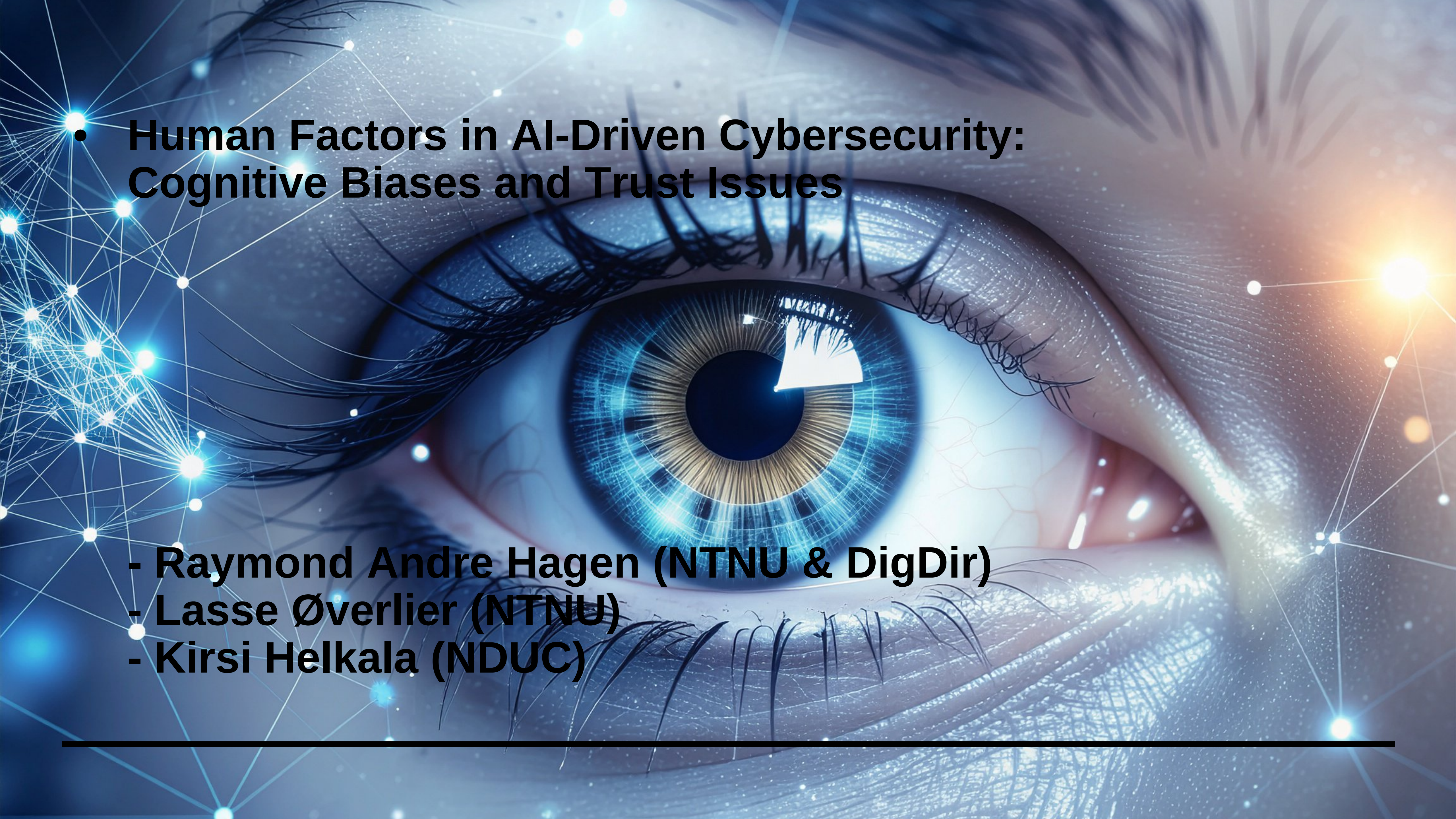


- **Human Factors in AI-Driven Cybersecurity: Cognitive Biases and Trust Issues**

- Raymond Andre Hagen (NTNU & DigDir)
- Lasse Øverlier (NTNU)
- Kirsi Helkala (NDUC)

Why this matters now

SOCs are adopting AI for triage, correlation, and response.

