BREAKTHROUGH INSTITUTE

PRESENTED AT PSAM 2026

AI-ASSISTED SAFETY ASSESSMENT

Automation Bias as an Emergent Risk

ADAM STEIN — DIRECTOR OF NUCLEAR ENERGY INNOVATION, THE BREAKTHROUGH INSTITUTE

A HUMAN-ORGANIZATIONAL RISK

Automation bias from AI in safety and regulatory review is a **human-organizational reliability problem** — not a failure of AI models.

It is a **well-documented behavioral phenomenon**: it affects expert practitioners, operates even when systems are highly accurate, and intensifies as repeated reliability builds trust.

Applied here to **regulatory review,** not design.

THE SHIFT

When a reviewer routinely receives AI output **before forming an independent judgment,** the cognitive architecture of review changes — gradually, and hard to detect from within.

MECHANISM

TWO ERROR MODES

OMISSION

The reviewer fails to act because the system issued no prompt. **Silence is read as evidence** — even when it only reflects detection limits or incomplete training data.

AI conformance checks and anomaly screens let issues the system never flagged slip past review.

COMMISSION

The reviewer follows incorrect AI guidance despite contradictory evidence. **The output displaces judgment** rather than supplementing it.

AI findings anchor reviewers toward a conclusion, especially when output looks complete and authoritative.

Both share one mechanism: automation becomes a heuristic replacement for vigilant information-seeking — not a tool subject to critical evaluation.

MECHANISM

HOW RELIANCE COMPOUNDS

Complacency

Monitoring intensity declines when the system has usually been right.

Reviewers may come to treat AI silence as meaningful evidence that no issue exists.

Learned Carelessness

Rare errors train users to accept output with minimal scrutiny.

The apparent efficiency of AI assistance may reduce independent engagement with the underlying record.

Hypothesis Narrowing

Once the system frames the issue, the search for alternatives shrinks.

The organization may preserve review throughput while losing the capability needed for independent error detection.

95%

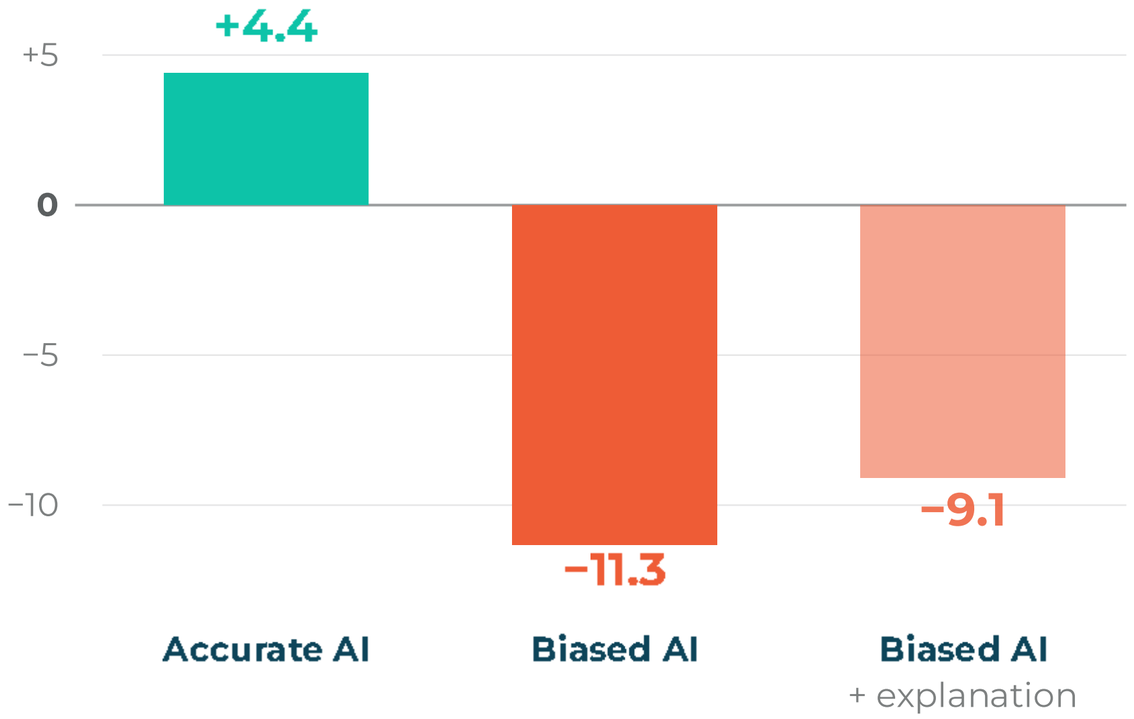
A 95%-accurate review aid produces its **most consequential failures** after users have learned that checking is usually unnecessary.

