Building Trustworthy AI in Financial Services

Security, Resilience and Governance Beyond Prompt Injection



Hemant Patkar, CISSP®

Cybersecurity Lead Designer/Architect | Financial Services

Convincing Signals Can Cause Financial Loss

US$1.28bn

Reported deepfake fraud losses in 2025



Trustworthy AI is not about spotting every fake. It is about designing controls so that a convincing fake cannot cause real harm.

Prompt Injection Is Only One Part of the Risk

Protect the complete decision system, not only one attack surface.

ONE DOOR

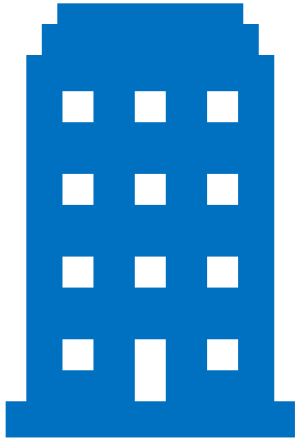


PROMPT INJECTION

A real and important attack technique

WIDER VIEW

THE WHOLE BUILDING



TRUSTWORTHY AI

SECURITY	Prevent system from misuse
RESILIENCE	Remain safe when things fail
GOVERNANCE	Explain and justify decisions

AI Is a Decision Supply Chain

The model is only one link. Every link can change the customer outcome.

01 DATA



What the system knows

Customer and credit information

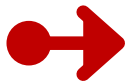
02 MODEL



How it interprets

Assessment or recommendation

03 ACTION



What can execute

Rules and approvals

04 OUTCOME



Who or what is affected

Approve or decline

ONE WEAK LINK CAN CHANGE THE OUTCOME

A trustworthy model inside an untrusted process still creates an untrusted outcome

What Makes the Full Decision System Trustworthy?

Trustworthy AI depends on security, resilience and governance working together.

SECURE × RESILIENT × GOVERNED = TRUSTWORTHY AI



01 **SECURE**

Can somebody misuse or manipulate the system?

02 **RESILIENT**

Will the system remain safe when something fails?

03 **GOVERNED**

Can we explain and justify the decision?

Trustworthy AI is not built on one control. It is built on security, resilience and governance working together.

Controls What AI May See and Do

SECURITY

Being intelligent does not remove the need for permissions, limits and independent approval.

01 SEE

Control the data

What may AI access?

02 DO

Control the action

What may AI execute?

03 APPROVE

Separate advice from authority

What must remain independent?

The principle is simple: AI can be flexible. Controls should not be.

Resilience Keeps the Service Safe When AI Fails

RESILIENCE

Not perfect uptime. A safe response when something goes wrong.

1 **ANTICIPATE**

Know what could go wrong

2 **CONTAIN**

Limit the impact

3 **RECOVER**

Return to a safe service

When AI fails, do we know what happens next?

Because resilience is not about proving that failures will never occur. It is about proving that the organisation remains safe when they do.

Governance Should Match the Consequence

GOVERNANCE

The same model can support a simple task or influence a customer's financial outcome. Governance should reflect the consequence.



LOWER IMPACT

Draft, summarise or assist

EXAMPLES

- Summarise meeting notes
- Draft an internal document
- Help an employee find information

Approved data • Quality checks • Monitoring • User review



HIGHER IMPACT

Recommend, decide or act

EXAMPLES

- Approve or release a payment
- Decline credit or price a product
- Freeze an account or take autonomous action

Independent validation • Human oversight • Outcome checks • Tested fallback

Not Innovation. Applying Assurance

Three Moves Build Operational Trust

01 MAP



Know where AI matters

- Important decisions
- Customer outcomes
- Critical services
- Named owner and dependencies

02 TEST



Test the whole workflow

- Permissions and limits
- Model and data changes
- Supplier failures
- Manual fallback and recovery

03 EXERCISE



Practise before pressure

- Run one realistic incident
- Include business, risk and cyber
- Include legal and communications
- Rehearse suppliers and hand-offs

Policies create intent. Operational readiness creates trust.

Five Questions Test the Complete Decision System

If you can answer these five questions clearly, you understand the decision risk.

01	**Should it decide?**	Purpose • owner • potential customer harm
02	**What may it see?**	Required data • source • sensitivity • retention
03	**What may it do?**	Permitted tools • actions • limits • approvals
04	**How will it fail?**	Safe state • stop mechanism • fallback • recovery
05	**Can we prove it?**	Reconstruct what happened, why and who approved it

ONE QUESTION TO TAKE BACK

Trust is Designed.

The demo is Flawless. The decision system earns trust.

DO NOT ASK ONLY

“Do we trust the model?”

ASK INSTEAD

“Can we trust the decision system?”

Can we explain?

- What AI can see?
- What AI can do?
- How it fails?
- Who owns outcome?
- What evidence remains?