Attack & Defence for First-Time Payee Decisioning in Real-Time Payments: Protecting Against APP Fraud

COMP1806 - Information Security, University of Greenwich (2025)

Research Task by Destiny Onyemerekwe

Table of Contents

Executive Summary

This project assesses the **first-time-payee decision service used in real-time payment (RTP) to prevent Authorised Push Payment (APP) fraud.** The scope covers the decision service, its CI/CD release process, facts feature loaders, and the payment routing based on decisions. I began by outlining the assets and mapped them to threats and attacks that impact the organisation and variables, allowing a comprehensive and narrowed risk assessment, risk management, defences, and planning.

Promotion of unsafe versions to live, within the decision service, I have assessed to be the top risk for this scenario. It corrupts the behaviour for the first-time-payee decision service until it's rolled back (causing a trickling impact on approval of numerous payments as long as it is live), persists until identified, and widens the landscape of multiple attack routes (pipeline abuse, data poisoning, or phish-driven approvals).

After the recommended controls, I score the residual risk to be approximately 40/100 (Medium) for this scenario, which is a significant decrease from the estimated score of 80/100 before implementing any controls. Controls such as gates, ensuring fail-closed, and verifying the appropriate use of tokens/MFA, present a 50% reduction in attack success probability. This reflects an efficiency to the robustness of the security of the decision service while accepting some remaining exposure, such as operational errors and edge-case misses in tests.

Recommended controls (with impact & possible hindrance):

1. **Two-person promotion + CI gates (tests/canary) with auto-rollback.**
    a. **Impact:** prevent unsafe releases or quickly undoes them (reducing further damage).
    b. **Hindrance:** release of promotions may be slower and a greater demand for coordination.
2. **Fail-closed routing with queue/hold - no pass-though if checker is down/suspicious**
    a. **Impact:** blocks bypass during incidents.
    b. **Hindrance:** temporary customer delay.
3. **Phishing-resistant access (MFA FIDO/WebAuthn) + short-TTL tokens with alerts (on creation/policy changes)**
    a. **Impact:** reduces chances of access takeover/persistence that leads to unsafe promotions.
    b. **Hindrance:** causes minor user friction and requires alert tuning.

Focusing on the integrity of the promotions of versions, through these recommended controls, as well as consistent monitoring and a trailed incident plan, lowers expected loss from fraud while keeping payment processing available.

Asset Inventory

No.	Asset	Type	Why it matters
1	Risk facts database	Data	If facts are wrong/missing/out-of-date it will lead to a wrong decisions (false negatives/positives)
2	Model/rules list	Software	Decides the logic for every payment; a bad version worsens all decisions until roll back
3	Hardware security module (HSM)	Hardware	Attackers can impersonate services, push fake models, or call payment APIs, if keys leak
4	Safety-check API	Software	If slow/unavailable, fallbacks maintain speed and availability of payments but raise chances of risky payments slipping through
5	Payment network connection	Network	Must enforce the safety-check's outcome; otherwise a payment may contradict the original decision
6	Confirmation of Payee (CoP)	Service	A mismatch and other risk signals should trigger a warning pop-up or hold
7	Customer account & saved payees	Data	If saved payee is edited by an attcker, a legitimate payment can be misdirected
8	Customer app session & device binding	Service	Attackers can impersonate customer (account takeover), bypassing checks by exploiting weak binding
9	On-call admin account	People	Access priviledge; compromise can diable protections or allow unsafe changes
10	Training examples (past payments + labels)	Data	If tampered, future versions learn incorrect patterns - future decisions systematically miss scams

Top 3 assets: they directly decide each outcome at the moment of payment: the facts used, the logic applied, and the keys that prove it's genuine; failures here immediately and system-wide increase missed scams

Table 1: Asset Inventory for First-Time Payee Decisioning

The top 3 assets, highlighted in yellow, directly decide each outcome at the moment of payment: the facts used, the logic applied, and the keys that prove it's genuine; failures here immediately and system-wide increase missed scams.

