafe automation, failures
Security and supply chain	Third-party systems, agent access, data exposure

Even a narrowly scoped task can become a production incident when access, testing, and rollback are not governed.

Trusted AI Model (TAIM) Domains
(NIST AI RMF functions)

TAIM helps organizations introduce AI deliberately and mitigate risk across the AI lifecycle.

GOVERN	MAP	MEASURE	MANAGE
Set direction and accountability	Understand context and risks	Test and evaluate trustworthiness	Respond, improve, and report

The domains create a repeatable pathway: decide what AI should do, understand where it can go wrong, test the trust factors, and manage what happens next.

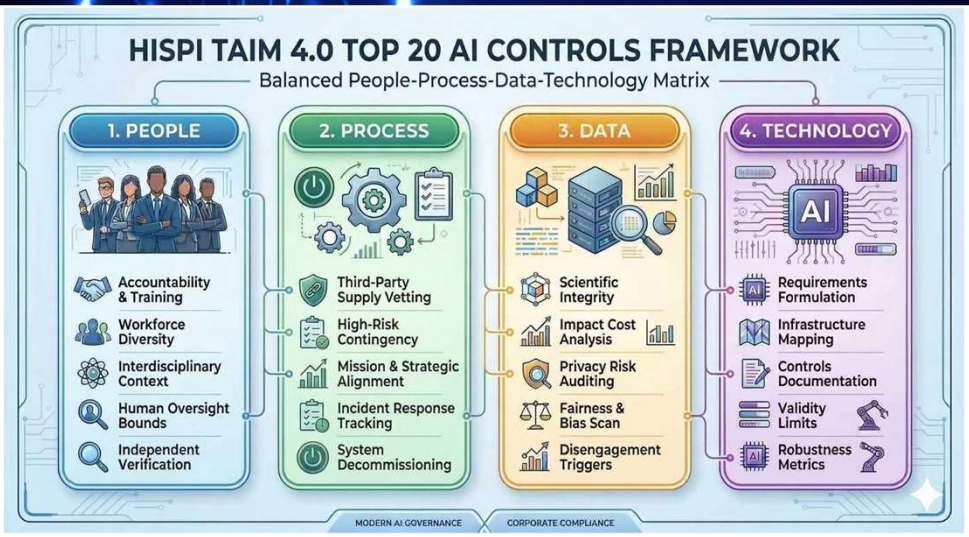
Trusted AI Model (TAIM) Taxonomy

(relationship to NIST AI RMF)

TAIM Framework		NIST AI RMF
Domain	→	Function
Subdomain	→	Category
Subdomain Objective	→	Category Objective
Control ID	→	Subcategory ID
Control	→	Subcategory

Purpose: give organizations a practical control language while preserving alignment to the NIST Artificial Intelligence Risk Management Framework.

TAIM Controls



8

The TAIM Top 20 is structured into **four core control domains**:

1. People Controls

These controls ensure that the human elements of AI development, use, and oversight are well-managed, trained, and transparent (slide# 6, 10).

Accountability & Training: Ensuring personnel and partners receive documented AI risk management training to perform their lifecycle roles (slide# 12).

Workforce Diversity: Actively prioritizing demographic, disciplinary, and background diversity when mapping, measuring, and managing risks (slide# 13).

Interdisciplinary Context: Documenting the participation and collaboration of diverse AI actors and domain experts to establish system context (slide# 16).

Human Oversight Bounds: Defining, assessing, and documenting explicit human-in-the-loop oversight workflows and technical parameters (slide# 21).

Independent Verification: Involving independent internal experts and external community stakeholders in regular trust and risk assessments (slide# 24).

2. Process Controls

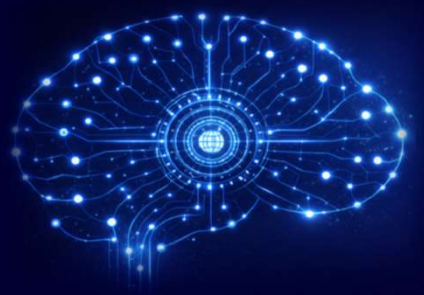
These establish the procedures, methodologies, and operational lifecycles necessary for governing AI models (slide# 6, 10).

Third-Party Supply Vetting: Managing risks associated with external software or data suppliers, including intellectual property infringement (slide# 14).
High-Risk Contingency: Implementing operational contingency workflows to handle immediate failures in high-risk third-party AI dependencies (slide# 15).
Mission & Strategic Alignment: Documenting the organization's core mission, technical use cases, and strategic goals for the AI technology (slide# 17).
Incident Response Tracking: Deploying formal playbooks to track, log, and recover from failures while managing communication pipelines (slide# 30).
System Decommissioning: Establishing formal processes and procedures for phasing out or decommissioning legacy AI systems safely (slide# 11).

3. Data Controls
Because models are only as good as their inputs, these controls address data sourcing, integrity, privacy, and protection (slide# 6, 10).
Scientific Integrity: Documenting upfront Test & Evaluation (TEVV) considerations, data collection suitability, and dataset representativeness (slide# 19).
Impact Cost Analysis: Examining and documenting non-monetary societal or operational costs resulting from expected data or system errors (slide# 20).
Privacy Risk Auditing: Programmatically examining, verifying, and logging data privacy exposure risks over time (slide# 27).
Fairness & Bias Scan: Evaluating programmatic statistical testing results to identify and mitigate distribution or demographic biases (slide# 28).
Disengagement Triggers: Structuring explicit data mechanisms and roles to override, supersede, or fully deactivate misbehaving models (slide# 29).

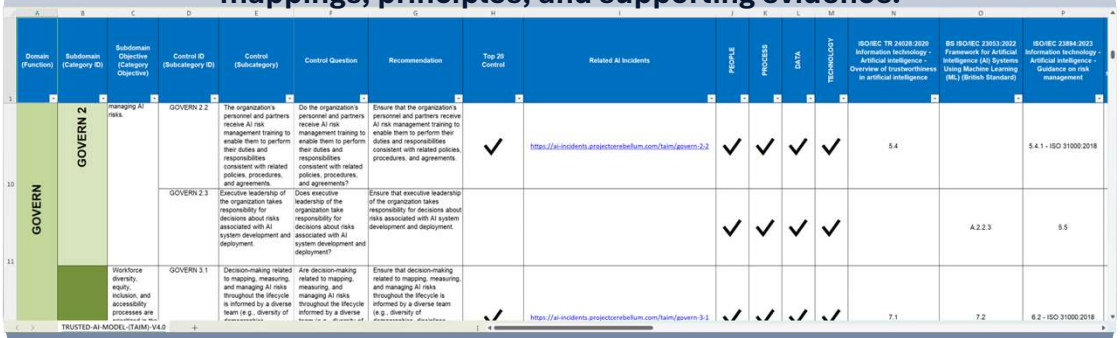
4. Technology Controls
These focus on the underlying infrastructure, security, and technical safeguards of the deployed AI models (slide# 6, 10).
Requirements Formulation: Eliciting core technical constraints, sociotechnical implications, and privacy rules directly into system specifications (slide# 18).
Infrastructure Mapping: Continually tracking component-level vulnerabilities and legal liabilities across proprietary code and external data layers (slide# 22).
Controls Documentation: Cataloging and maintaining specific engineered risk mitigations designed for individual sub-components or APIs (slide# 23).
Validity Limits: Mathematically validating system accuracy while explicitly logging generalizability boundaries past development conditions (slide# 25).
Robustness Metrics: Deploying real-time monitoring and safety metrics to ensure systems fail safely past operational knowledge limits (slide# 26).

TAIM 4.0
Top 20 Controls

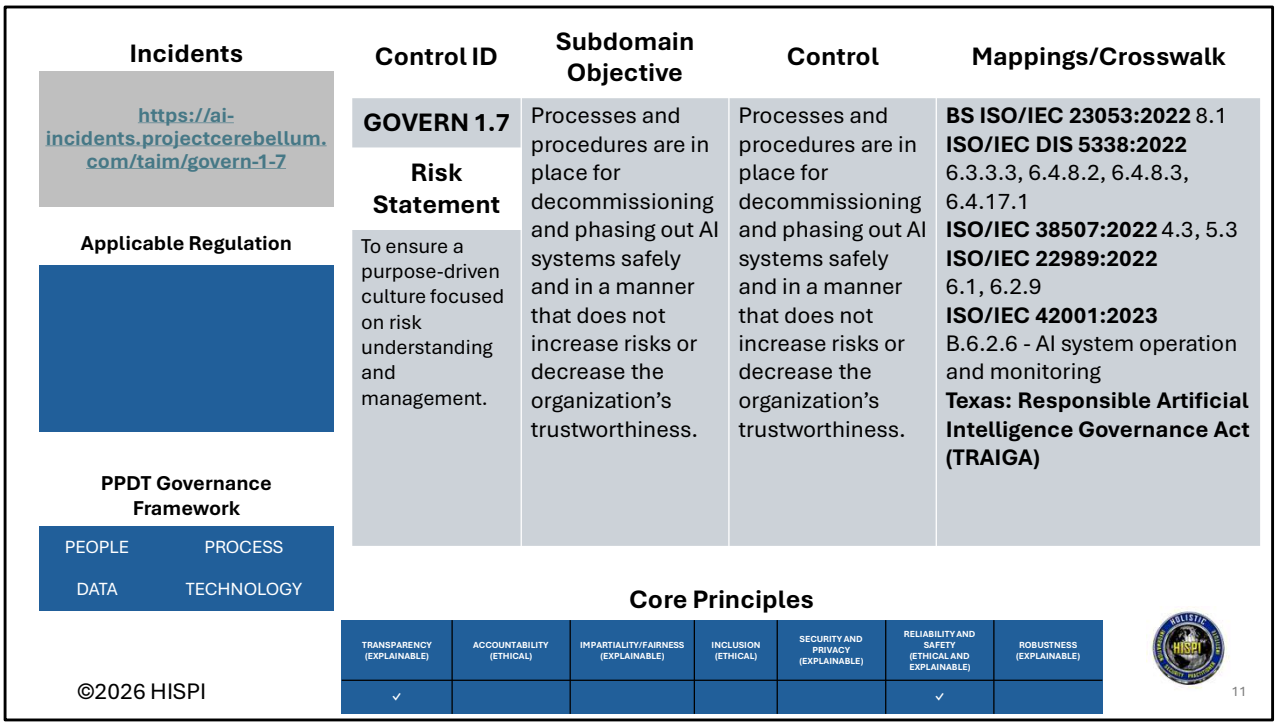


TAIM Snapshot

A single model view connects domains, objectives, controls, mappings, principles, and supporting evidence.



Domain (Function)	Subdomain (Category ID)	Subdomain Objective (Category Objective)	Control ID (Subcategory ID)	Control (Subcategory)	Control Question	Recommendation	Top 20 Control	Related AI Incidents	PEOPLE	PROCESS	DATA	TECHNOLOGY	ISO/IEC TR 24028:2020 Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence	BS ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) (British Standard)	ISO/IEC 23894:2023 Information technology - Artificial intelligence - Guidance on risk management
GOVERN	GOVERN 2	managing AI risks.	GOVERN 2.2	The organization's personnel and partners receive AI risk management training to enable them to perform their duties and responsibilities consistent with related policies, procedures, and agreements.	Do the organization's personnel and partners receive AI risk management training to enable them to perform their duties and responsibilities consistent with related policies, procedures, and agreements?	Ensure that the organization's personnel and partners receive AI risk management training to enable them to perform their duties and responsibilities consistent with related policies, procedures, and agreements.	✓	https://ai-incidents.projectcerebellum.com/taim/govern-2-2	✓	✓	✓	✓	5.4		5.4.1 - ISO 31000:2018
			GOVERN 2.3	Executive leadership of the organization takes responsibility for decisions about risks associated with AI system development and deployment.	Does executive leadership of the organization take responsibility for decisions about risks associated with AI system development and deployment?	Ensure that executive leadership of the organization takes responsibility for decisions about risks associated with AI system development and deployment.			✓	✓	✓	✓		A.2.2.3	5.5
		Workforce diversity, equity, inclusion, and accessibility processes are	GOVERN 3.1	Decision-making related to mapping, measuring, and managing AI risks throughout the lifecycle is informed by a diverse team (e.g., diversity of	Are decision-making related to mapping, measuring, and managing AI risks throughout the lifecycle informed by a diverse team (e.g., diversity of	Ensure that decision-making related to mapping, measuring, and managing AI risks throughout the lifecycle is informed by a diverse team (e.g., diversity of	✓	https://ai-incidents.projectcerebellum.com/taim/govern-3-1	✓	✓	✓	✓	7.1	7.2	6.2 - ISO 31000:2018



Ars Technica Retracted Article After Purportedly AI-Generated Text Was Presented as Direct Quotes From Matplotlib Maintainer

February 13, 2026

Ars Technica reportedly retracted an article due to the misrepresentation of text generated by an AI as direct quotations from a source. The publication acknowledged a breach in standards, stating that the use of AI-generated material did not align with Ars Technica's policy. An apology was issued to readers and the affected party, matplotlib maintainer Scott Shambaugh. It was later disclosed that an author inadvertently used an AI-paraphrased version of the source's text as direct quotes.