EVIDENCE · MEDICINE

THE CLEAREST SIGNAL

Change in clinician diagnostic accuracy with AI assistance

Percentage points vs. unaided baseline



Randomized clinical-vignette study (Jabbour et al., 2023)

Accurate AI helps modestly. Biased AI hurts about three times as much — and explanations solve the problem.

−9.1 Adding **explanations** to biased AI still left accuracy well below baseline — it persuaded rather than exposed the error.

57→48 A separate **ECG** study: resident accuracy fell nine points with a manipulated tool.

> The reviewers were **trained experts** — the profile of a regulatory reviewer, not a novice.

EVIDENCE · HIGH-CONSEQUENCE DOMAINS

WHEN THE STAKES ARE FATAL

AVIATION

AF447

Automation mode confusion turned a recoverable failure fatal.

In simulations, pilots followed false alerts they had said they would ignore.

DEFENSE

3 deaths, 2 friendly aircraft (2003)

Patriot batteries fired on automated target classifications.

Training and culture had made human oversight nominal.

WHAT THEY TELL US

- The failure pathway is **behavioral and organizational**, not a model error.
- Stated vigilance is **not real vigilance** under a confident cue.
- What decides the outcome is retained **authority and accountability** to override.

Even experts. Even when the system is accurate. Even when told it is imperfect.

NUCLEAR REGULATORY REVIEW

THE SLOW-DECISION CONTEXT

Regulatory review runs over **months and years** , not seconds. The absence of acute time pressure changes the risk profile — in both directions.

DELIBERATION BUFFERS

+ Multiple review layers create distributed verification
+ Documentation externalizes and audits reasoning
+ No acute time pressure forcing intuitive judgment

BUT NEW RISKS OPEN

- Habituation across months of reliable output
- Cognitive offloading of dense submissions
- Organizational skill atrophy
- Diffusion of responsibility across the team

THE CENTRAL QUESTION

When the AI makes an error, will anyone catch it?

NOT: *"Will the AI make an error?"*

This is a question about skill, mandate, accountability, and independence — not about model performance alone.

EXISTING FRAMEWORKS

THE FOUNDATION

NUREG-0711	Function allocation — what AI does, what humans must retain
NUREG-0700	Interface design — salience, engagement, uncertainty
NUREG-2261	AI readiness and workforce development
IAEA 2025	Explainability and graded, risk-informed deployment

WHERE THEY NEED EXTENSION

- The slow-decision review context
- Long-term behavioral monitoring
- Team accountability design
- Internalized vs. procedural accountability

The discipline exists. It must be turned inward, onto the regulator's own review.

FIVE GOVERNANCE CONTROLS

#	CONTROL	MECHANISM IT ADDRESSES
1	**Independent verification**	Anchoring & commission errors — form a view before seeing the AI
2	**Uncertainty & limits, up front**	Overconfidence and false precision in AI output
3	**Traceability to evidence**	Passive acceptance and inability to audit reasoning
4	**Periodic AI-off review**	Skill atrophy and loss of manual competence
5	**Accountability & monitoring**	Responsibility diffusion, habituation, learned carelessness

THE BOTTOM LINE

BREAKTHROUGH INSTITUTE

The goal is not to slow AI — it is to direct it.

- These errors are **easy to make** and **hard to detect** in process.
- A human signature at the end of an AI workflow is **not the same** as substantive human review of AI-assisted analysis.
- In-the-loop matters only if the human retains the **skill, authority, accountability, and independence** to detect error.
- Controls should be **designed into workflows and processes** early.

Adam Stein

adam@thebreakthrough.org

Full paper: https://www.iapsam.org/PSAM18/papers/AD175-PSAM18.pdf