Weighted-Factor Valuation

No.	Asset	Revenue	Profitability	Public Image	Weighted Value
1	Risk facts database	95.0	90.0	85.0	90.0
2	Model/rules list	90.0	80.0	80.0	83.0
3	Hardware security module (HSM)	90.0	85.0	90.0	88.0
4	Safety-check API	90.0	75.0	70.0	78.0
5	Payment network connection	85.0	80.0	70.0	78.5
6	Confirmation of Payee (CoP)	75.0	70.0	70.0	71.5
7	Customer account & saved payees	80.0	75.0	85.0	79.5
8	Customer app session & device binding	80.0	70.0	80.0	76.0
9	On-call admin account	85.0	70.0	65.0	73.0
10	Training examples (past payments + labels)	80.0	85.0	90.0	85.0

Factors & weights: Impact to Revenue = 0.3, Profitability = 0.4, Public Image = 0.3 (sum = 1.0)
Method: score each asset 0-100, then computer WV = (0.3 * Revenue) + (0.4 * Profitability) + 0.3 * Public Image)

5 top assets (by highest weighted values):

1. **Risk facts database — 90.0**
   Working: (0.3 * 95) + (0.4 * 90) + (0.3 * 85) = 90.0
2. **Hardware security module (HSM) — 88.0**
   Working: (0.3 * 90) + (0.4 * 85) + (0.3 * 90) = 88.0
3. **Training examples (past payments + labels) — 85.0**
   Working: (0.3 * 80) + (0.4 * 85) + (0.3 * 90)
4. **Model/rules list — 83.0**
   Working: (0.3 * 90) + (0.4 * 80) + (0.3 * 80)
5. **Customer account & saved payees — 79.5**
   Working: (0.3 * 80) + (0.4 * 75) + (0.3 * 0.85) = 79.5

Table 2: Weighted-Factor Asset Valuation

Weighted-Factor Valuation Arithmetic

Factors & weights:

- Impact to Revenue = **0.3**
- Profitability = **0.4**
- Public Image = **0.3**

  (Sum = 1.0)

Method (per asset):

$$WV = (0.3 \times \text{Revenue}) + (0.4 \times \text{Profitability}) + (0.3 \times \text{Public Image})$$

Top 5 Assets (with working shown)

Risk facts database – 90.0

$$(0.3 \times 95) + (0.4 \times 90) + (0.3 \times 85) = 90.0$$

Hardware security module (HSM) – 88.0

$$(0.3 \times 90) + (0.4 \times 85) + (0.3 \times 90) = 88.0$$

Training examples (past payments + labels) – 85.0

$$(0.3 \times 80) + (0.4 \times 85) + (0.3 \times 90) = 85.0$$

Model/rules list – 83.0

$$(0.3 \times 90) + (0.4 \times 80) + (0.3 \times 80) = 83.0$$

Customer account & saved payees – 79.5

$$(0.3 \times 80) + (0.4 \times 75) + (0.3 \times 85) = 79.5$$

TVA Grid

ID	Threat	Vulnerability	Confidentiality	Integrity	Availability
A1-T1	Tamper/poison facts	No integrity checks/validation on ingest	Low	High	Medium
A1-T2	Facts out-of-date	No freshness limit/alerts; no "hold if out-of-date" rule	Low	High	High
A1-T3	Bulk read of data	Unrestricted read permission of data; weak audit on large exports of data	High	Low	Low
A2-T1	Wrong version of payment logic promoted	Weak approvals; poor change records; no checks	Low	High	High
A2-T2	Rollback blocked/slow	No one-click to revert; poor rollback process	Low	High	High
A2-T3	Admin phished	Weak admin multi-factor authentication (MFA); social engineering	Medium	High	Low
A3-T1	Key extraction/abuse	Misconfigured HSM policy; keys stored in software; admin interfaces left open	High	High	High
A3-T2	Long-lived tokens	No Time-To-Live (TTL) limit	High	High	Low
A3-T3	Key leak in backups	Unencrypted backups; poor key-export controls	Medium	Medium	Low