But trust, explainability, and bias shape whether AI helps or harms.

For incident commanders and forensics, defensibility ≠ speed alone.

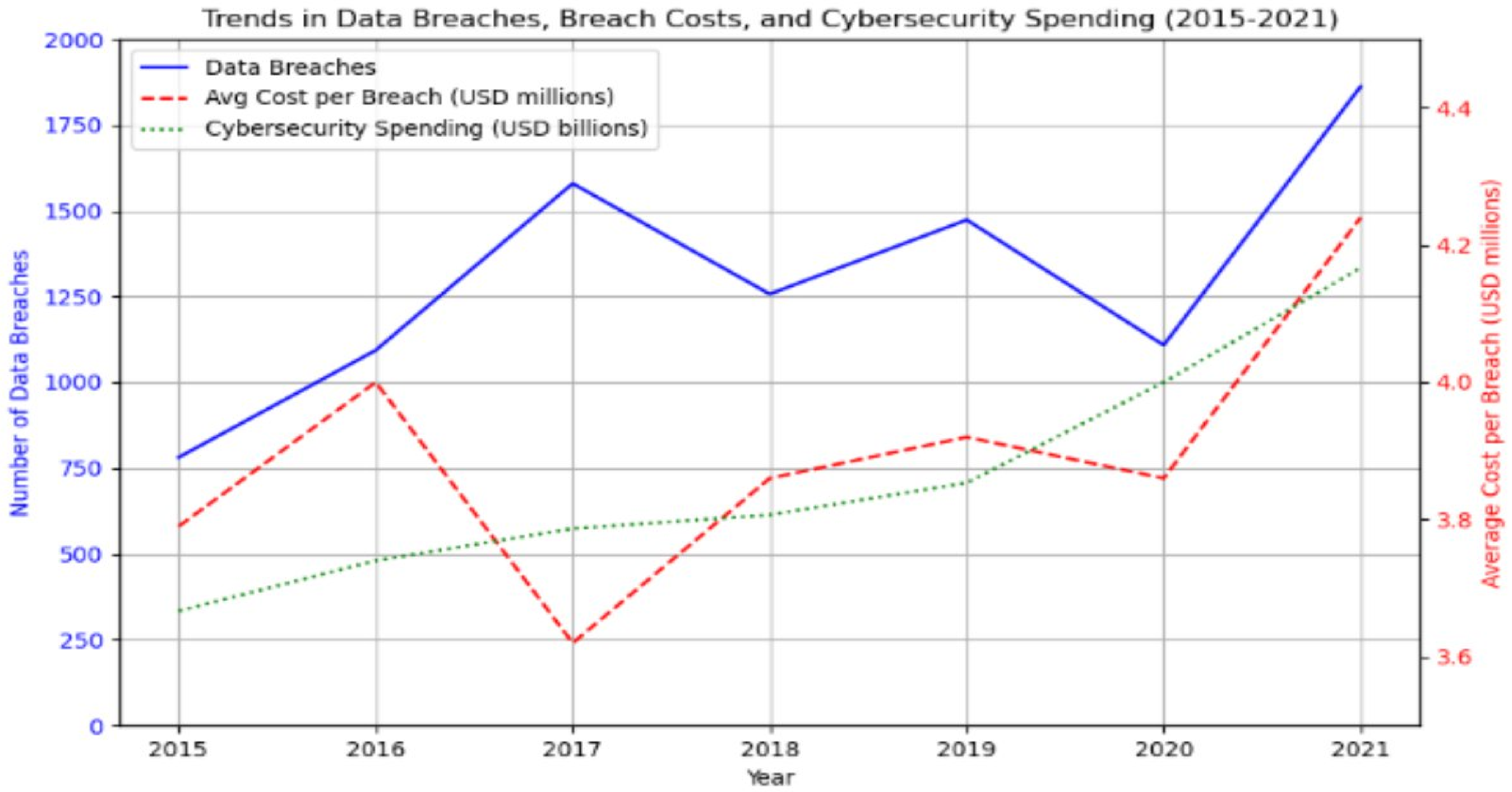
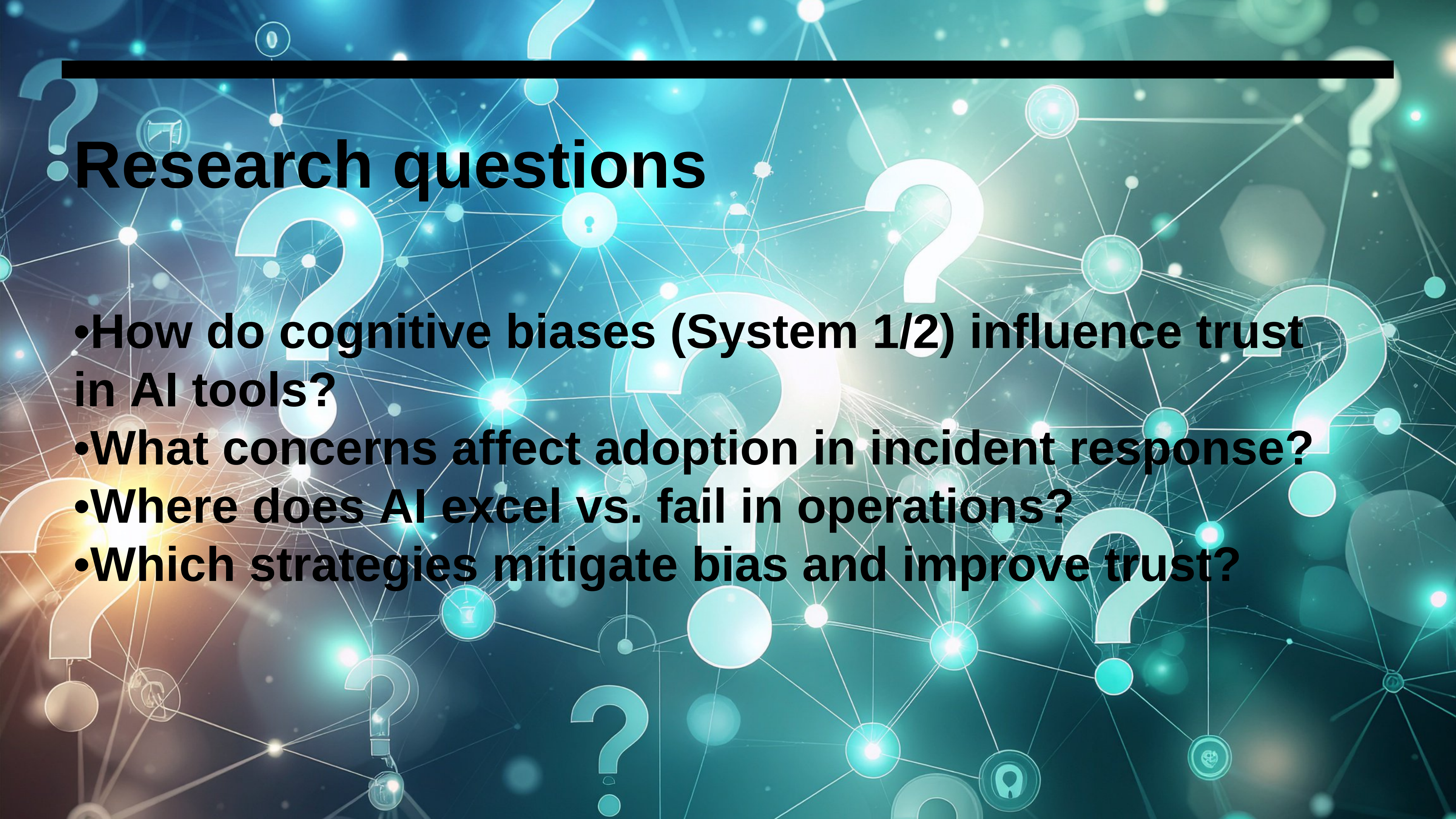


Fig. 1. Trends in Data Breaches, Breach Costs, and Cybersecurity Spending (2015-2021). Data on breaches is sourced from the Identity Theft Resource Center[10], breach costs from IBM[11], cybersecurity spending from Gartner[12], and cybercrime cost projections from Cybersecurity Ventures[13].