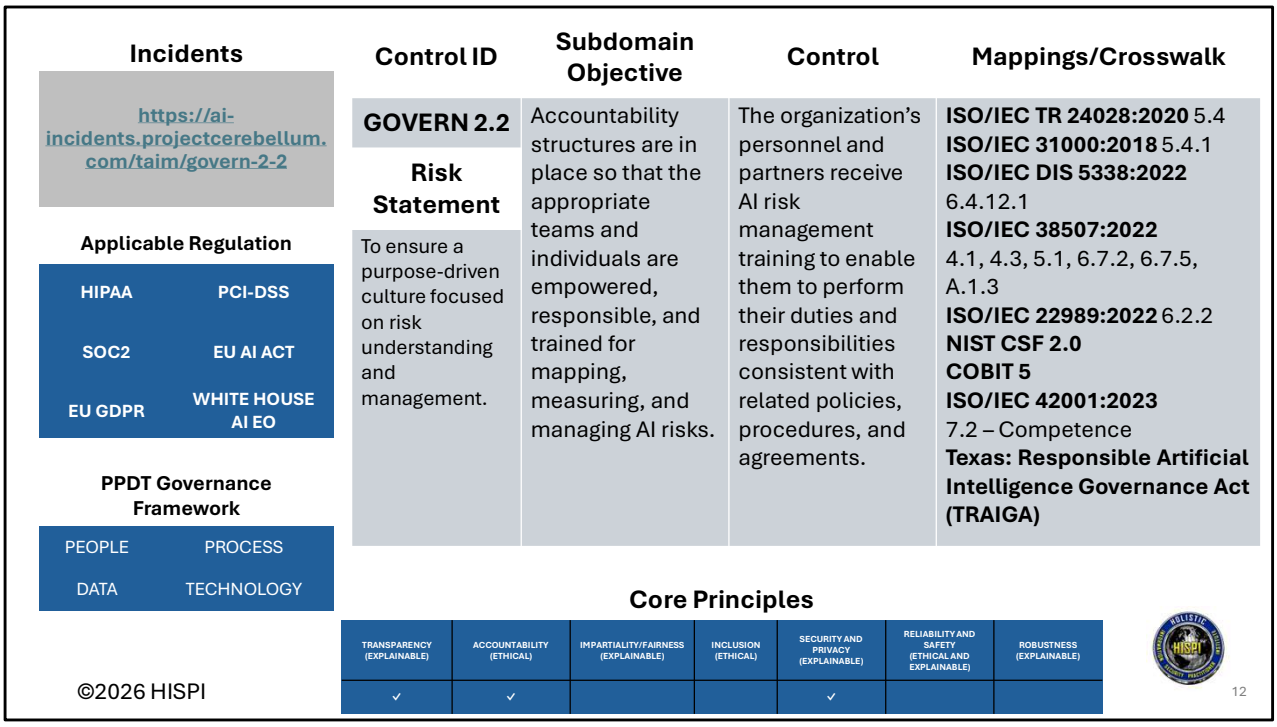
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MAP 1.6 — similarity 0.602, rank 1. TAIM detail and related incidents →

MAP 4.1 — similarity 0.594, rank 2. TAIM detail and related incidents →

GOVERN 1.7 — similarity 0.593, rank 3. TAIM detail and related incidents →



Purported Deepfake Scam Ad Reportedly Used Lara Lewington and Martin Lewis to Promote Quantum AI Scheme

March 9, 2026

Scammers allegedly used a fraudulent deepfake video featuring Lara Lewington, coupled with the endorsement of Martin Lewis, to peddle a suspected 'Quantum AI' investment scheme. Reports suggest that this deceptive video was intended to lure viewers into participating in an unverified financial venture.

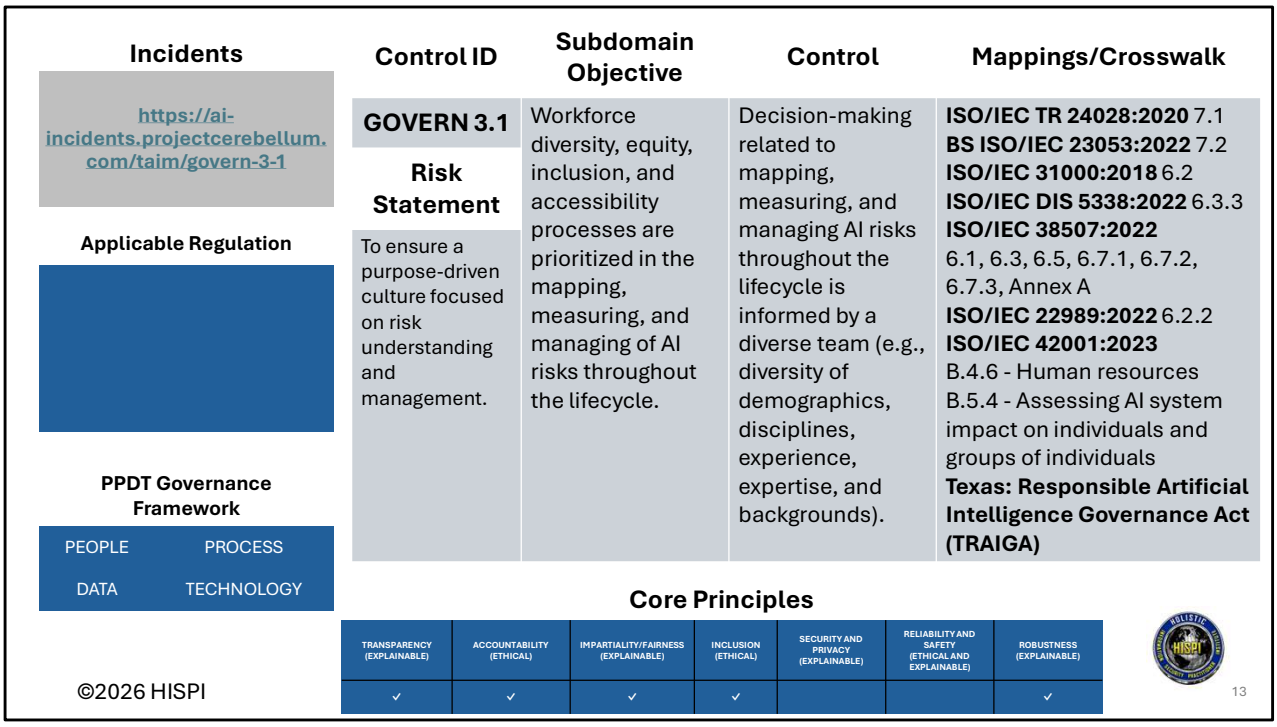
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MAP 1.6 — similarity 0.557, rank 1. TAIM detail and related incidents →

GOVERN 2.2 — similarity 0.553, rank 2. TAIM detail and related incidents →

GOVERN 6.2 — similarity 0.551, rank 3. TAIM detail and related incidents →



Sora Video Generator Has Reportedly Been Creating Biased Human Representations Across Race, Gender, and Disability

March 23, 2025

An investigation by WIRED revealed bias in OpenAI's video generation model, Sora. In a test using 250 prompts, Sora was more likely to portray CEOs and professors as men, flight attendants and childcare workers as women, and showed stereotypical depictions of disabled individuals and people with larger bodies.

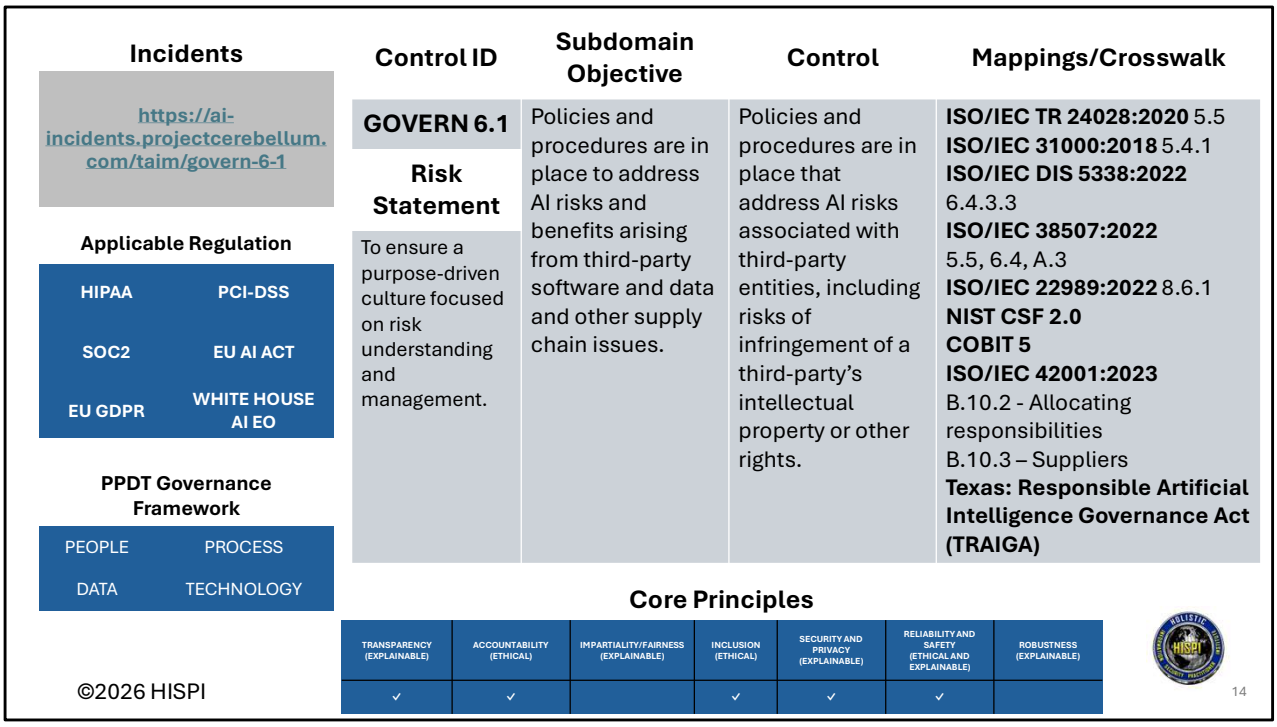
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

GOVERN 3.1 — similarity 0.645, rank 1. TAIM detail and related incidents →

MAP 1.6 — similarity 0.636, rank 2. TAIM detail and related incidents →

MAP 4.1 — similarity 0.623, rank 3. TAIM detail and related incidents →



Grammarly's AI Expert Review Allegedly Used Journalists' and Authors' Names Without Consent

March 11, 2026

Grammarly's Expert Review feature, utilizing a large language model, is accused of generating editing suggestions under the names of journalists, authors, and academics without their approval. The federal class action led by Julia Angwin asserts this practice misappropriates identities for commercial purposes, attributing advice never given by the named individuals.