Table 3: Threat Vulnerability Asset (TVA) Grid for Top Assets

Quantitative Risk Calculation

Vulnerability	Threat	Likelihood	Attack Success Probability (ASP)	Probable Loss (%)	Asset Value (£)	Confidence	Risk	Interpretation
Risk facts database	T1 - Tamper/poison facts	0.3	0.5	40.0	500,000.00	Medium	30000	Wrong facts drive wrong allow/hold decisions until they're detected
Model/rules list	T1 - Wrong version promoted	0.2	0.6	60.0	700,000.00	Medium	50400	One bad version affects all customers until it's rolled back

Final risk arithmetic
1. **Risk facts database:** (0.3 * 0.5) * (0.4 * £500,000) = £30,000
2. **Model/rules list:** (0.2 * 0.6) * (0.6 * £700,000) = £50,400

Table 4: Quantitative Risk Calculation for Selected High-Value Assets

Final Risk Arithmetic

1. **Risk facts database**

$$\text{Risk} = (0.3 \times 0.5) \times (0.4 \times £500{,}000) = £30{,}000$$

2. **Model/rules list**

$$\text{Risk} = (0.2 \times 0.6) \times (0.6 \times £700{,}000) = £50{,}400$$

CVSS-Lite Vulnerability Scoring

Vulnerability (asset)	Exploitability	Impact	Scope (1.0/1.5)	Vulnerability Score (VS)	Attack Success Probability (ASP): before and after	Risk: before and after (£)
Weak approval on promotion (models/rules list)	8	9	1.5	12.75	0.6 to 0.45	50,400 to 37,800

CVSS-lite vulnerability scoring arithmetic

1. **Vulnerability score** = (Exploitability + Impact) / 2) * Scope
2. **ASP delta** = min(1, VS / 15)
3. **Recomputed risk** = (Likelihood * ASP delta) * (Probable Loss * Asset Value)
   (Using A4 values: Likelihood = 0.20, Probable Loss = 0.60, Asset Value = £700,000)

Table 5: CVSS-Lite Vulnerability Scoring for Key Asset Vulnerability

CVSS-Lite Vulnerability Scoring Arithmetic

1. **Vulnerability Score (VS)**

$$VS = (Exploitability + Impact) / 2 \times Scope$$

(using 8, 9, scope 1.5)

$$VS = (8 + 9) / 2 \times 1.5 = 12.75$$

2. **ASP delta**

$$ASP\ \Delta = \min(1, VS/15)$$

Recomputed Risk (before & after controls)

Risk = (Likelihood × ASP Δ) × (Probable Loss × Asset Value)

Using A4 values:

- Likelihood = 0.20
- Probable Loss = 0.60
- Asset Value = £700,000

Before controls:

$$Risk = 0.20 \times 0.60 \times 0.60 \times 700{,}000 = £50{,}400$$

After controls:

$$Risk = 0.20 \times 0.45 \times 0.60 \times 700{,}000 = £37{,}800$$

Attack Path

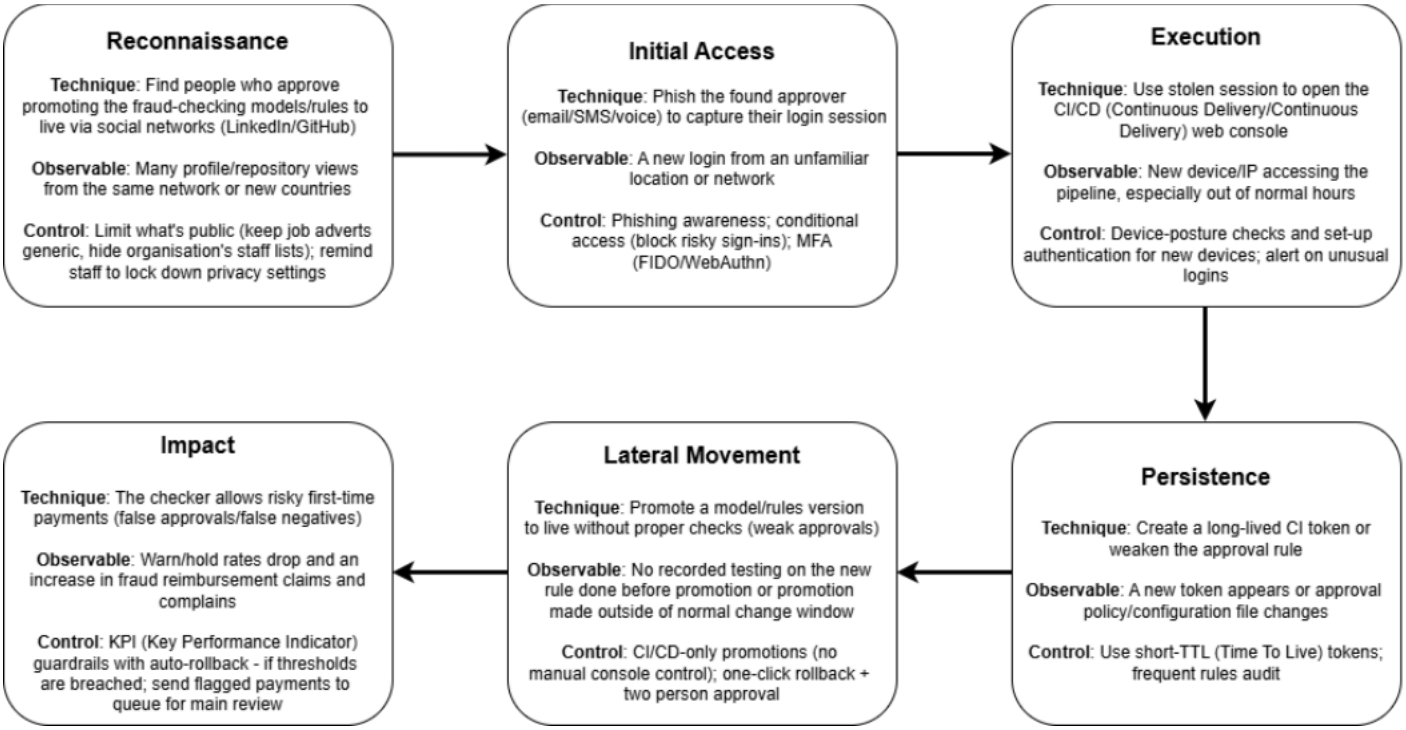

Figure 1: MITRE-Style Attack Path for Unsafe Promotion Scenario

Attack Grid

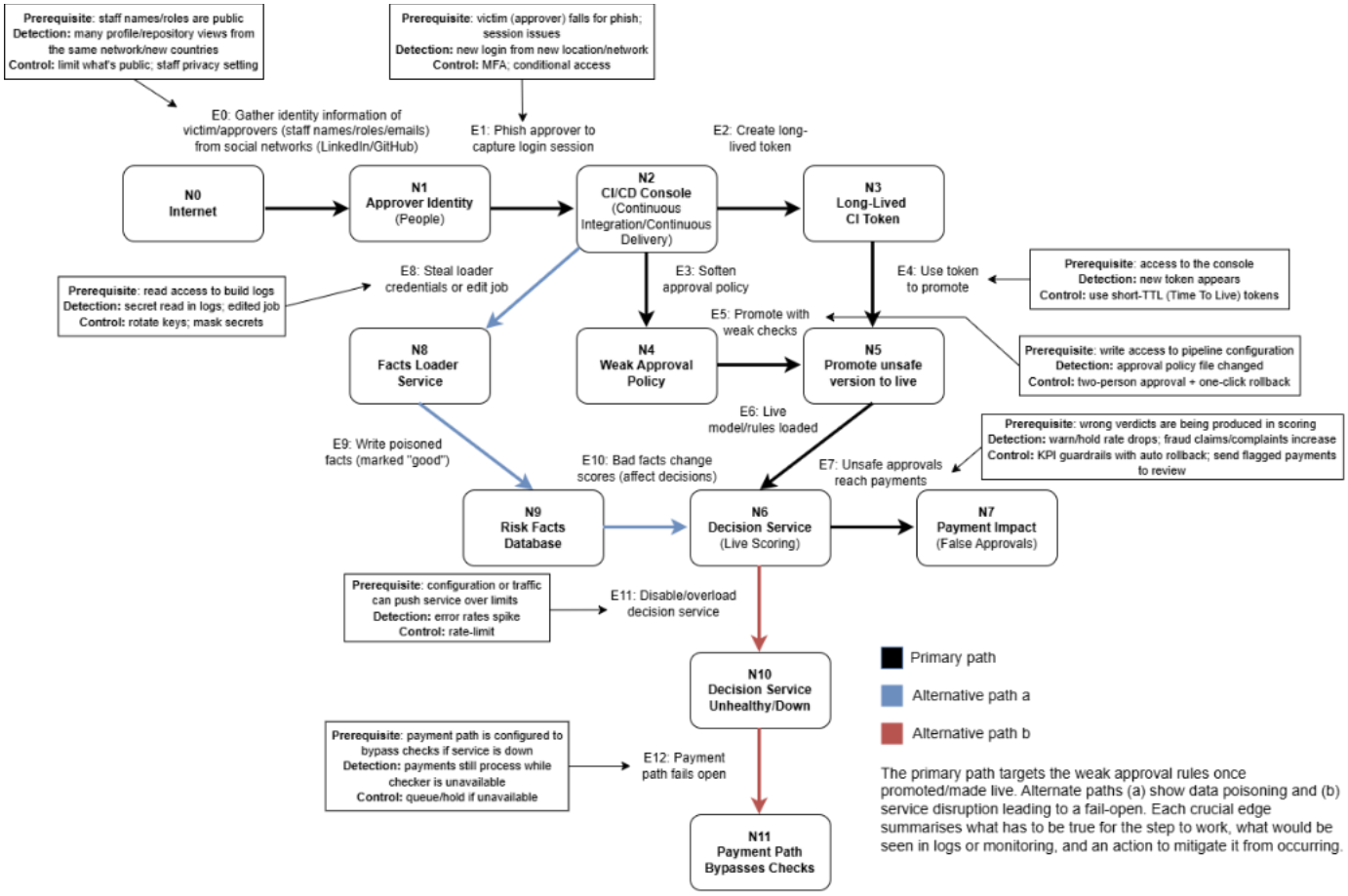

Figure 2: Attack Graph Showing Branching Paths, Preconditions, and Multiple End-States

Defences and Controls Map

Attack Step	Technical Control	Organisational Control	Estimated Effectiveness (%)	Trade-off
E1: Phish approver	Enforce MFA (Multi-Factor Authentication) + conditional access in the IdP (block risky sign-ins/new device or location)	Targeted anti-phish awareness training for staff	60.0	Extra effort for login processes and some time for training , but it eliminates many early easy wins for attackers
E2: Create long-lived token	Issue short TTL (Time-To-Live) tokens - shrinking "useful" time for attackers + alert on token creation	Regular audit and verification of access tokens by DevOps	50.0	Consistent periodic audit and verification can become time-consuming for DevOps teams
E3/5 to N5: Weaken approbal policy/promote unsafe version	Mandatory CI gates (checkpoints), defined by passing canary tests and auto-rollback reverts on bad metrics, before promoting	Implement a two-person promotion policy (author cannot approve their own promotion) - SoD (Separation of Duties)	70.0	Release of promotions may be slower and need more coordination - but avoid damage from one "push"
E8: Steal loader credentials/edit job	Mask secrets in logs and require approvals for loader job changes	Implement a key rotation schedule and ownership defined in a SOP (Standard Operational Procedure)	55.0	Requires more maintanance to rotate keys and extended process in logging and auditing ownership/rotation
E10: Bad facts change scores	Hold process decided by freshness and performance (within expected ranges) checks	A SOP document containing data quality (ranges for freshness and performance) that define whether to trigger alert to hold	50.0	Some good traffic can be held if the tresholds are too strict; needs consitent tuning
E11/12: Decision service unhealthy/down to fail-open	Protect the checker during overload: slow traffic, add capacity, and block pass-through when unhealthy	Implement a rulebook to follow that follows: queue/hold payments if checker is down	70.0	During incidents, queue may build and payments slow; autoscaling adds cost