Research questions

- How do cognitive biases (System 1/2) influence trust in AI tools?
- What concerns affect adoption in incident response?
- Where does AI excel vs. fail in operations?
- Which strategies mitigate bias and improve trust?

Study design & data

- 19 semi-structured interviews (multi-sector, global)
- Comparative analysis of four vendor approaches (Microsoft, CrowdStrike, Darktrace, IBM)
- Thematic coding (deductive + inductive) with NVivo

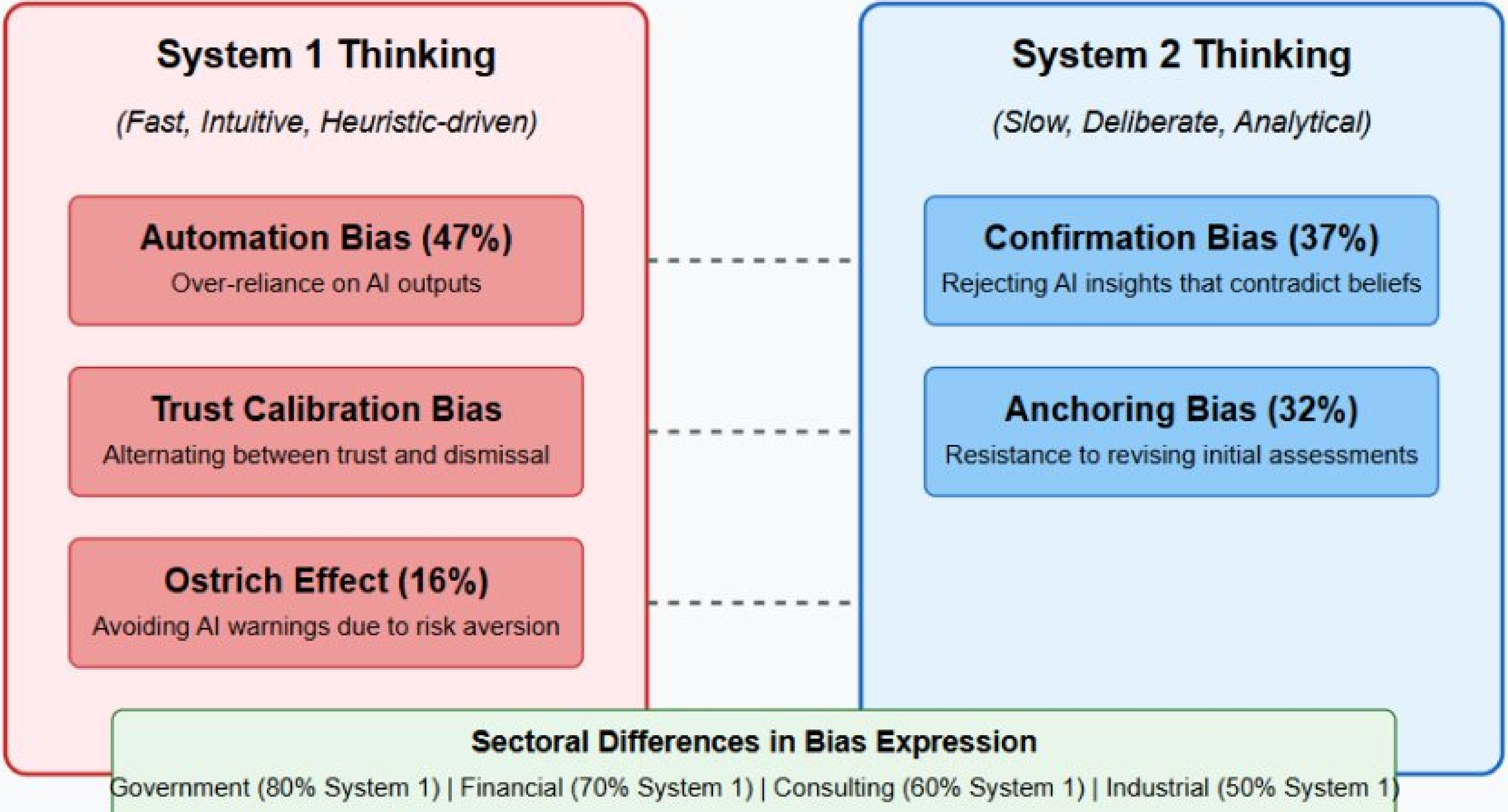
Theoretical lens: System 1 vs. System 2

SYSTEM 1: FAST, INTUITIVE, EFFICIENT — *BIAS-PRONE UNDER PRESSURE*

SYSTEM 2: SLOW, ANALYTICAL, DEFENSIBLE — *RESOURCE-INTENSIVE*

INCIDENT TEMPO PULLS ANALYSTS TOWARD SYSTEM 1; AUDITS DEMAND SYSTEM 2

System Thinking and Cognitive Biases in Cybersecurity



System 1 Thinking

(Fast, Intuitive, Heuristic-driven)

Automation Bias (47%)
Over-reliance on AI outputs

Trust Calibration Bias
Alternating between trust and dismissal

Ostrich Effect (16%)
Avoiding AI warnings due to risk aversion

System 2 Thinking

(Slow, Deliberate, Analytical)

Confirmation Bias (37%)
Rejecting AI insights that contradict beliefs

Anchoring Bias (32%)
Resistance to revising initial assessments

Sectoral Differences in Bias Expression

Government (80% System 1) | Financial (70% System 1) | Consulting (60% System 1) | Industrial (50% System 1)