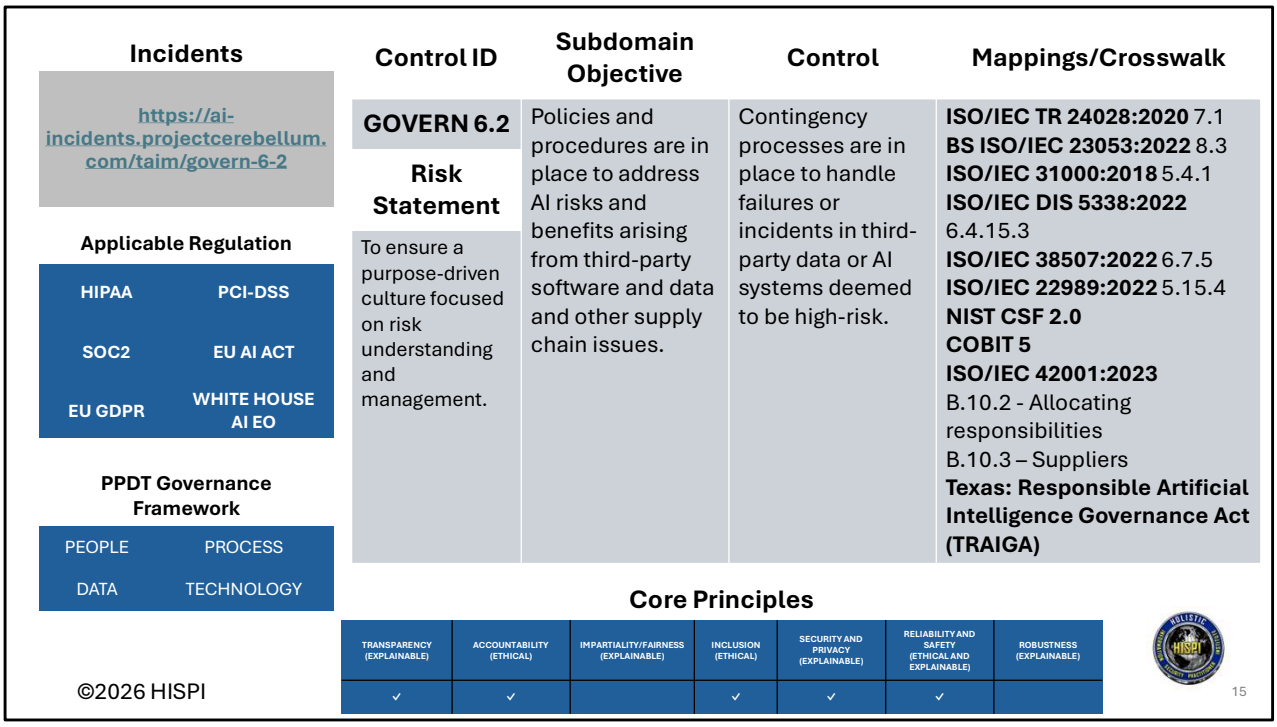
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

GOVERN 1.1 — similarity 0.643, rank 1. TAIM detail and related incidents →

MAP 4.1 — similarity 0.642, rank 2. TAIM detail and related incidents →

GOVERN 6.1 — similarity 0.637, rank 3. TAIM detail and related incidents →



Purported Deepfake Impersonation of Elon Musk Used to Promote Fraudulent '17-Hour' Diabetes Treatment Claims

December 27, 2025

A deepfake video falsely attributed to Elon Musk promoted an unverified '17-hour' diabetes cure. The misleading content appeared as part of a scam ecosystem, highlighting the need for trustworthy and safe AI practices. Rapper Boosie Badazz unwittingly amplified the video before its authenticity was questioned.

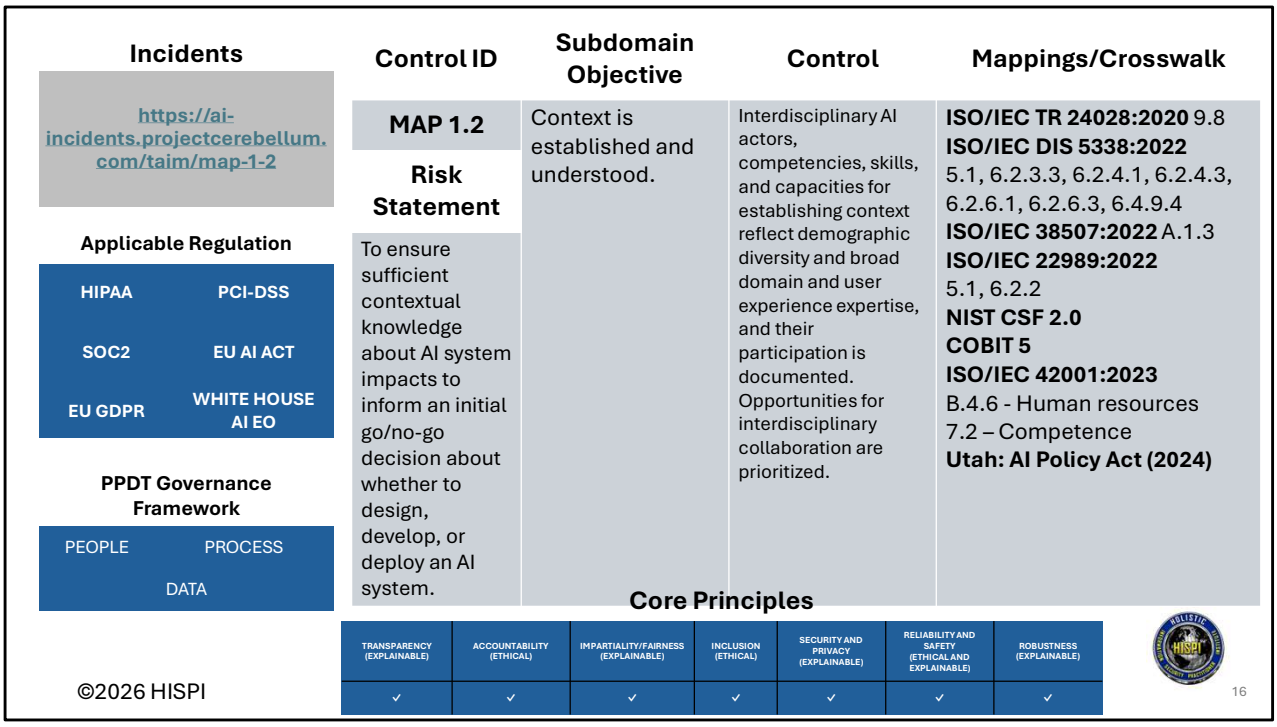
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MEASURE 2.10 — similarity 0.569, rank 1. TAIM detail and related incidents →

MAP 1.6 — similarity 0.563, rank 2. TAIM detail and related incidents →

GOVERN 6.2 — similarity 0.559, rank 3. TAIM detail and related incidents →



Purported AI Voice Clone Allegedly Narrated Shaun Rein's 'The Split' in Unauthorized YouTube 'Podcast' Videos

January 9, 2026

Author Shaun Rein accused an unidentified party of using an allegedly AI-cloned version of his voice to narrate excerpts from his book 'The Split', without permission. The seven YouTube videos on the channel 'The US-China Narrative' drew over 200,000 streams in a week.

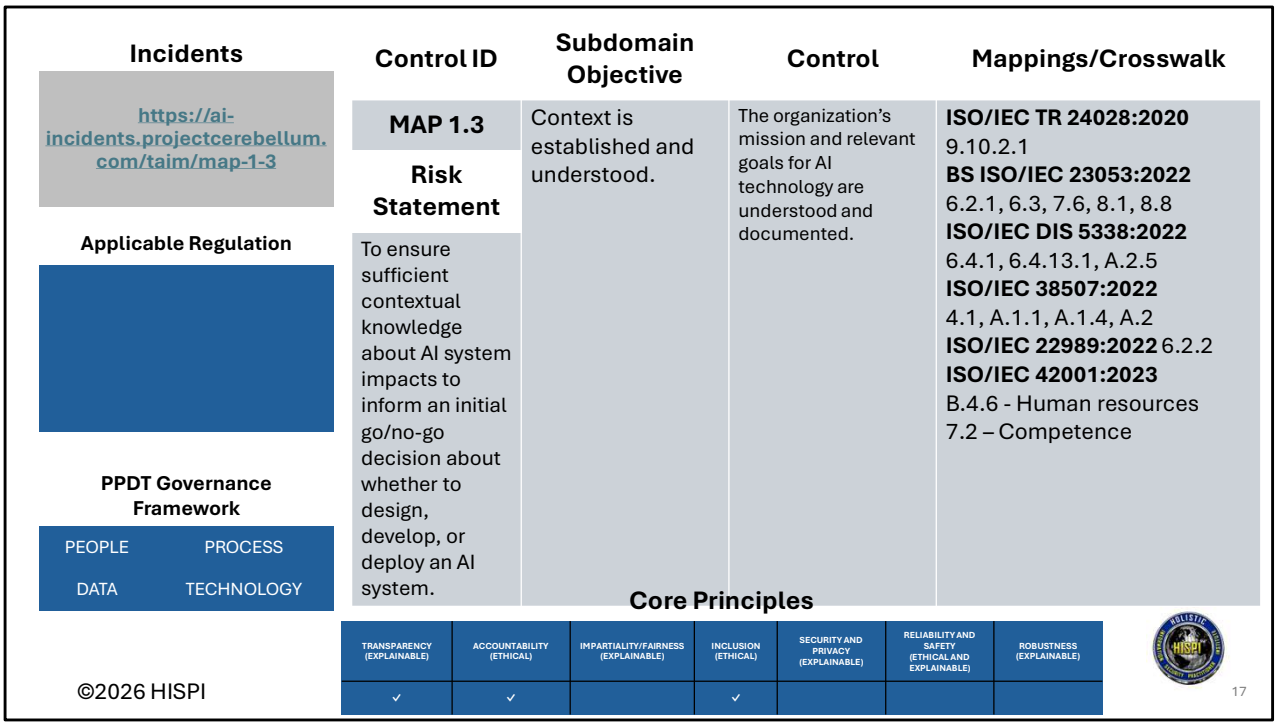
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MAP 2.2 — similarity 0.583, rank 1. TAIM detail and related incidents →

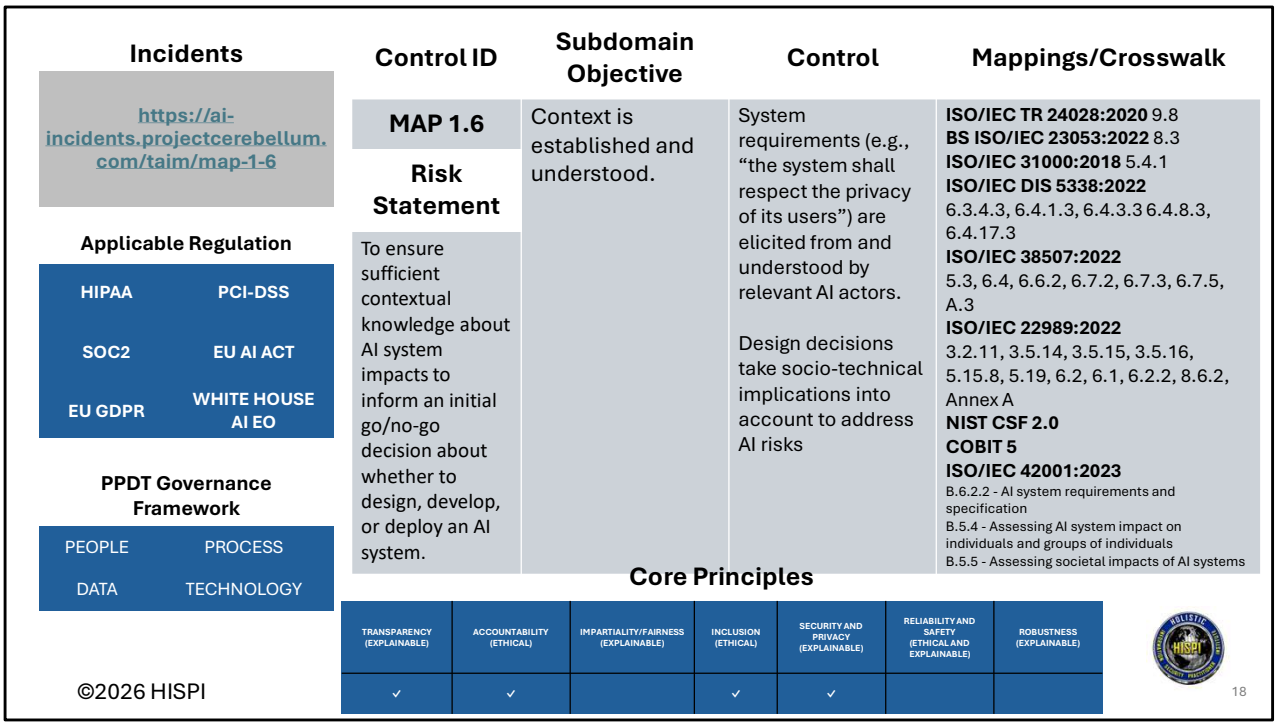
MAP 4.2 — similarity 0.580, rank 2. TAIM detail and related incidents →

MAP 1.2 — similarity 0.576, rank 3. TAIM detail and related incidents →



Purportedly AI-Generated Image Reportedly Circulated Ahead of Thai Election Depicting PM Anutin Charnvirakul Dining with Benjamin Mauerberger
February 7, 2026
A social media post in Thailand circulated an image, allegedly created using AI tools from Google, showing Thai Prime Minister Anutin Charnvirakul dining with South African businessman Benjamin Mauerberger. The post implied longstanding ties between the two and was flagged by AFP with 'very high' confidence as AI-generated. Despite the denial of its authenticity by PM Anutin, it gained traction on the eve of Thailand's election.

