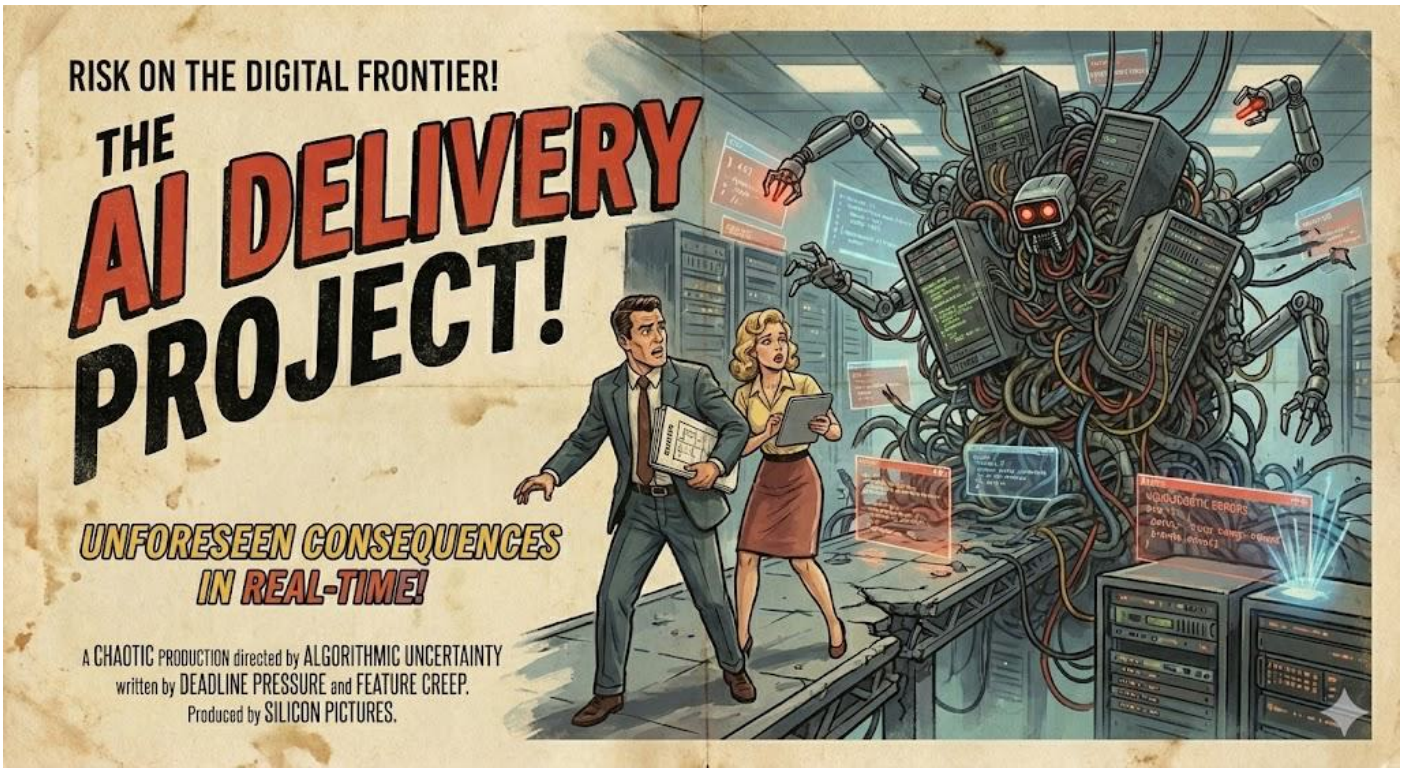
Organisations That are Underestimating AI Risk

Not slightly. Structurally.



The evidence: Recent research shows AI-assisted attacks removing the last meaningful barriers to entry. What used to require specialist teams, time, and money can now be executed by a single motivated actor in days. Sometimes hours. Link

That is not an abstract threat, but an acceleration problem.

Rishi Sunak's warning about "vibe hacking" matters because it reframes cyber risk. Not as sophistication, but as *access*. When AI compresses skill, effort, and cost, volume explodes. Attack frequency rises. Detection becomes harder. Insurance, regulation, and response models fall behind. Link.

Check Point's VoidLink case makes this uncomfortably concrete. A single individual produced a malware framework that previously would have required multiple teams and months of coordinated effort. AI did not invent new attacks. It collapsed the delivery timeline. Link.

Is that the pattern people are missing?

Zooming Out

Is the same acceleration happening inside organisations?

Are AI agents being deployed into production environments with live data access, tool integrations, and decision authority?

Are they often built quickly? Are they often owned loosely? Are they not owned at all?

Attackers are no longer just targeting infrastructure; they are targeting the agents themselves.

Because prompts are not harmless text:

- They define behaviour.
- They encode authority.
- They embed assumptions about what the system is allowed to see, decide, and trigger.

Exposing a prompt could be the equivalent to exposing operating logic. In some cases, **business logic**. In regulated environments, that is not a vulnerability. That is a control failure.

This is already being observed in the market. Prompt extraction. Guardrail bypassing. Tool misuse. The are all documented and part of daily news. Link 1 , Link 2, Link 3

Then there is the part nobody likes to admit: **Most AI failures are not sophisticated attacks, but basic carelessness**.

Nearly 200 AI-powered iOS apps leaking sensitive user data. Not through exotic exploits. Through open databases and misconfigured storage. Chat histories. Emails. Location data. Millions of users exposed. Link

This is not an AI technology problem, but a delivery discipline problem.

I have seen this exact failure mode long before AI. Data platforms rushed live. Dashboards built without access models. Automations wired directly into production systems because “we will tidy it up later”.

“Later” rarely comes.

This is where experience stops being a “nice-to-have”, but a “rather not”

In corporate, particularly in regulated financial services, every new capability starts from one assumption: It *can* fail!

The question is how, where, and what blast radius will be.

The AI initiatives that survived scrutiny shared common traits:

- Clear ownership, not shared enthusiasm.
- Explicit constraints on scope and authority.
- Human checkpoints where decisions carried risk.
- Evidence that the system could be stopped, audited, and explained.

They were slower to launch, but they scaled safely.

The **Philosophical** Perspective

The real risk is not that AI will become malicious.

It is that organisations deploy decision-making capability faster than they can govern it, monitor it, or even understand how it behaves under pressure.

Security is not something you attach on once the model works operationally, because by then, the risk is already embedded in the solution.

If you are using AI today, ask yourself something uncomfortable.

If this system is abused, misused, or manipulated, how quickly would you know?

And who is accountable when you do?

If those answers are vague, the risk is already real.

That is not fearmongering, but delivery reality.

#ArtificialIntelligence #CyberSecurity #RiskManagement #AIgovernance #Leadership #OperationalRisk

Disclaimer and Disclosure

Third-party Content and AI Assistance: This article references tools and software that are publicly available and proprietary to their respective creators. The author does not claim ownership or affiliation with these third-party products. This article was written by the author with assistance from Generative AI Language Models.

Transparency Notice: While every effort has been made to ensure accuracy, readers should verify information independently and consult official sources or documentation for the mentioned tools and software. The use of AI in the writing process is disclosed in the interest of transparency, but all opinions and analyses are the author's own unless otherwise stated.