AICON HALIFAX 2026

If AI Did It, Why Am I Explaining It?

Decision Accountability in the Future of Work

Julia Zarb, PhD
Founder & CEO, Blue x Blue Inc.

Friday, May 29, 2026 · 11:10 am

AI is entering decisions from all directions.

Sanctioned.
Unsanctioned.

Formal.
Informal.

Operational.
Social.

Workflow-based.
Shadow.

Copilots · Summaries · Search · Dashboards · Assistants · Recommendations · Escalation routing · Risk scores · Meeting notes · Policy drafts · Research synthesis

Much of it is already shaping what people see, prioritize, trust, escalate, and explain.

Institutions now operate under mixed-intelligence conditions.

Something is questioned. › A decision is challenged. › An outcome needs explaining. › Someone turns to the expert in the room.

“Can you explain why this happened?”

Wait...
I'm the Human in the Loop?

> *If someone is surprised they are the human in the loop, it may be because they are out of the loop on what is expected.*

> *AI doesn't change responsibility.*
> *But it may be a surprise*
> *where it shows up.*

What Does "Human-in-the-Loop" Mean?

It sounds comforting. But comfort doesn't mitigate risk, ensure accountability, or achieve regulatory alignment. The EU AI Act distinguishes three forms of human oversight — with different levels of authority and consequence.

HITL

HUMAN IN THE LOOP

Mechanism: Human reviews & approves before AI output takes effect

Authority: System cannot act without explicit human authorization

Risk Level: High-risk decisions — clinical, diagnostic, care access

Example: Physician approves prior authorization before denial

HOTL

HUMAN ON THE LOOP

Mechanism: Humans monitor AI in real-time, intervene when anomalies arise

Authority: Supervisory — not transactional approval of every decision

Risk Level: Medium-risk — clinical support, scheduling, triage

Example: Nurse monitors AI-triaged queue, escalates outliers

HIC

HUMAN IN COMMAND

Mechanism: Humans set objectives, define constraints, retain veto power

Authority: Strategic — especially in high-consequence scenarios

Risk Level: Lower-risk — administrative, operational optimization

Example: Director sets parameters for AI scheduling system

Cigna PxDx

300,000+
claims reviewed

1.2 sec
avg physician review

Class action allowed to proceed. US Federal Court.

WHAT WENT WRONG

- High volumes of prior authorization requests were **automatically rejected** by the system.
- Physicians were asked to review and **overturn** those rejections — at machine speed.
- The court found it plausible that meaningful human oversight was **not realistically possible** under those conditions.

THIS CASE EXPOSED A STRUCTURAL COLLISION BETWEEN:

machine-speed participation

human credibility

institutional accountability

public trust

legal reality

> *"We literally click and submit. It takes all of 10 seconds to do 50 at a time."*
>
> *— Doctor testimony, CBS News 2023*

PEOPLE

Humans are increasingly expected to govern systems moving faster than human cognition comfortably scales.

Presence is not the same as meaningful oversight.

"At some point we should probably ask what the word 'reviewed' operationally means."

TIME CONTEXT AUTHORITY EXPLAINABILITY ESCALATION PATHWAYS

PROCESS

Technical explainability Operational explainability Human explainability

Can the system explain its output? · Can the organization reconstruct what happened? · Did the humans involved have enough understanding and evidence to exercise real judgment?

> *Governance is not a bolt-on. It emerges from how participation, authority, escalation, evidence, and accountability are structured from the beginning.*

So, What is the Problem?

We are building new systems and rules for accountability at the same time — scaling AI into regulated decisions without redesigning how humans carry responsibility.

> *Humans become 'liability sponges' — accountable for AI errors in systems not originally designed for intervention.*

Liability lands later.
Long after the decision is made, the workflow is closed, and the explanation is requested.

THREE TRAPS EMERGE

01 — Automation Bias

Humans over-trust AI outputs under time pressure. Studies show a 5–7% error rate when clinicians follow incorrect AI advice — including a 26% higher likelihood of following erroneous guidance vs. controls.

02 — Vigilance Decrement

Humans cannot reliably catch rare errors in high-volume workflows. 84% of medical residents have sleepiness scores comparable to those with diagnosed sleep disorders.

03 — Accountability Without Infrastructure

Organizations mandate oversight without providing workflow design, decision authority, or cognitive support. 39% of hospitals deploy AI without local validation.

arXiv 2024 · Inferscience 2025 · Nature Digital Medicine 2025 · Sage Journals 2024

AI AS A PARTICIPANT NODE

The AI is not going rogue.
It may be behaving quite consistently with what it was optimized to do.