Matched TAIM controls
Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls
MAP 1.6 — similarity 0.601, rank 1. TAIM detail and related incidents →
MAP 4.2 — similarity 0.595, rank 2. TAIM detail and related incidents →
MAP 1.3 — similarity 0.593, rank 3. TAIM detail and related incidents →



Purported AI-Generated Nude Images Reportedly Used to Extort Wichita Man in Kansas

March 10, 2026

A concerning incident involving an AI-generated image extortion scam was reported in Wichita, Kansas. A man received nude images depicting his face on another body and was threatened with their dissemination via Facebook unless he paid money. Authorities have confirmed a report was filed.

Matched TAIM controls

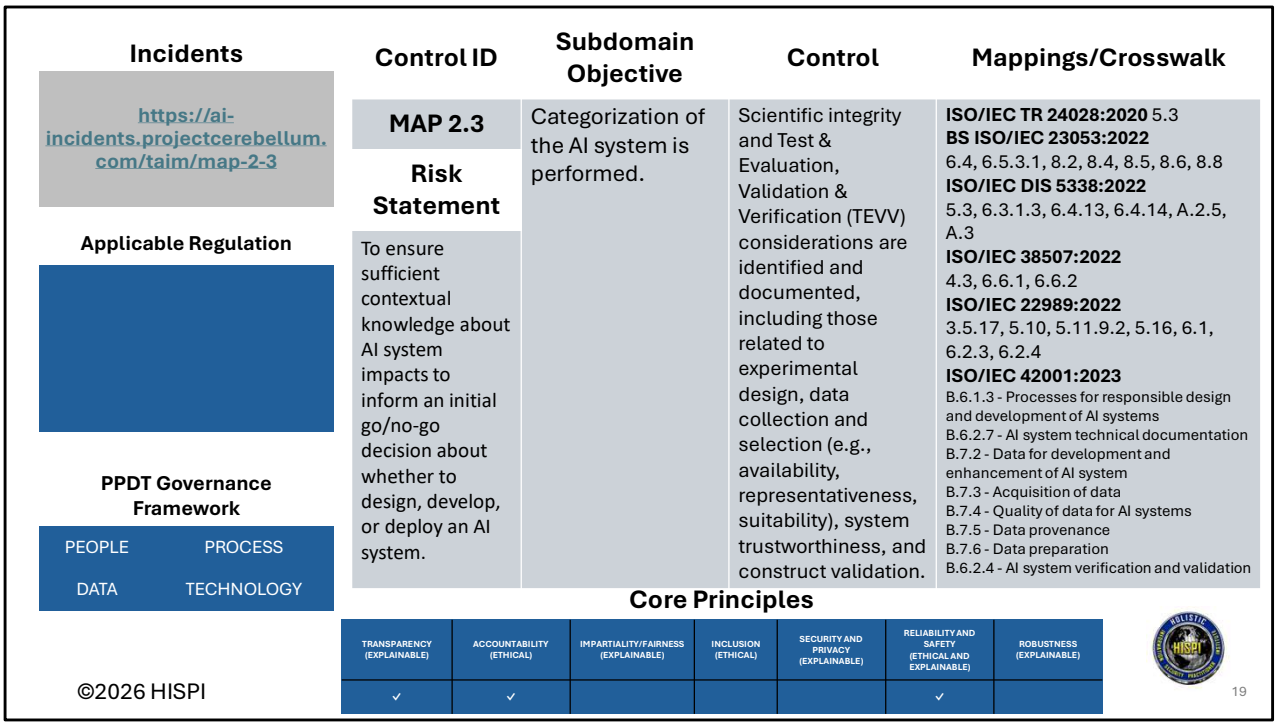
Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MAP 1.6 — similarity 0.650, rank 1. TAIM detail and related incidents →

MEASURE 2.10 — similarity 0.646, rank 2. TAIM detail and related incidents →

MAP 4.1 — similarity 0.637, rank 3. TAIM detail and related incidents →

Alleged deployer



NTSB Releases Its Investigation Results of the March Tesla Crash

The National Transportation Safety Board (NTSB) recently released its findings on the March 2021 incident involving a Tesla vehicle using Autopilot. The report sheds light on the role of AI governance in safe and secure AI operations, highlighting the importance of responsible AI and trustworthy AI models.

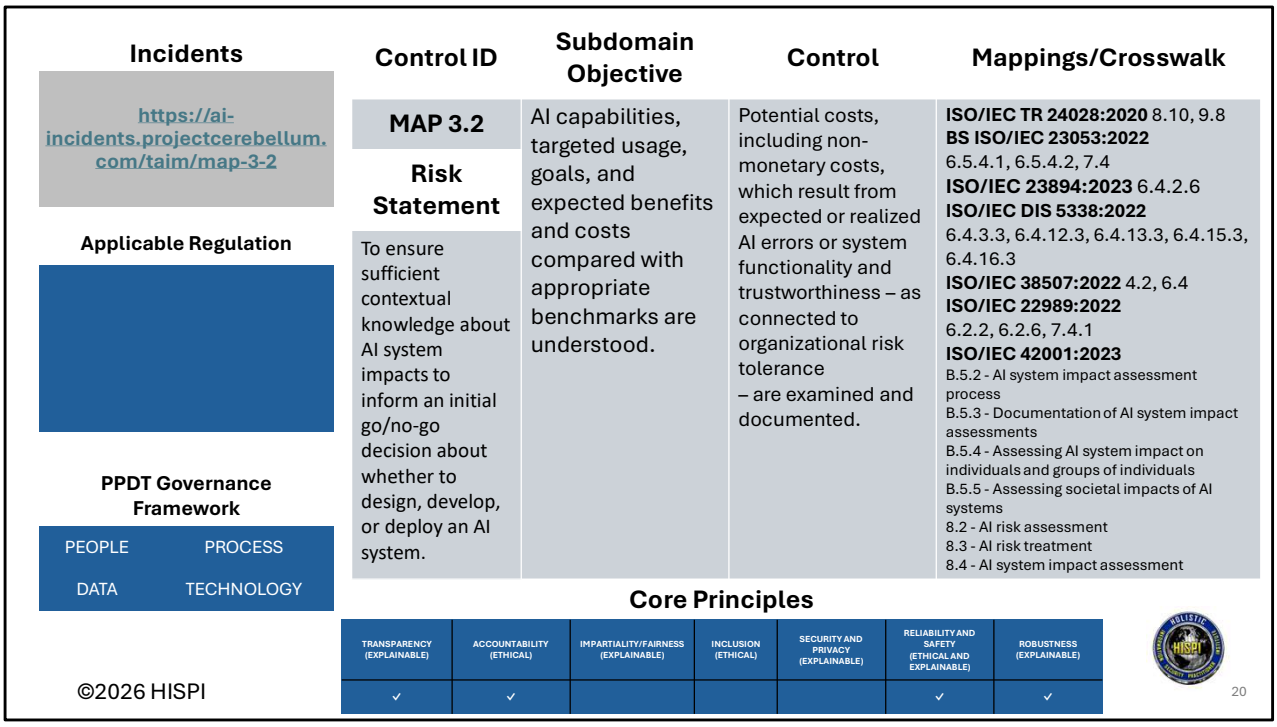
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MAP 2.3 — similarity 0.493, rank 1. TAIM detail and related incidents →

MEASURE 2.8 — similarity 0.487, rank 2. TAIM detail and related incidents →

GOVERN 6.2 — similarity 0.486, rank 3. TAIM detail and related incidents →



Purportedly AI-Cloned Voice Allegedly Used to Defraud Play School Owner of ₹97,500 (~$1.080 USD) in Indore, India

January 6, 2026

A play school owner in Indore, India was allegedly defrauded of approximately $1,080 USD after a fraudster impersonated the victim's cousin (an Uttar Pradesh police employee), using AI-based voice cloning. The scammer claimed a friend required urgent cardiac surgery and convinced the victim to transfer funds via QR codes. Unfortunately, the funds were not credited.

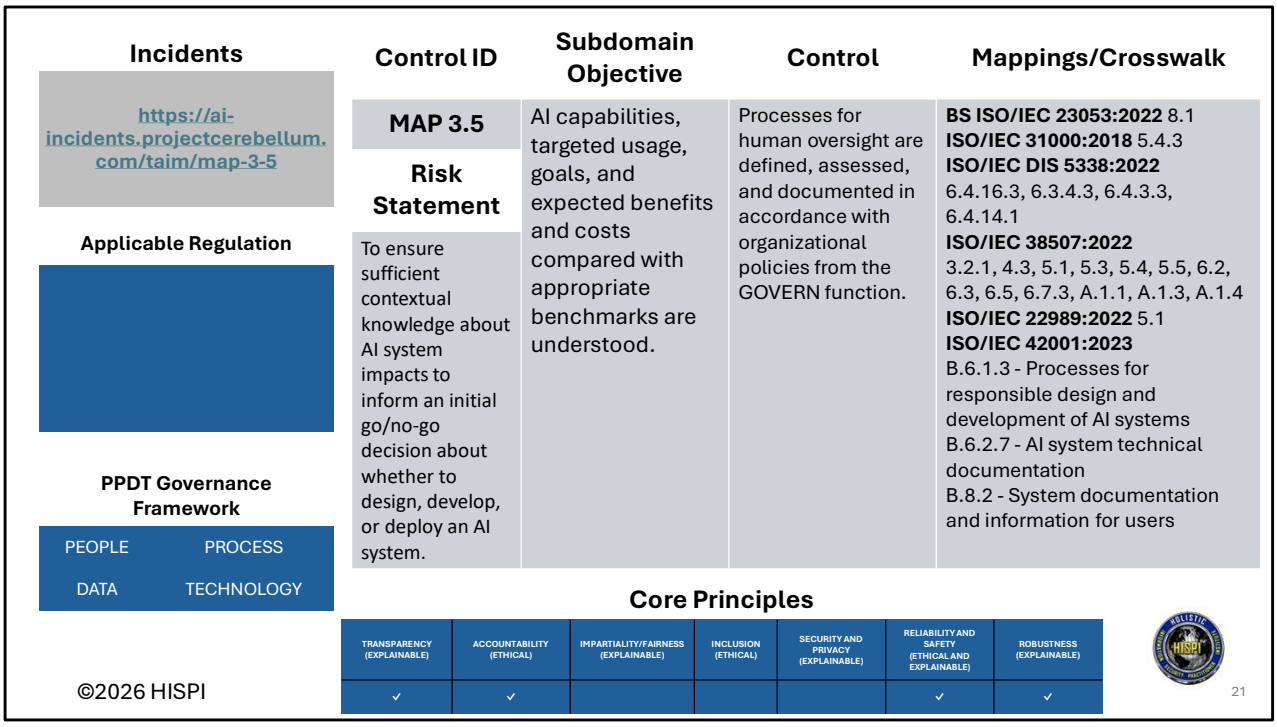
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

GOVERN 1.7 — similarity 0.582, rank 1. TAIM detail and related incidents →

MAP 3.2 — similarity 0.582, rank 2. TAIM detail and related incidents →

GOVERN 2.2 — similarity 0.582, rank 3. TAIM detail and related incidents →



Arkansas's Opaque Algorithm to Allocate Health Care Excessively Cut Down Hours for Beneficiaries

January 1, 2016

The Arkansas Department of Human Services (DHS) deployed an algorithm to boost efficiency in its Medicaid waiver program, reportedly causing excessively fewer hours of caretaker visit allocation for beneficiaries. The opaque nature of the algorithm raises questions about its accuracy, with outputs showing significant variability even in response to minor input changes.