Bias landscape in SOC/IR

Automation bias: over-trusting AI outputs

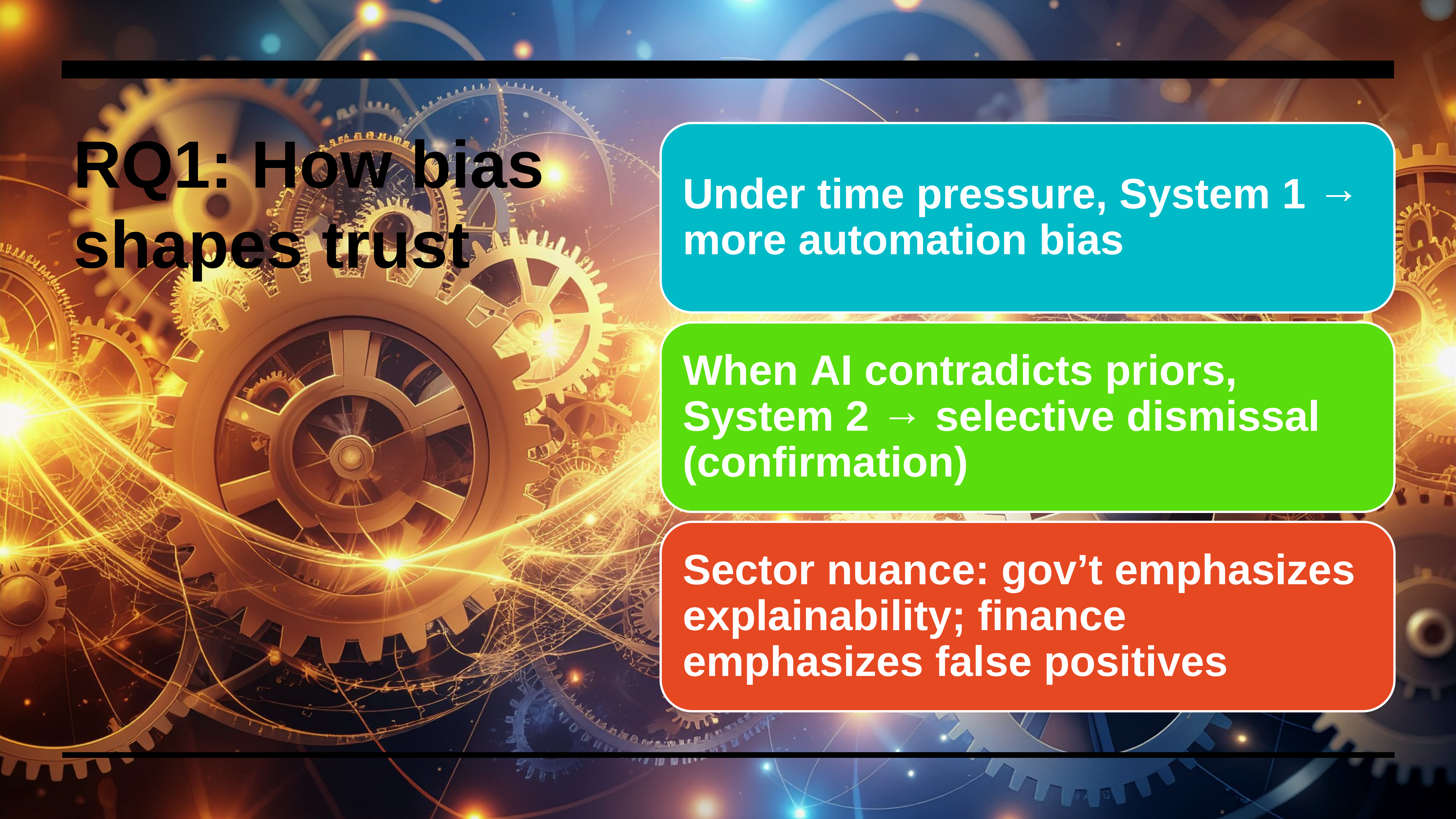
Confirmation bias: discounting AI that challenges priors

Anchoring: sticking to first hypothesis despite new evidence

Ostrich effect: avert warnings when stakes/uncertainty are high

Data snapshot: Trust is brittle

- Mixed trust in AI alerts; skepticism common in regulated sectors
- Dynamic *trust calibration*: swing between over-reliance and rejection
- Explainability gaps and false positives drive most skepticism