Table 6: Defences and Controls Mapped to Attack Stages

Incident Response Plan

Immediate containment

1. **Rollback & freeze promotions:** trigger a "one-click" rollback that flips back to the last known good version; freeze promotions - block push of new versions to live in CI/CD pipeline.
2. **Fail-closed payments:** block payments and route to queue/hold or manual review, if the decision service is missing or wrong.
3. **Revoke access:** for approver that's linked to the promotion's session/token invalidate sessions/tokens and re-enforce authentication.

Evidence to preserve

1. **CI/CD audit:** shows what has been edited (policy changes/token creation) and who approved (who's session/token has been used).
2. **Identity logs (IdP) logs:** who has logged in and MFA details during promotion window.
3. **Behaviour of the decision service/KPIs:** snapshots of the impact of the incident; how the behaviour of the system has changed and change in KPIs.

Stakeholder notification list & timelines

0-30 mins:

- **DevOps:** build/develop the pipelines and deploy the credentials; rollback and freeze promotions.
- **Features Team:** analysts/developers of the model/rules; validate the safe version and monitor behaviour of the decisions service; approve the intended live version, advise on when the behaviour returns to safe, through consistent monitoring and canary.
- **SRE (Site Reliability Engineering):** run health checks, scaling, and performance to ensure reliability; enforce fail-closed; monitor health.
- **Fraud Operations:** while technical teams work on recovering the incident, increase manual review to catch fraudulent transactions.
- **SecOps (Security Operations):** security/on-call team; revoking sessions and resetting MFA for an infiltrated approver account.

0-60 mins:

- **Product Owner (Payments):** responsible for the vision for the payment product (features) and aligns it with the business' goals; approve containment and customer impact trade-offs.
- **Compliance/Risk:** ensure regulations and policies are followed by the company to mitigate company damage; assess the damage (financial loss, customer harm) and decide regulations/partner notices based on policy.
- **Communications Team (Internal/External):** prepare messages to notify customers that their payments are temporarily held whilst on-going checks/when resolved.

Same business day:

- **Senior Management:** oversee and support the acceptance of risk; approve freeze/unfreeze strategy and customer impact.
- **Customers (if needed):** inform affected users.

Remediation steps

1. **Re-secure access (CI/CD + IdP):** revoke CI token/session of affected approvers, reset login credentials, re-enrol MFA.

2. **Renew pipeline controls:** two-person approval, conduct tests/canary, create alerts on policy changes/promotions.
3. **Harden runtime metrics:** set KPI guardrails, that monitor behaviour drifting, with auto-rollback, queue/hold payments is decision checker is missing/unhealthy.

Rollback timeline:

- **Stabilising (0-30 mins):** rollback to last known good version, freeze promotions, revoke CI token.
- **Containment (30-120 mins):** reset accounts, introduce two-person approval, implement "alert-on-promotion".
- **Restoration (120+):** shadow/canary healthy version, monitor the health and KPIs, unfreeze when metrics are back to expected performance