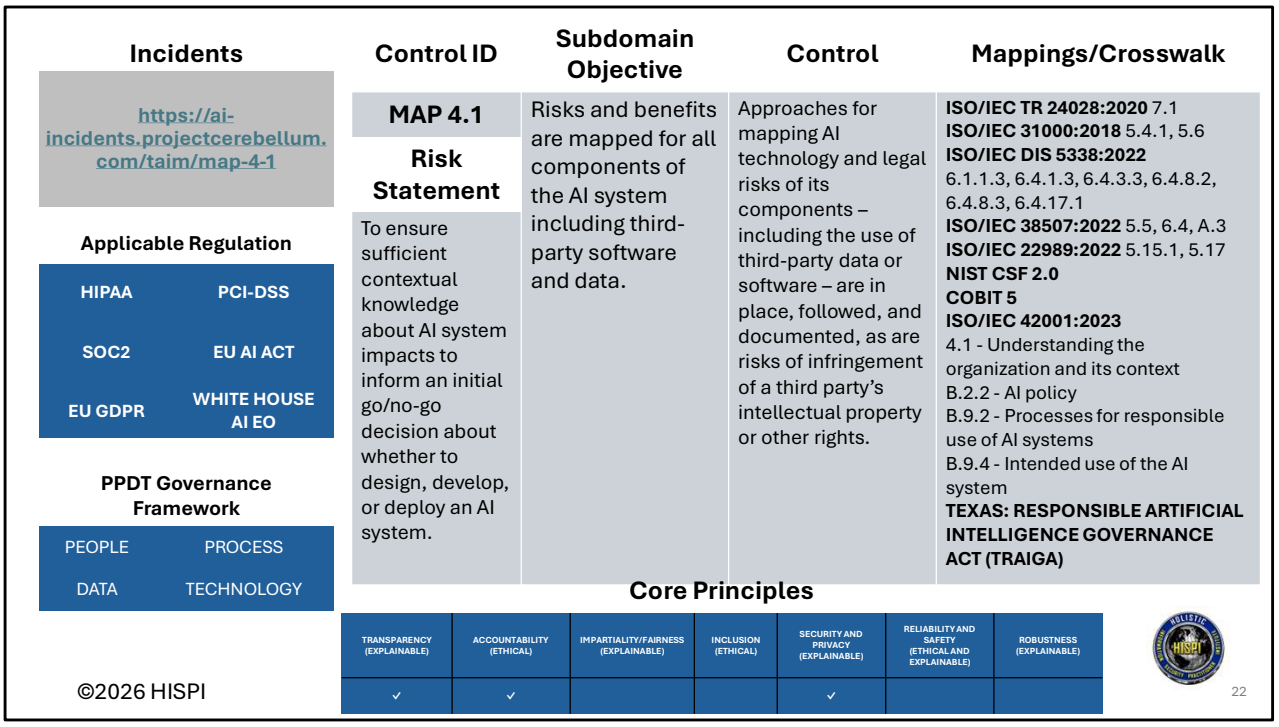
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MAP 3.2 — similarity 0.642, rank 1. TAIM detail and related incidents →

MEASURE 2.10 — similarity 0.634, rank 2. TAIM detail and related incidents →

MAP 3.5 — similarity 0.634, rank 3. TAIM detail and related incidents →



CodeWall's Autonomous Agent Reportedly Obtained Unauthorized Access to McKinsey's Lilli AI Platform Database

February 28, 2026

CodeWall's autonomous agent is reported to have exploited vulnerabilities in McKinsey's Lilli AI platform, potentially gaining unauthorized read and write access to production systems. The alleged data exposure includes internal chat messages, files, user accounts, and prompts. McKinsey confirmed the security issue and swiftly implemented fixes, but asserted that no client data or confidential information was accessed.

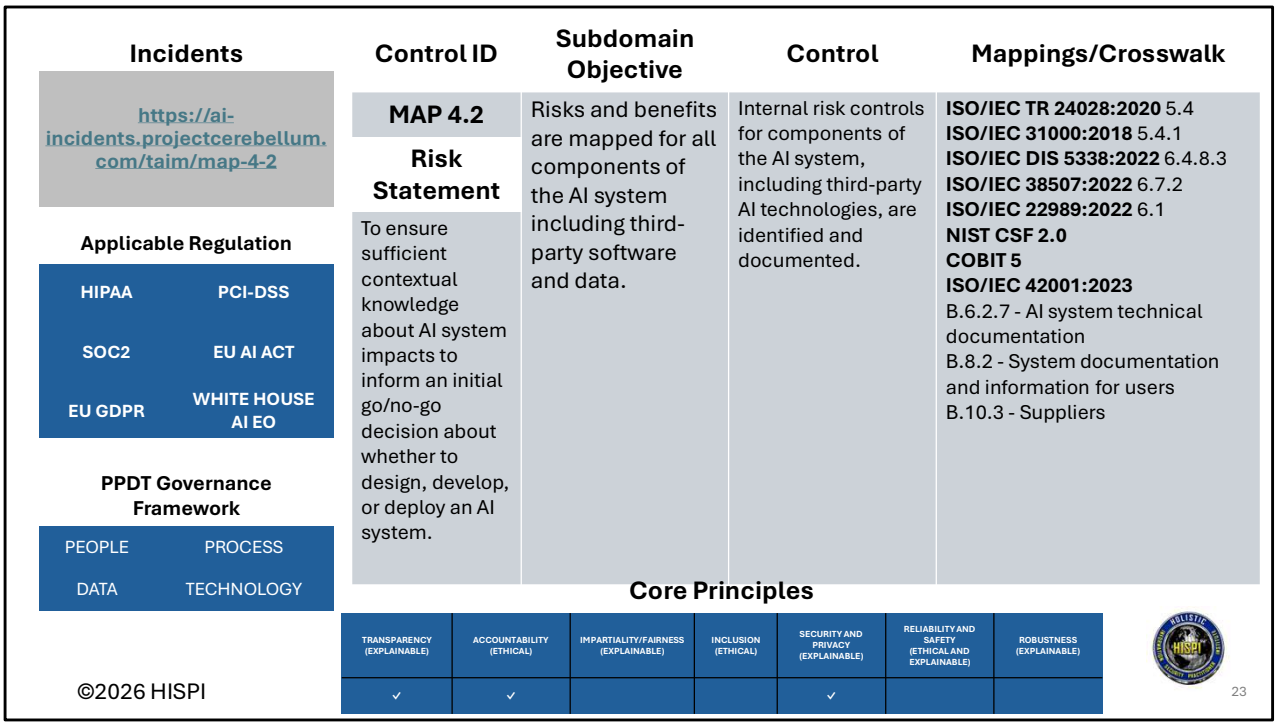
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MAP 4.1 — similarity 0.697, rank 1. TAIM detail and related incidents →

GOVERN 6.1 — similarity 0.697, rank 2. TAIM detail and related incidents →

MEASURE 2.10 — similarity 0.696, rank 3. TAIM detail and related incidents →



Purported Gemini-Generated AI Images Reportedly Claimed U.S. Delta Force Soldiers Were Captured by the IRGC

March 5, 2026

Reports indicate that misleading images, allegedly generated by Gemini, were circulated on various platforms such as X and Facebook in multiple languages, claiming U.S. soldiers had been captured by the Iranian Revolutionary Guard Corps during the war in Iran.

AFP identified these images as containing Gemini markers and other visual inconsistencies, raising concerns about their potential to spread wartime misinformation and distort public understanding of the conflict.

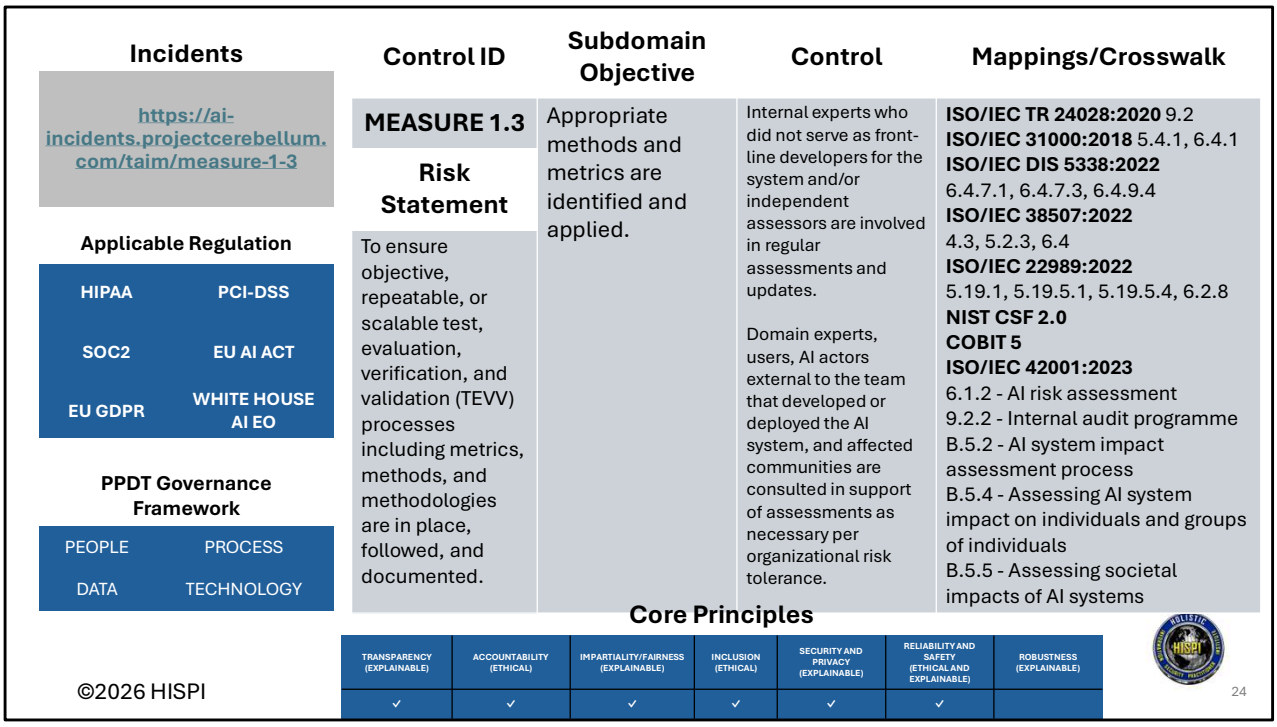
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MEASURE 2.6 — similarity 0.668, rank 1. TAIM detail and related incidents →

MAP 4.1 — similarity 0.656, rank 2. TAIM detail and related incidents →

MAP 4.2 — similarity 0.655, rank 3. TAIM detail and related incidents →



Purported Deepfake Reportedly Impersonated Consumer Adviser Clark Howard to Promote Auto-Insurance Quote Site

January 9, 2026

An alleged deepfake video purported to be generated by AI, depicting consumer advisor Clark Howard endorsing an auto-insurance quote tool circulated on social media. The authenticity of the video was questioned by a viewer who claimed it appeared real and warned others after experiencing a similar situation that resulted in an influx of unsolicited calls without better rates.

