AIXPERT

Accountability by Design

From Post-Hoc Audit to Architectural Traceability
in AI-Enabled Public Services

Shaina Raza, PhD

Vector Institute / AIXPERT (EU Horizon)

Toronto, Canada

April 2026

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

www.aixpert-project.eu

The Accountability Gap in Public-Sector AI

01 Opaque Decisions

Citizens receive AI-influenced outcomes - benefit denials, triage scores, transit routing - with no accessible rationale for the specific decision made about them.

02 Retroactive Audits

Accountability happens after harm. Post-hoc audits cannot reconstruct multi-step reasoning chains, especially in agentic systems with tool calls and intermediate inferences.

03 Compliance ≠ Trust

Legal compliance satisfies regulators but doesn't answer the citizen's fundamental question: "Why was this decision made about me, and how can I contest it?"

The Necessary Reframe

FROM

"Is this output correct?"

"Is this system compliant?"

→

TO

"Is this interaction trustworthy?"

Trace

Every reasoning step logged in real time

Contest

Citizens can interrogate any decision

Intervene

Humans can override at any point

Accountability is not a post-deployment checkbox : it is a property of how the system reasons, acts, and explains itself.

Transparency in Agentic AI: Three Pillars

Interpretability

Can we see inside?

Understanding the internal reasoning mechanisms of agentic AI such as attention patterns, tool selection logic, planning traces, and intermediate representations.

Explainability

Can we communicate it?

Translating internal reasoning into stakeholder-appropriate explanations , for citizens, clinicians, case workers, and auditors, not just ML engineers.

Governance

Can we enforce boundaries?

Runtime mechanisms to constrain autonomous actions ,such as tool call permissions, data access scoping, escalation triggers, and human override protocols

https://vectorinstitute.github.io/Agentic-Transparency/

Scenario: AI-Assisted Healthcare Triage

WITHOUT Trace Accountability

1. Patient submits symptoms via portal
2. AI assigns triage priority: "Low"
3. Patient waits 6+ hours
4. Condition worsens - no record of why AI scored "Low"
5. Post-hoc audit finds model used postcode as a proxy for acuity

WITH Trace Accountability

1. Patient submits symptoms via portal
2. RRI flags: reasoning step weighted postcode → low equity score
3. CoT-RS: chain inconsistency detected - symptom severity contradicts triage output
4. AGR: model queried demographic data outside sanctioned scope
5. System escalates to human clinician before assigning priority

Trace-level metrics catch bias, inconsistency, and scope violations BEFORE harm but not after.

Mapping to the EU AI Act

EU AI Act Requirement	What It Demands
Art. 13 - Transparency	Users must understand AI system outputs and interpret them appropriately
Art. 14 - Human Oversight	Humans must be able to understand, monitor, and override AI decisions
Art. 9 - Risk Management	Continuous identification and mitigation of risks throughout lifecycle
Art. 10 - Data Governance	Training and operational data must be relevant, representative, and free of bias
Art. 15 - Robustness	Systems must perform consistently and handle errors or inconsistencies

Trace-level metrics operationalise EU AI Act obligations as runtime evidence, not just documentation.

What This Means for Public Services

For System Designers

Embed trace logging and metric evaluation into your architecture from day one - not as a compliance layer, but as core infrastructure. Think fire exits, not fire insurance.

For Policymakers

The EU AI Act mandates transparency and oversight. Trace-level metrics give you something concrete to audit: runtime evidence of how decision #47,382 was actually made.

For Citizens

The goal is contestable AI. Every automated decision should come with a reasoning trace a non-expert can interrogate - not a 200-page model card, but a clear answer to 'why me?'

"We should not ask whether AI systems are trustworthy.
We should ask whether their reasoning is."

shaina.raza@vectorinstitute.ai
Open for discussion

Our Consortium

ATHENA RESEARCH CENTER

Amsterdam UMC

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

cnrs SCIENCES INFORMATIQUES

Furhat Robotics

INFINITIVITY DESIGN LABS

itml
innovation applied

KYKLOS

MARTEL innovate

novelcore
RESEARCH & INNOVATION

ORFIUM

PHILIPS

university of groningen

SORBONNE UNIVERSITÉ

UNIVERSITAT DE BARCELONA

VECTOR INSTITUTE | INSTITUT VECTEUR

workable

AIXPERT

Thank you!

aixpert-project.eu

info@aixpert-project.eu

mastodon.social/@AIXPERT_project

/company/aixpert-project/

@AIXPERT_project

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Funded by the European Union [AIXPERT, 101214389]. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).