Quick Policy – APP Fraud Decisioning & Release

The primary intent of this policy is to assure safe operation of the first-time-payee decision service and its pipeline: CI/CD (build/release), facts/feature loaders, runtime service, and payment routing. Approver/admin access must use MFA with FIDO/WebAuthn via the IdP. CI/CD tokens must be short lived (short TTL) with alerts on creation/policy changes. Approval of promotions require two independent persons (separation of duties), pass mandatory gates (tests and canary/shadow), and have auto-rollback enabled on adverse or drifting metrics. In the instance that the decision service is missing, unhealthy, or returns mistrustful verdicts, the payment path must be fail-closed: queue/hold or route payments to manual review (no pass-through). Risk-facts quality must be enforced at load and at use: freshness thresholds and sanity/range checks are required; vetoed or stale inputs must trigger hold and alert. The following must be continuously monitored with alerts: decision-service health (errors/latency), drift in metrics or input, and business KPIs (approval %, warn/hold %, claims/reimbursements). Upon detection of an unsafe promotion or fail-open risk, the CSIRT (Cyber Security Incident Response Team) must execute the incident response plan: immediate rollback and freeze promotions, preserve evidence (IdP, CI/CD audit, snapshots of decision service behaviour/KPI), notify stakeholders on defined timelines, and contained recovery with verified tests. The responsibilities of assuring the processes of safe end-to-end operations of the decision service fall on: DevOps (promotions, approval gates and tests, developing promotion/policy-change alerts, freezing controls), Feature Team (KPI guardrails, managing integrity of model/rules, drift tuning), SRE (fail-closed routing, monitor health of the service, incident bridge, between the incident and business continuity), SecOps (MFA with FIDO/WebAuthn and conditional access, short TTL tokens - continually reviewed, session/token revocation), Data Engineering (risk-facts quality - fresh and within integral range), and Fraud Operations (business KPIs, manual-review workflow and threshold). This policy is reviewed every 6 months or after any incident affecting first-time-payee decision service by the Product Owner (Payments) and the CSIRT Lead.

Reflective Learning Statement

Through this project I was able to go further than just explain security at a surface level and examine risks and threats posed in technology through scoping in on real use-case scenarios that have a meaningful impact to various components – allowing me to explore information security under a professional lens. I learned how narrowing into a specific scenario – first-time payees – enables the ability to gain a more concrete understanding of what and who needs to be protected, and why, rather than attempting to assess risk and implement defences to vague ideas. Through this, the TVA grid, weighted evaluation, and attack graph were more precise and defensible. Most importantly, this project gave me the opportunity to develop a deeper and industry-preparing level of cyber security within the financial industries – specifically ensuring security of critical ML models, which are increasingly being implemented in this industry (as seen in American Express' partnership with Nvidia for their LSTM fraud detection model). Exploring the processes of information security demanded me to research the various systems/platforms (CI/CD, version promotion, IdP), processes (tokens (long-TTL and short-TTL), canary/shadow, types of MFA (WebAuthn/FIDO), fail-closed/open), and teams involved (DevOps, Sec Ops, SRE). If I were to revisit this project, what I would do differently would be to outline the baseline for the KPIs and conceptualise the rollback/fail-closed sooner so that my residual-risk is estimated by real thresholds rather than elementary assumptions.

References

AuditBoard (2025). *NIST Incident Response: Your Guide to the 4 Phases of IR.* Available at: https://auditboard.com/blog/nist-incident-response

CISA (2022). *Implementing Phishing-Resistant MFA.* Available at: https://www.cisa.gov/sites/default/files/publications/fact-sheet-implementing-phishing-resistantmfa-508c.pdf

Thornton & Lowe (n.d.) *The Complete Guide to Policy Writing.* Available at: https://thorntonandlowe.com/guide-to-policy-writing/

M.E. Whitman, H. M. (2017). *Principles of Information Security. 6th edn*.

MITRE (n.d.). *ATT&CK® Matrix for Enterprise.* Available at: https://attack.mitre.org/matrices/enterprise/

OWASP (n.d.) *Top 10 CI/CD Security Risks.* Available at: https://owasp.org/www-project-top-10-ci-cdsecurity-risks/

NVIDIA (n.d.). *American Express Prevents Fraud and Foils Cybercrime with NVIDIA AI Solutions*. Available at: https://www.nvidia.com/en-gb/customer-stories/american-express-prevents-fraud-andfoils-cybercrime-with-nvidia-ai-solutions/

Declaration of AI use

I have used AI while undertaking my assignment in the following ways:

- To develop research questions on the topic
- To explain concepts
- To support my use of language