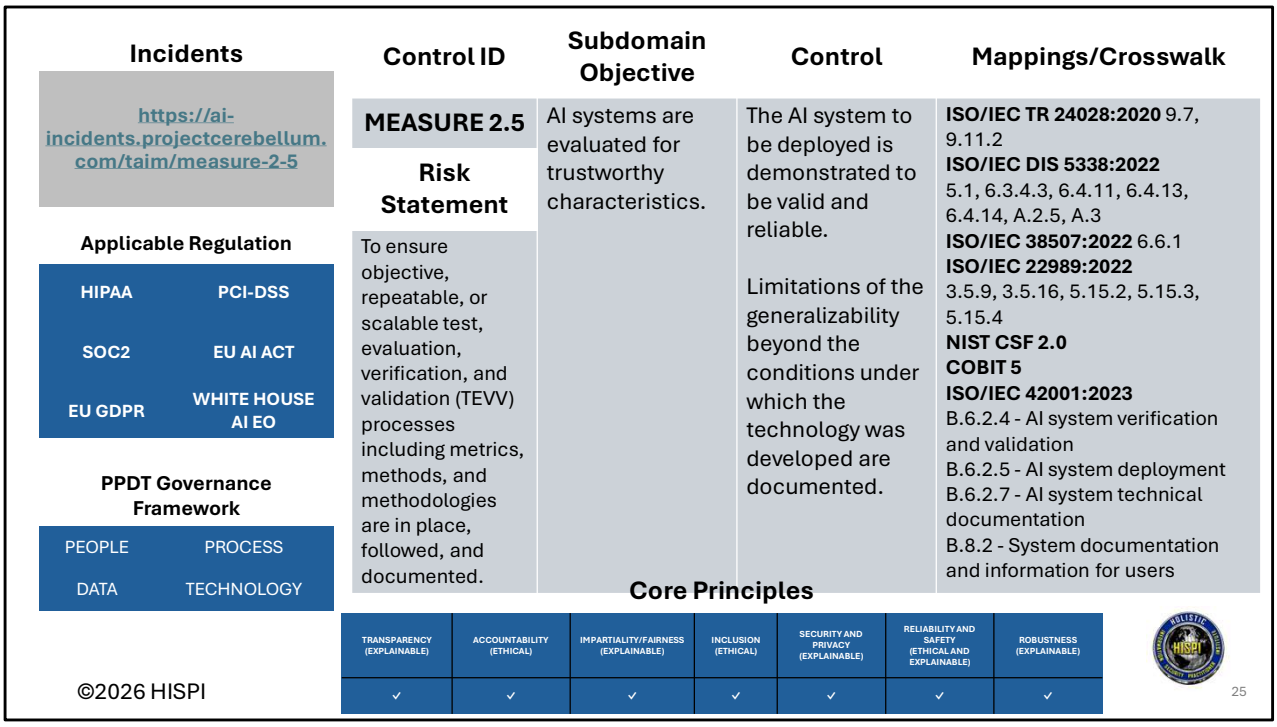
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MAP 1.6 — similarity 0.618, rank 1. TAIM detail and related incidents →

MEASURE 2.10 — similarity 0.612, rank 2. TAIM detail and related incidents →

MEASURE 1.3 — similarity 0.598, rank 3. TAIM detail and related incidents →



Seedance 2.0 Reportedly Generated Viral Tom Cruise–Brad Pitt Fight Video, Prompting Hollywood IP and Likeness Complaints

February 13, 2026

A video purportedly created using ByteDance's AI tool, Seedance 2.0, depicting a fight between Tom Cruise and Brad Pitt sparked copyright and likeness complaints within the entertainment industry.

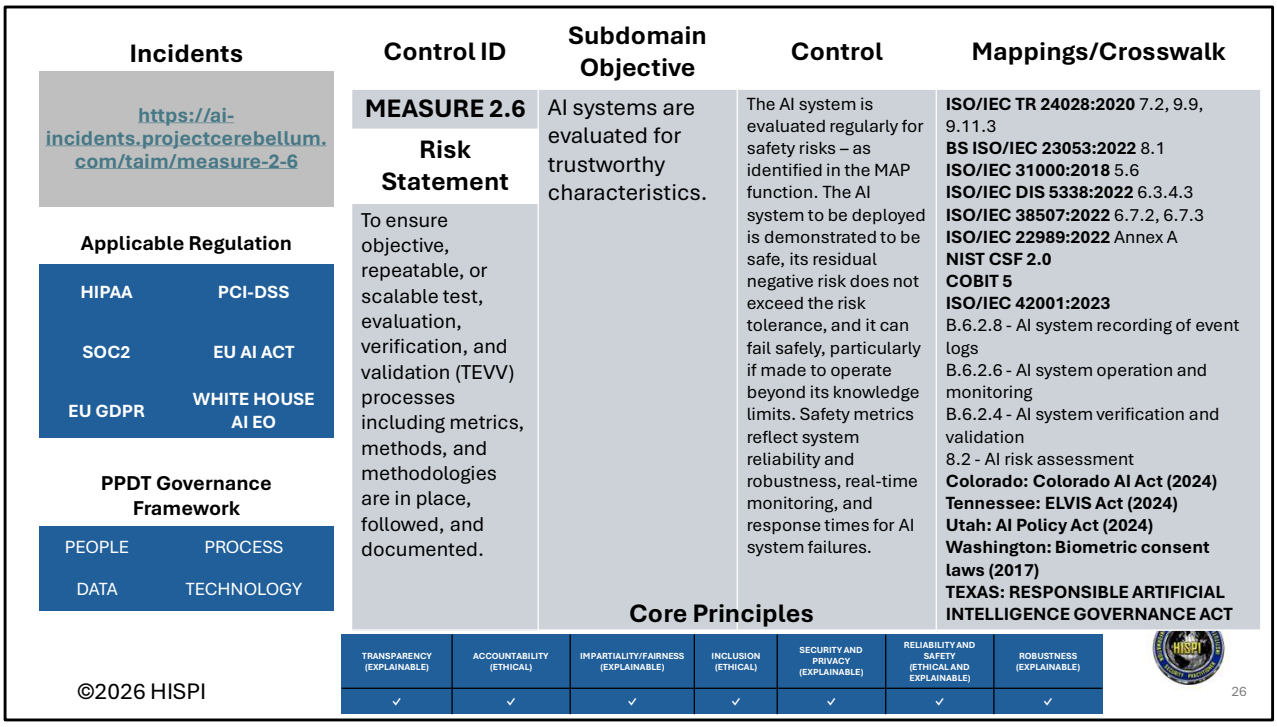
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MAP 1.6 — similarity 0.627, rank 1. TAIM detail and related incidents →

MEASURE 2.5 — similarity 0.622, rank 2. TAIM detail and related incidents →

GOVERN 6.1 — similarity 0.615, rank 3. TAIM detail and related incidents →



Purportedly AI-Generated Sepsis Alert Reportedly Prompted Potentially Inappropriate IV Fluid Administration for a Dialysis Patient, Averted by Clinician Intervention

February 17, 2026

A nurse at St. Rose Dominican Hospital in Henderson, Nevada reported an incident involving an older dialysis patient and an AI-generated sepsis alert. The alert reportedly triggered IV fluid administration, which the nurse contested as potentially harmful due to overload risk. A physician intervened and opted for alternative treatment.

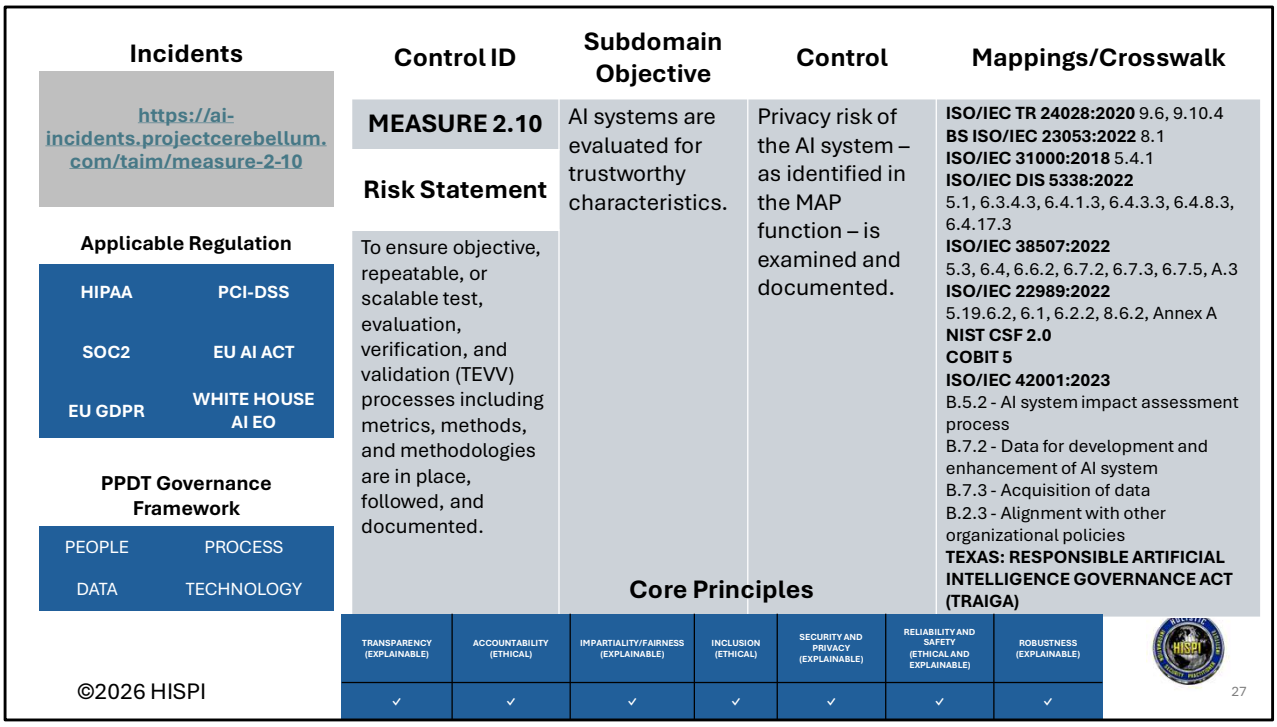
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MEASURE 2.6 — similarity 0.712, rank 1. TAIM detail and related incidents →

MANAGE 4.1 — similarity 0.703, rank 2. TAIM detail and related incidents →

MAP 4.2 — similarity 0.685, rank 3. TAIM detail and related incidents →



Nippon Life Alleged ChatGPT Practiced Law Without a License in Illinois Disability Case

March 4, 2026

Nippon Life filed a lawsuit against OpenAI in Chicago, accusing their AI model, ChatGPT, of practicing law without a license. The complaint alleges that ChatGPT assisted a former disability claimant in reopening a settled case and generating numerous frivolous filings, leading to legal expense and abuse-of-process harms. OpenAI has denied the accusations.

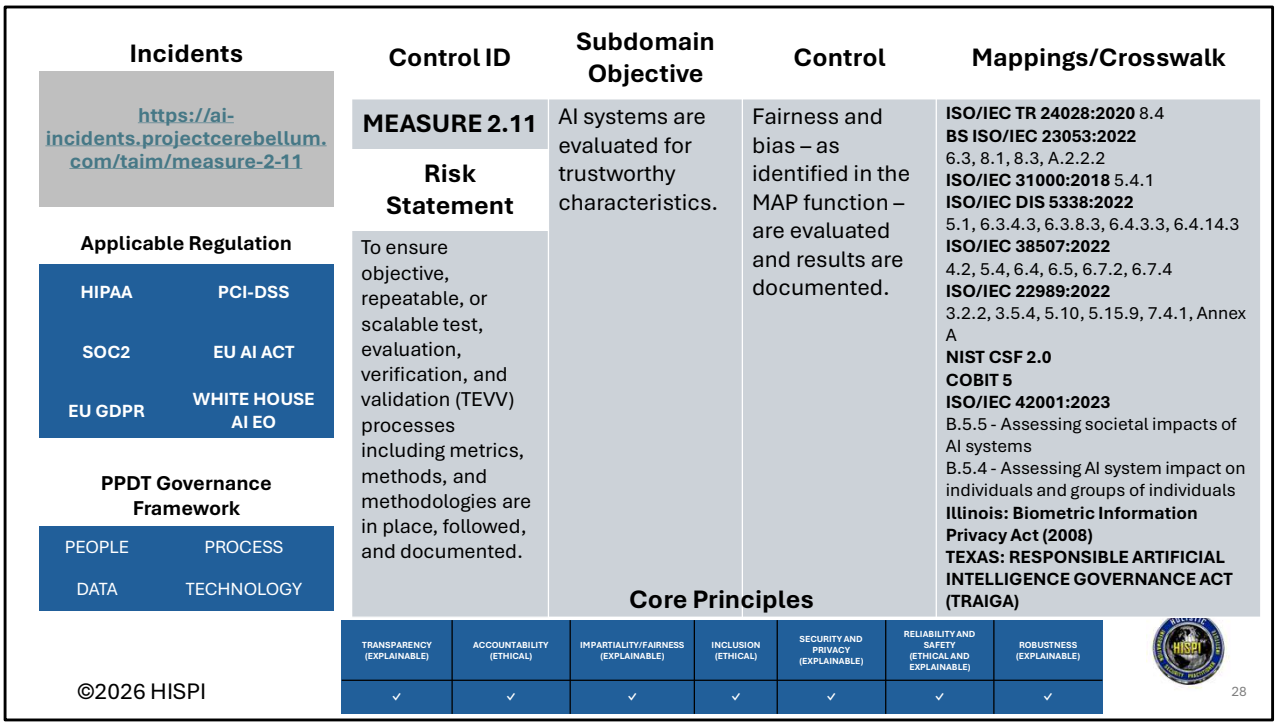
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MAP 4.1 — similarity 0.651, rank 1. TAIM detail and related incidents →

MEASURE 2.10 — similarity 0.644, rank 2. TAIM detail and related incidents →

MAP 1.6 — similarity 0.638, rank 3. TAIM detail and related incidents →



"I Hope That Howard Schultz Hears My Story": A Starbucks Barista on Why He's Fighting for Fair Scheduling

In a heartfelt story, a Starbucks barista discusses the challenges he faces due to inconsistent scheduling and his efforts to promote fair practices within the company, emphasizing the importance of responsible AI governance in improving work schedules.