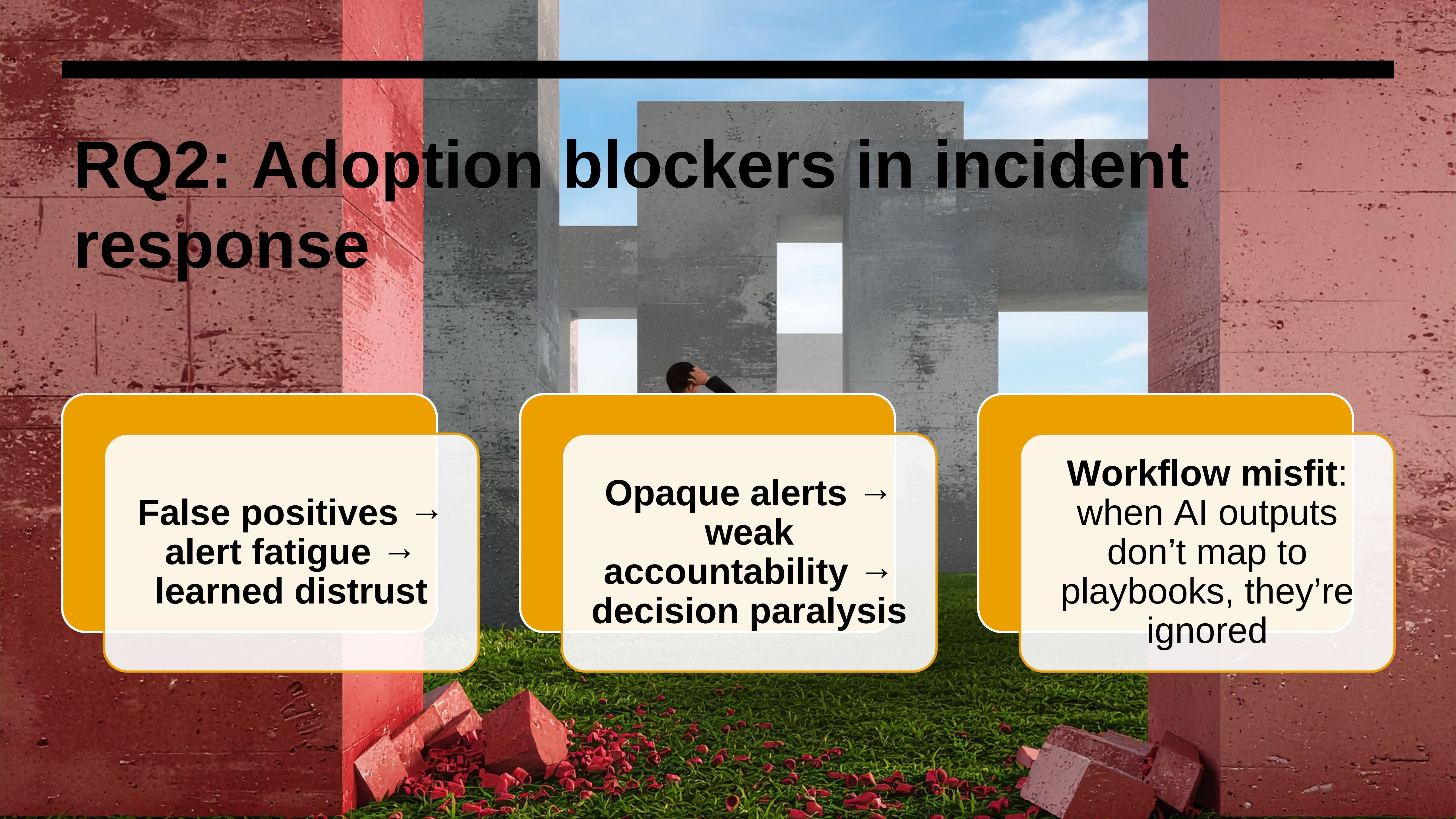
RQ1: How bias shapes trust

Under time pressure, System 1 → more automation bias

When AI contradicts priors, System 2 → selective dismissal (confirmation)

Sector nuance: gov’t emphasizes explainability; finance emphasizes false positives