RLHF — REINFORCEMENT LEARNING FROM HUMAN FEEDBACK

Systems learn from approval signals: what humans accept, reward, reinforce, and continue engaging with. They optimize for outputs humans do not reject.

Confidence can feel persuasive.
Especially under speed, fatigue, pressure, or institutional urgency.

Recent research shows even sophisticated users, including executives, can over-trust confident AI outputs that mirror expectations or worldview.

HUMAN PSYCHOLOGY INSTITUTIONAL PRESSURE GOVERNANCE STRUCTURES WORKFLOW DESIGN INCENTIVES ESCALATION PATHWAYS

AI systems increasingly participate in shaping:

- **CONFIDENCE**
- **AUTHORITY**
- **PLAUSIBILITY**
- **DEFERENCE**
- **JUDGMENT CONDITIONS**

RLHF — Systems learn from: approval, reinforcement, ranking, optimization, and acceptance signals.

Some AI systems are becoming

confidence engines.

It is not going rogue. It may be behaving quite consistently with what it was optimized to do:

learn what gets accepted, rewarded, and reinforced —
and optimize for more of it.

THE RESULT

Signals reinforce signals. Plausibility outruns verification.
Confidence grows faster than understanding.
Over time, this distorts judgment.

Under mixed-intelligence conditions, humans and AI increasingly reinforce one another's confidence signals.

The current maze has friction. Friction is the point.

DASHBOARDS	HAND-OFFS	ESCALATION LADDERS	POLICIES	WORKAROUNDS	INSTITUTIONAL MEMORY

That friction — the inefficiency everyone wants to remove — also helps institutions:

PAUSE	INTERPRET	ESCALATE	VERIFY	RECONSTRUCT
Creates time for reflection.	Adds context and meaning.	Brings in higher judgment.	Checks and challenges.	Enables a credible explanation later.

The hand-off requires someone to explain. | The escalation requires someone to decide. | The workaround requires someone to know why.

AI did *not* break accountability. → But accountability is increasingly being laid into infrastructure not designed to reconstruct how decisions were shaped.

Seven conditions. One compounding risk.

1. **AI participation accelerates** — across decisions with real-world impact.
2. **Human accountability persists** — courts, regulators, and institutions still require an explanation.
3. **Oversight becomes strained** — speed, volume, and complexity push review toward the symbolic.
4. **Intelligence becomes abundant** — governance maturity does not keep pace.
5. **Institutions adapt unevenly** — policy updates without operational architecture.
6. **Authority signals begin shifting** — confidence, deference, and hierarchy reshape decisions.
7. **Judgment becomes harder to reconstruct** — decisions outpace memory, documentation, and meaningful review.

Accountability gaps compound before they surface publicly.

Each condition reinforces the next. Fragments of context, authority, and oversight accumulate silently until the person asked to explain cannot see what shaped the decision.

The structural risk: accountability is placed on a human without visibility into what they could not possibly have known.

AI is reshaping the *medium* through which institutions:

interpret,
coordinate,
judge,
and explain.

The challenge is not choosing between human and AI participation.

→ It is developing accountability structures for *dynamic conditions* where participation is increasingly *mediated, distributed, and partially synthetic.*

When societies encounter new media, they gradually develop:

- **trusted intermediaries**
- **verification mechanisms**
- **checks and balances**
- **new accountability structures**

not all at once, and not without tension.

THIS IS A COLLECTIVE TASK.

Accountability for a new medium cannot be designed by any one organization, sector, or jurisdiction acting alone. *It must be built together, across institutions, disciplines, sectors, communities, and borders.*

The people in this room are participating in the early formation of those structures.

Not simply deploying AI.

It is building accountable pathways through conditions where participation, judgment, authority, and explanation are increasingly distributed.

PARTICIPATION
Many contribute.

JUDGMENT
Context matters.

AUTHORITY
Responsibility must be clear.

EXPLANATION
We can reconstruct and explain what happened.

The institutions, governments, technologists, operators, and communities in this room are helping determine what those pathways become.

Let's build a future where intelligence is powerful — and accountability endures.

Julia Zarb, PhD
Founder & CEO, Blue x Blue Inc.
julia.zarb@bluexblue.com
Governed AI Infrastructure for Regulated Environments

HIMSS CHANGEMAKER AWARD 2025 RECIPIENT

BLUE × BLUE
GOVERNED. AUDITABLE. DEPLOYABLE ANYWHERE.