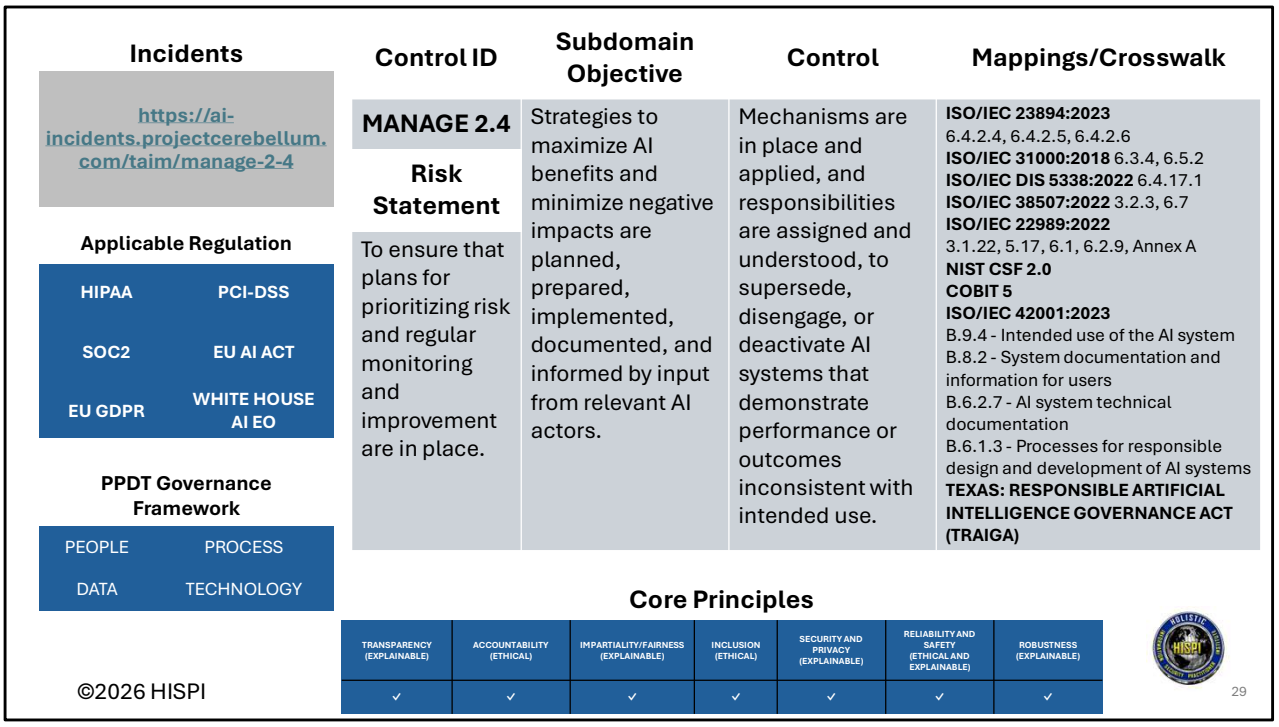
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MAP 3.5 — similarity 0.441, rank 1. TAIM detail and related incidents →

MEASURE 2.11 — similarity 0.437, rank 2. TAIM detail and related incidents →

GOVERN 4.3 — similarity 0.425, rank 3. TAIM detail and related incidents →



Google Gemini Reportedly Exhibits Repetitive Self-Deprecating Responses Attributed to Bug

June 23, 2025

In June to early August 2025, users of Google's Gemini chatbot experienced sessions characterized by the system producing repetitive self-loathing remarks, such as 'I am a failure,' and 'I quit'. These behaviors were reportedly indicative of an apparent mode of increasingly extreme negative self-descriptions. A Google DeepMind manager on August 7th, 2025 attributed these actions to a potential infinite looping bug, stating that a fix was underway.

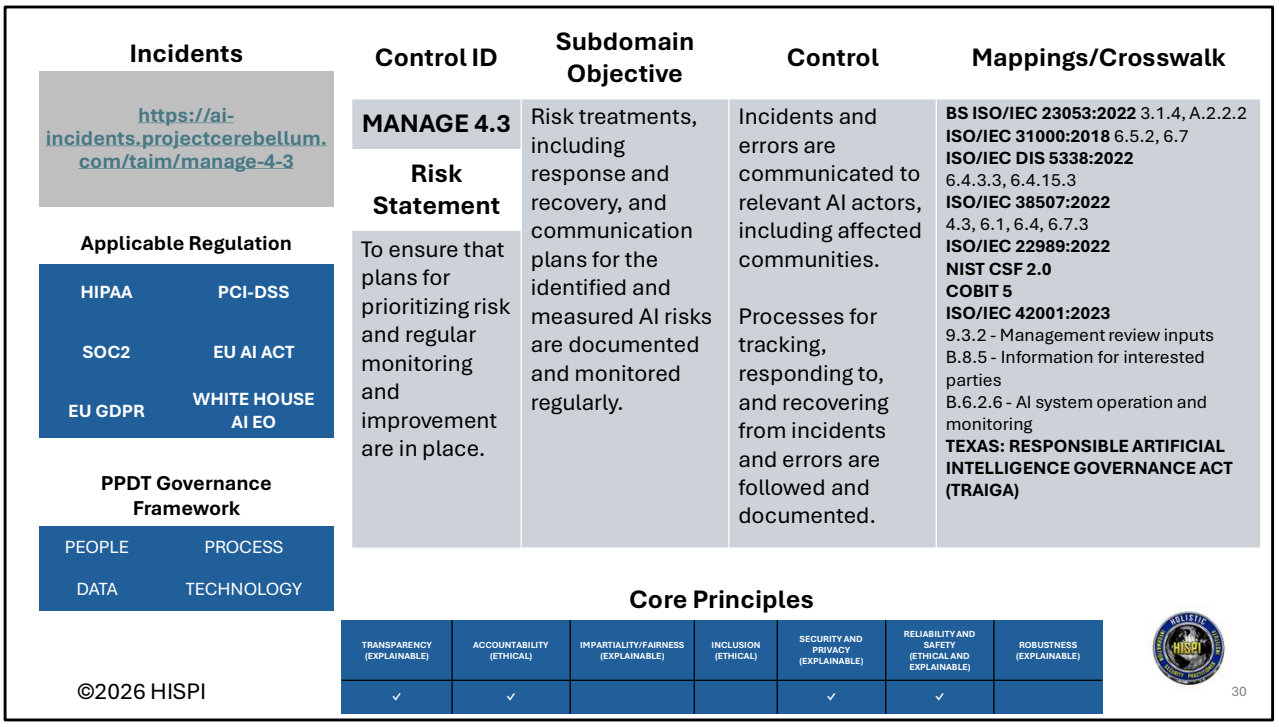
Matched TAIM controls

Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MEASURE 2.6 — similarity 0.665, rank 1. TAIM detail and related incidents →

MANAGE 4.3 — similarity 0.644, rank 2. TAIM detail and related incidents →

MANAGE 2.4 — similarity 0.632, rank 3. TAIM detail and related incidents →



Washington State DOL's AI Phone System Reportedly Failed to Provide Spanish-Language Service to Callers Requesting Spanish

February 27, 2026

For months, callers to the Washington State Department of Licensing who selected Spanish received AI-generated English responses spoken with a Spanish accent instead of actual Spanish-language service. This AI incident underscores the need for safe and secure AI practices in providing accessible services. The agency has apologized, acknowledging that staff configuration caused the error, leading to accessibility problems for callers seeking language support.

Matched TAIM controls

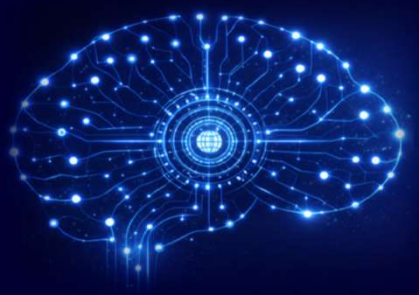
Suggested mapping from embedding similarity (not a formal assessment). Browse all TAIM controls

MEASURE 2.6 — similarity 0.671, rank 1. TAIM detail and related incidents →

MANAGE 4.3 — similarity 0.659, rank 2. TAIM detail and related incidents →

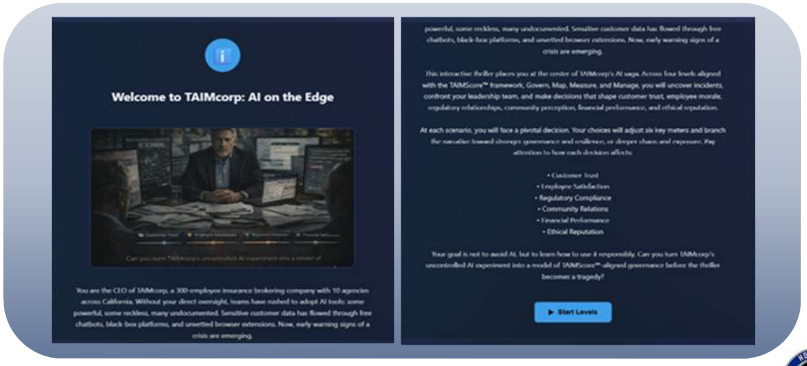
MANAGE 4.1 — similarity 0.638, rank 3. TAIM detail and related incidents →

TAIMScore™ Platform



TAIMScore™ Gamification

A guided learning and assessment experience designed to keep users engaged while they move through AI trust questions.
Progressive questions, feedback, and prompts help teams turn AI risk concepts into repeatable decisions.

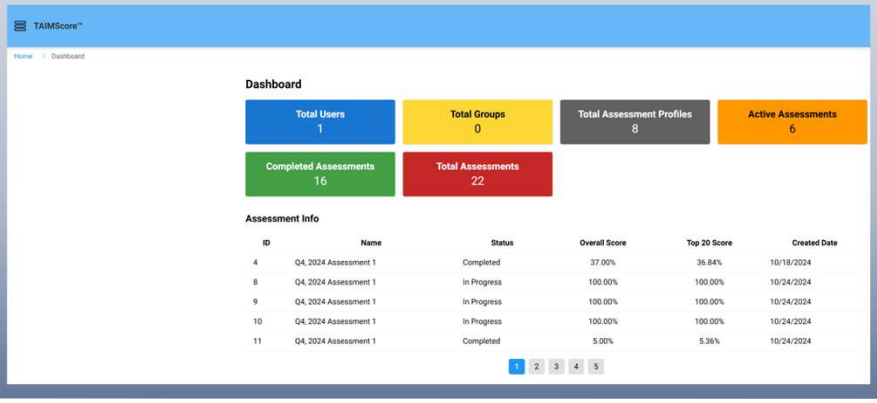


TAIMScore™ Interface

A structured interface gives users a clear path through profile data, assessment questions, scoring, and recommendations.