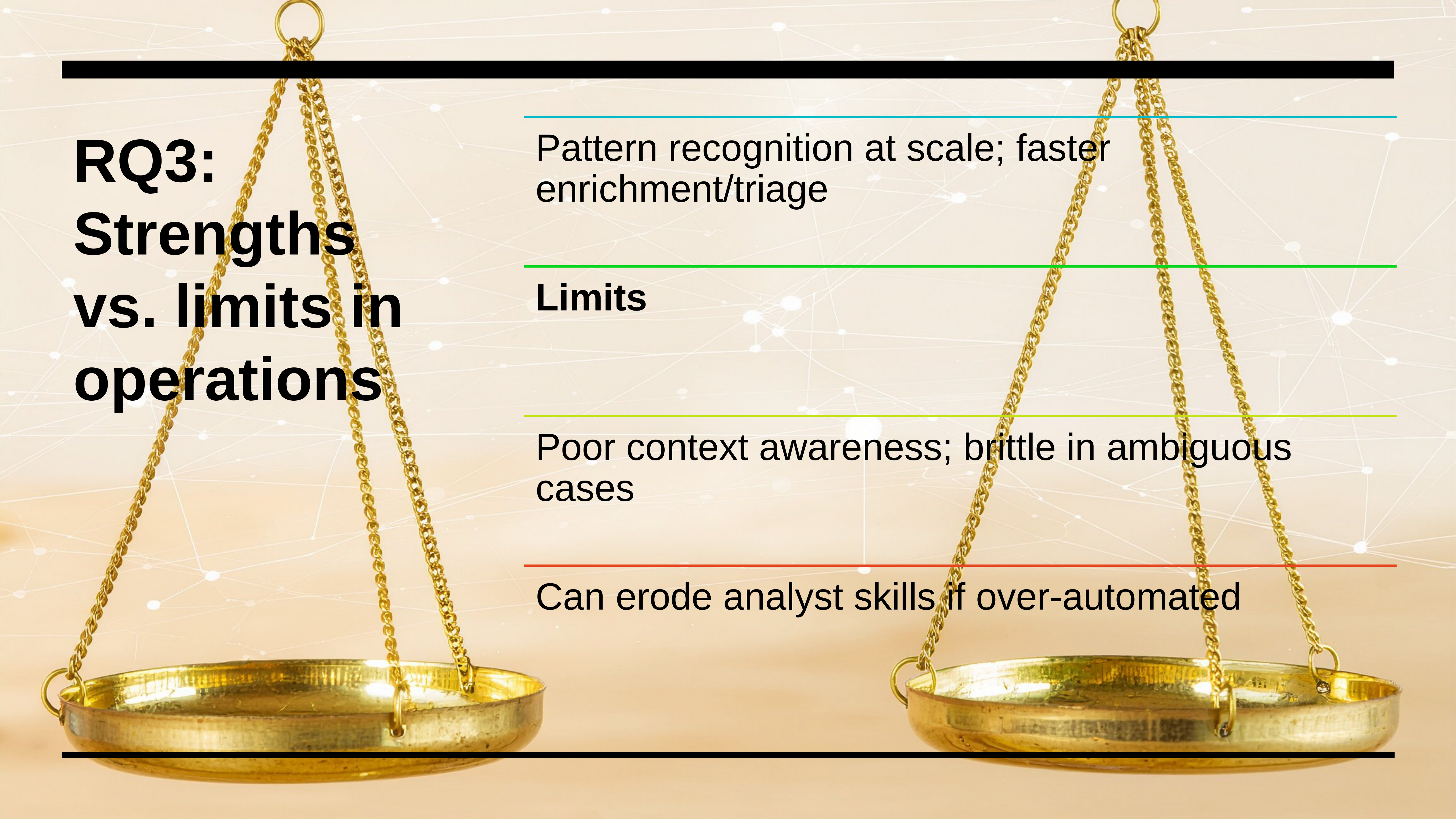
RQ2: Adoption blockers in incident response



False positives → alert fatigue → learned distrust

Opaque alerts → weak accountability → decision paralysis

Workflow misfit: when AI outputs don't map to playbooks, they're ignored



RQ3: Strengths vs. limits in operations

Pattern recognition at scale; faster enrichment/triage

Limits

Poor context awareness; brittle in ambiguous cases

Can erode analyst skills if over-automated

Q4: What actually helps (mitigations)

- **Explainable AI (XAI)**: surface evidence, logic paths, uncertainty
- **Bias-aware training**: name the traps; rehearse counter-moves
- **Adaptive trust calibration**: learn from analyst feedback; adjust thresholds

Design implications (for tool builders & SOC leads)

BUILD **EXPLANATIONS FIRST**, NOT LAST: PROVENANCE, LOGIC TREES, COUNTERFACTUALS

EMBED **FEEDBACK LOOPS**: ANALYST ACCEPTS/OVERRIDES → MODEL LEARNS

HUMAN-IN-CONTROL DEFAULTS FOR HIGH-IMPACT ACTIONS (CONTAINMENT, PURGE)

Contemplation …

Where would you place a feedback loop in your SOC today?

Which bias shows up most in your team’s last major incident?

What evidence would raise your trust in an AI alert tomorrow?

So what did I find ….