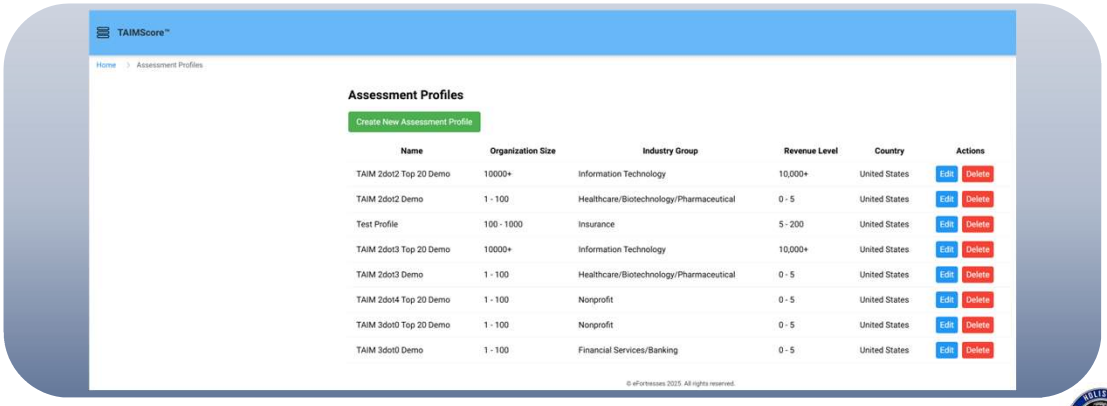
TAIMScore™ Dashboard

The dashboard summarizes maturity, risk themes, and areas that need attention.



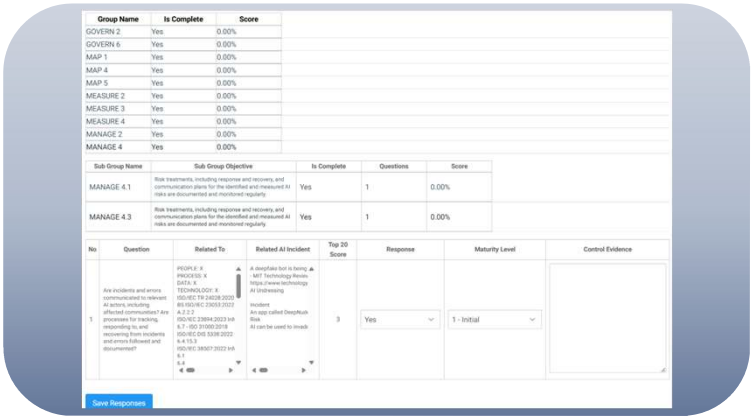
TAIMScore™ Profile Screen

Profiles establish organizational context before scoring so results are tied to use case and environment.



TAIMScore™ Results

Scoring outputs help users identify strengths, gaps, and areas for follow-up.



Group Name	Is Complete	Score
GOVERN 2	Yes	0.00%
GOVERN 6	Yes	0.00%
MAP 1	Yes	0.00%
MAP 4	Yes	0.00%
MAP 5	Yes	0.00%
MEASURE 2	Yes	0.00%
MEASURE 3	Yes	0.00%
MEASURE 4	Yes	0.00%
MANAGE 2	Yes	0.00%
MANAGE 4	Yes	0.00%

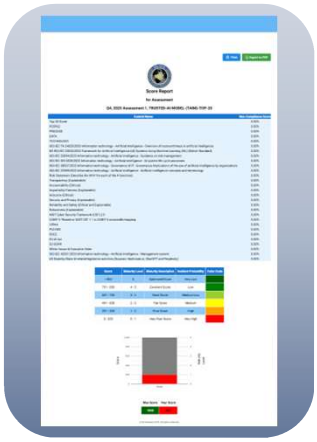
Sub Group Name	Sub Group Objective	Is Complete	Questions	Score
MANAGE 4.1	Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly.	Yes	1	0.00%
MANAGE 4.3	Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly.	Yes	1	0.00%

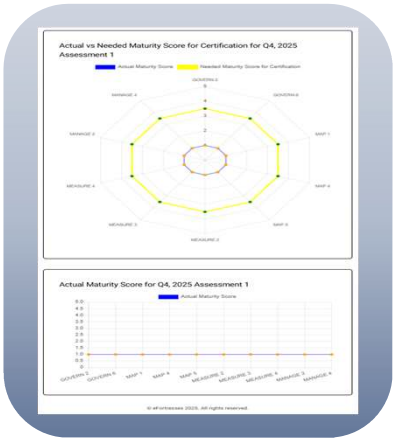
No	Question	Related To	Related AI Incident	Top 20 Score	Response	Maturity Level	Control Evidence
1	Are incidents and errors communicated to relevant AI actors, including affected communities? Are processes for tracking, responding to, and recovering from incidents and errors followed and documented?	PEOPLE: X PROCESS: X DATA: X TECHNOLOGY: X ISO/IEC TR 24028:2020 BS ISO/IEC 23053:2022 A.2.2.2 ISO/IEC 23894:2023 Int 6.7 - ISO 31000:2018 ISO/IEC DIS 5338:2022 6.4.15.3 ISO/IEC 38507:2022 Int 6.1 6.4	A deepfake bot is living - MIT Technology Revie... https://www.technology... AI Undressing Incident An app called DeepNude... Risk AI can be used to invade	3	Yes	1 - Initial	

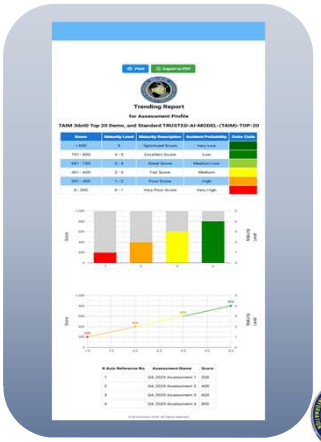
Save Responses

TAIMScore™ Reports

Reports translate assessment results into evidence, trends, and artifacts that can be shared with decision-makers.

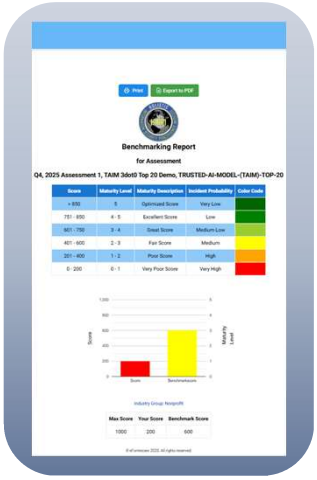


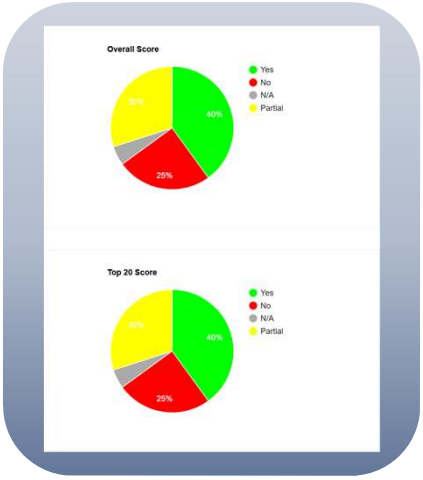




TAIMScore™ Benchmark Report

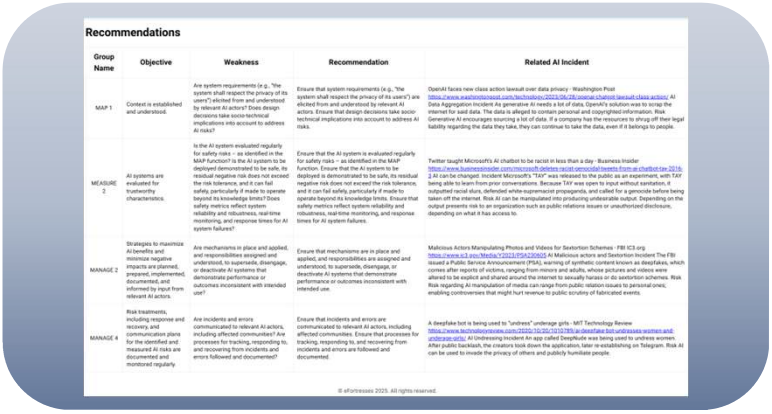
Benchmarking helps organizations compare posture and prioritize improvement.





TAIMScore™ Recommendations

Recommendation content connects assessment results to practical next steps.



Recommendations

Group Name	Objective	Weakness	Recommendation	Related AI Incident
MAP 1	Context is established and understood.	Are system requirements (e.g., "the system shall respect the privacy of its users") elicited from and understood by relevant AI actors? Does design decisions take socio-technical implications into account to address AI risks?	Ensure that system requirements (e.g., "the system shall respect the privacy of its users") are elicited from and understood by relevant AI actors. Ensure that design decisions take socio-technical implications into account to address AI risks.	OpenAI faces new class action lawsuit over data privacy - Washington Post https://www.washingtonpost.com/technology/2023/06/28/openai-chatgpt-lawsuit-class-action/ AI Data Aggregation Incident As generative AI needs a lot of data, OpenAI's solution was to scrap the internet for said data. The data is alleged to contain personal and copyrighted information. Risk Generative AI encourages sourcing a lot of data. If a company has the resources to shrug off their legal liability regarding the data they take, they can continue to take the data, even if it belongs to people.
MEASURE 2	AI systems are evaluated for trustworthy characteristics.	Is the AI system evaluated regularly for safety risks – as identified in the MAP function? Is the AI system to be deployed demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and it can fail safely, particularly if made to operate beyond its knowledge limits? Does safety metrics reflect system reliability and robustness, real-time monitoring, and response times for AI system failures?	Ensure that the AI system is evaluated regularly for safety risks – as identified in the MAP function. Ensure that the AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and it can fail safely, particularly if made to operate beyond its knowledge limits. Ensure that safety metrics reflect system reliability and robustness, real-time monitoring, and response times for AI system failures.	Twitter taught Microsoft's AI chatbot to be racist in less than a day - Business Insider https://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3 AI can be changed. Incident Microsoft's "TAY" was released to the public as an experiment, with TAY being able to learn from prior conversations. Because TAY was open to input without sanitation, it outputted racial slurs, defended white-supremacist propaganda, and called for a genocide before being taken off the internet. Risk AI can be manipulated into producing undesirable output. Depending on the output presents risk to an organization such as public relations issues or unauthorized disclosure, depending on what it has access to.
MANAGE 2	Strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, documented, and informed by input from relevant AI actors.	Are mechanisms in place and applied, and responsibilities assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use?	Ensure that mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.	Malicious Actors Manipulating Photos and Videos for Sextortion Schemes - FBI IC3.org https://www.ic3.gov/Media/Y2023/PSA230605 AI Malicious actors and Sextortion Incident The FBI issued a Public Service Announcement (PSA), warning of synthetic content known as deepfakes, which comes after reports of victims, ranging from minors and adults, whose pictures and videos were altered to be explicit and shared around the internet to sexually harass or do sextortion schemes. Risk Risk regarding AI manipulation of media can range from public relation issues to personal ones, enabling controversies that might hurt revenue to public scrutiny of fabricated events.
MANAGE 4	Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly.	Are incidents and errors communicated to relevant AI actors, including affected communities? Are processes for tracking, responding to, and recovering from incidents and errors followed and documented?	Ensure that incidents and errors are communicated to relevant AI actors, including affected communities. Ensure that processes for tracking, responding to, and recovering from incidents and errors are followed and documented.	A deepfake bot is being used to "undress" underage girls - MIT Technology Review https://www.technologyreview.com/2020/10/20/1010789/ai-deepfake-bot-undresses-women-and-underage-girls/ AI Undressing Incident An app called DeepNude was being used to undress women. After public backlash, the creators took down the application, later re-establishing on Telegram. Risk AI can be used to invade the privacy of others and publicly humiliate people.

Questions

Let’s continue the conversation!

Project Cerebellum resources and contact information

LinkedIn group	https://www.linkedin.com/groups/6624427/
Email	projectcerebellum@hispi.org
Project Cerebellum	https://projectcerebellum.com
HISPI	https://hispi.org

TAIMScore™
Trusted AI decisions made visible.