# AI Hallucination & Confabulation Playbook

*Detecting, Containing & Remediating Confident AI Fabrication of Facts*

AIPB-001 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-001 |
| **Playbook Owner** | Chief AI Officer / AI Operations Lead |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Reputational, Legal & Compliance Risk |

## 2. Scenario Description

### Overview

AI hallucination (confabulation) occurs when an AI system generates plausible-sounding but factually incorrect, fabricated, or entirely invented content — presented with apparent confidence. This includes citing non-existent sources, generating false statistics, fabricating legal or medical guidance, inventing names or events, or producing internally inconsistent outputs. The risk is highest when AI outputs are used in customer-facing, regulatory, medical, legal, or financial contexts without adequate human verification.

### Detection Indicators / Trigger Conditions

- Users or staff report AI-generated outputs containing incorrect facts, dates, citations, or references
- AI system cites sources that do not exist or cannot be verified
- AI output contradicts known, verified organisational data or publicly available facts
- Downstream decisions or communications have been made based on unverified AI content
- Monitoring alerts flag high uncertainty scores or outputs outside known knowledge boundaries
- Customer complaints received about incorrect information provided via AI-assisted channels
- Internal audit flags AI-generated content used in regulatory filings or legal documents without verification

## 3. Immediate Response Actions

## PHASE 1 — DETECT & TRIAGE (0–30 mins)

1. Receive and log the hallucination report in the AI Incident Management System
2. Assign severity: P1 if hallucination has caused customer harm, regulatory breach, or financial loss; P2 if identified before external impact; P3 if internal only
3. Identify the specific AI system, model version, and prompt/context that generated the fabricated output
4. Preserve all evidence: prompt inputs, raw AI outputs, model version, timestamp, user session ID
5. Immediately alert the AI Operations Lead and Content/Quality team

## PHASE 2 — CONTAIN (30 mins – 2 hrs)

1. Quarantine and remove the hallucinated content from all affected channels (website, customer comms, internal systems)
2. Issue a hold on any pending communications or decisions that relied on the hallucinated content
3. If the AI system is a customer-facing chatbot or assistant, assess whether to temporarily disable the affected capability
4. Contact affected customers or stakeholders immediately if harm has occurred; provide accurate corrected information
5. Notify Legal if the hallucinated content appeared in contracts, regulatory submissions, or public statements

## PHASE 3 — INVESTIGATE (2–48 hrs)

1. Conduct prompt analysis: identify the query type, context window, and system prompt conditions that triggered hallucination
2. Review model version, temperature settings, and decoding parameters that may have increased hallucination risk
3. Analyse whether the hallucination pattern is isolated or systematic (test additional similar prompts)
4. Assess scope: how widely was the hallucinated content distributed and who acted on it?
5. Determine if any automated downstream processes consumed and acted on the hallucinated output
6. Document root cause: knowledge boundary issue, over-confident model, insufficient retrieval grounding, or prompt design flaw

## PHASE 4 — REMEDIATE (48 hrs – 7 days)

1. Implement prompt engineering improvements: add explicit uncertainty instructions, factual grounding requirements, and source citation mandates
2. Deploy Retrieval-Augmented Generation (RAG) or ground truth verification layer if not already in place
3. Lower model temperature or adjust sampling parameters to reduce hallucination frequency
4. Implement output validation pipeline: fact-checking module, source verification, confidence scoring
5. Update human review requirements for high-risk output categories (legal, medical, financial, regulatory)
6. Retrain or fine-tune model if systematic hallucination in specific domains is confirmed
7. Conduct end-to-end regression testing before restoring full service

## PHASE 5 — RECOVER & CLOSE (7–14 days)

1. Restore full AI service with enhanced controls; monitor closely for 72 hours post-restoration
2. Issue corrected information to all affected parties; document all remediation actions taken
3. Complete the post-incident review and update detection playbook with new hallucination signatures
4. Update AI system documentation, risk register, and model card with hallucination risk profile
5. Report incident to AI Governance Committee; present lessons learned

# Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|
| **Detect & log incident** | AI Ops / User who detected | AI Ops Lead | Legal / Compliance | CTO, CAO, Comms |
| **Quarantine affected content** | Content / Product Team | AI Ops Lead | Legal | Business Owner, CX |
| **Technical root cause analysis** | ML Engineer | AI/ML Lead | Data Science | CTO |
| **Customer/stakeholder notification** | Comms / CX Team | CEO / CAO | Legal | Regulators if required |
| **Remediation implementation** | ML Engineer | AI/ML Lead | QA / Security | CAO, Business Owner |
| **Post-incident review** | AI Governance Team | CAO | All leads | Board if P1 |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to detect hallucination** | <4 hours from occurrence | Incident log timestamp |
| **Time to quarantine affected content** | <2 hours from detection | Content system audit |
| **Hallucination rate post-remediation** | <0.1% of outputs in tested category | Automated evaluation pipeline |
| **Customer complaint rate (AI accuracy)** | <0.05% of interactions | CX system reports |
| **Time to post-incident review** | <5 business days from closure | Governance tracker |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |

## Related Documents

- AI Acceptable Use Policy (AIMP-AUP-001)
- AI Risk Management Policy (AIMP-RMP-006)
- AI Incident Recording and Reporting Policy (AIMP-IRP-003)
- AI System Development Policy (AIMP-SDP-005)
- AI Communication Management Policy (AIMP-CMM-012)

## Approval & Sign-Off

| Chief AI Officer | Chief Technology Officer | Chief Compliance Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Bias & Discrimination Playbook

*Identifying, Containing & Remediating Discriminatory AI Outputs*

AIPB-002 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-002 |
| **Playbook Owner** | Chief AI Ethics Officer / Chief AI Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | CRITICAL — Legal, Regulatory, Reputational & Human Rights Risk |

## 2. Scenario Description

### Overview

AI bias occurs when an AI system produces outputs that systematically disadvantage or discriminate against individuals based on protected characteristics including race, gender, age, disability, religion, national origin, sexual orientation, or socioeconomic status. Discrimination may be direct (explicit bias) or indirect (proxy discrimination through correlated features). High-risk contexts include AI-assisted hiring, credit scoring, insurance pricing, medical triage, law enforcement, and benefits eligibility.

### Detection Indicators / Trigger Conditions

- Fairness metrics (demographic parity, equalised odds) fall below acceptable thresholds
- Audit or external review identifies disparate impact across protected groups
- Employee, customer, or civil society complaint alleges AI-driven discrimination
- Regulatory body inquiry or investigation initiated regarding AI system outcomes
- Internal analysis reveals disproportionate rejection/approval rates correlated with protected attributes
- Media or advocacy group report highlights discriminatory AI outputs
- Legal team receives threat of litigation related to AI-driven decision outcomes

## 3. Immediate Response Actions

## PHASE 1 — DETECT & ESCALATE (0–1 hr)

1. Log the bias complaint or alert in the AI Incident Management System; classify as P1 (regulatory/legal breach) or P2 (internal detection)
2. Immediately notify the Chief AI Ethics Officer, Legal, and Compliance
3. Suspend production use of the affected AI system for the specific use case if there is active potential for ongoing harm
4. Preserve all evidence: model version, training data version, decision outputs, demographic breakdowns, and individual case records
5. Do not alter, delete, or modify any records pending investigation — document preservation order issued by Legal

## PHASE 2 — ASSESS SCOPE (1–24 hrs)

1. Conduct rapid fairness audit: measure demographic parity, equal opportunity, and equalised odds across all relevant protected groups
2. Identify the temporal scope: when did the biased model enter production; how many decisions were affected?
3. Identify affected individuals or groups and assess actual harm caused (denial of service, financial loss, reputational harm)
4. Engage external fairness/bias expert or auditor if internal capability is insufficient for the assessment
5. Assess regulatory notification obligations: does this constitute a reportable discrimination incident?

## PHASE 3 — CONTAIN & NOTIFY (24–72 hrs)

1. Implement human review override for all pending decisions from the affected AI system
2. Notify the Data Protection Officer and Legal of any personal data implications
3. Notify affected individuals in accordance with legal obligations and organisational policy
4. Engage regulators proactively if the organisation has regulatory reporting obligations (e.g. financial services, public sector)
5. Prepare internal communication to leadership and the AI Ethics Committee
6. If discrimination relates to employment decisions, engage HR and Employment Legal immediately

## PHASE 4 — ROOT CAUSE & REMEDIATE (3–30 days)

1. Conduct thorough root cause analysis: training data bias, feature selection, proxy variables, label bias, or model architecture
2. Implement bias mitigation: re-sampling, re-weighting, adversarial debiasing, or post-processing threshold adjustment
3. Re-run comprehensive fairness testing across all protected groups before any re-deployment
4. Review and remediate affected historical decisions where legally and practically possible
5. Implement ongoing bias monitoring with real-time fairness metric dashboards and automated alerts
6. Update the AI Data Bias Assessment process and checklist based on findings

## PHASE 5 — RECOVER & GOVERN (30–90 days)

1. Re-deploy the AI system only after bias remediation is validated by an independent review
2. Report full findings to the AI Ethics Committee and Board
3. Engage legal counsel regarding individual redress for those affected by biased decisions
4. Update the AI Risk Register with bias risk profile; schedule quarterly fairness reviews
5. Publish a transparency report if public trust restoration is required

# Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|
| **Bias detection & initial assessment** | AI Ethics Officer / Fairness Team | CAO / Chief Ethics Officer | Legal, Compliance, DPO | CTO, CEO, Board |
| **Suspend/modify affected AI system** | AI/ML Lead | CTO | Legal, Ethics Officer | Business Owner, CAO |
| **Regulatory notification** | Legal / Compliance | CCO | DPO, CAO | Regulators |
| **Affected individual notification** | Legal / Comms | CEO | HR, DPO | All affected individuals |
| **Bias root cause analysis** | ML Engineer / Data Scientist | AI/ML Lead | Ethics Officer, External Auditor | CTO, CAO |
| **Model remediation & re-deployment** | ML Engineer | AI/ML Lead | QA, Ethics Officer, Legal | CAO, Business Owner |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to suspend biased AI system** | <4 hours from P1 detection | Incident log |
| **Fairness metric improvement post-remediation** | All fairness metrics within defined thresholds | Fairness audit report |
| **Affected decisions remediated** | 100% of adverse biased decisions reviewed | Case management tracker |
| **Time to regulatory notification** | Within regulatory deadline | Legal compliance tracker |
| **Time to post-incident ethics review** | <10 business days | Governance tracker |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |

| | | |
|---|---|---|
| | | |

## Related Documents

- AI Ethics Policy (AIMP-ETH-013)
- AI Data Privacy and Protection Framework (AIMP-DPF-008)
- AI Risk Management Policy (AIMP-RMP-006)
- AI System Data Bias Assessment Checklist (AIDBAC-010)
- AI Incident Recording and Reporting Policy (AIMP-IRP-003)

## Approval & Sign-Off

| Chief AI Ethics Officer | Chief AI Officer | Chief Legal Officer | CEO |
|---|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Data Breach & Privacy Playbook

*Responding to Personal Data Breaches Involving AI Systems & Training Data*

AIPB-003 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-003 |
| **Playbook Owner** | Data Protection Officer / Chief Information Security Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | CRITICAL — Regulatory, Legal & Reputational Risk |

## 2. Scenario Description

### Overview

An AI data breach occurs when personal data used in AI training, inference, or model outputs is unlawfully accessed, disclosed, altered, or destroyed. This includes exposure of training datasets containing personal data, model inversion attacks extracting training data from deployed models, unauthorised access to AI inference logs, and AI systems inadvertently memorising and reproducing personal data in outputs. Regulatory notification timelines (e.g. 72 hours under GDPR, 3 days under PDPA) are mandatory triggers.

### Detection Indicators / Trigger Conditions

- Unauthorised access to AI training data storage, model registry, or inference logs detected
- Model inversion attack suspected — AI outputs appear to reproduce specific personal data from training set
- Data exfiltration alert from AI platform or cloud environment
- Personal data of identifiable individuals appears in AI-generated outputs
- Security scan reveals breach of AI data pipeline or feature store
- Third-party AI vendor notifies of security incident affecting shared data
- Staff report unexpected AI outputs containing personal or confidential information

## 3. Immediate Response Actions

## PHASE 1 — DETECT & DECLARE (0–1 hr)

1. Declare the data breach incident — activate the AI Data Breach Response Team (DPO, CISO, Legal, AI Lead)
2. Start the regulatory notification clock: GDPR — 72 hours to supervisory authority; PDPA — 3 days to PDPC
3. Log all known facts: what data, what AI system, when discovered, how many records, who is affected
4. Issue an immediate access lockdown on the affected AI system, data store, and related infrastructure
5. Preserve forensic evidence: do not alter, patch, or restart affected systems until forensic image is captured

## PHASE 2 — CONTAIN (1–6 hrs)

1. Revoke all API keys, access tokens, and credentials associated with the breached AI system
2. Isolate the affected AI environment from production networks
3. Disable the AI model's public API endpoint if model inversion attack is suspected
4. Coordinate with cloud provider (if applicable) to preserve logs and restrict access
5. Issue internal hold notice — no staff to discuss breach externally pending Legal approval
6. Engage forensic IR team to conduct technical investigation

## PHASE 3 — INVESTIGATE (6–48 hrs)

1. Determine the full scope: which datasets were exposed, what categories of personal data (including any special category data), number of data subjects affected
2. Determine attack vector: external breach, insider threat, misconfiguration, model vulnerability, supply chain
3. Assess whether personal data is memorised in model weights and can be extracted — conduct model extraction testing
4. Review AI inference logs for signs of deliberate personal data extraction via querying
5. Establish whether any third parties received the breached data
6. Assess risk to affected individuals: identity theft, financial loss, discrimination, reputational harm

## PHASE 4 — NOTIFY (24–72 hrs)

1. Submit regulatory notification to the relevant data protection authority within the mandatory timeframe
2. Notify affected data subjects where the breach poses high risk to their rights and freedoms
3. Notify any third-party data processors or controllers who shared the affected data
4. Prepare and issue internal communication to leadership, board, and relevant staff
5. Prepare holding statement for media/public if breach becomes public — approve through Legal and Comms

## PHASE 5 — REMEDIATE & RECOVER (3–30 days)

1. Implement technical remediation: patch vulnerabilities, re-encrypt data, rotate credentials, harden model APIs
2. If model memorisation is confirmed: consider model retraining with differential privacy or data sanitisation; assess model withdrawal
3. Re-validate AI system security posture before returning to production
4. Implement enhanced monitoring: data loss prevention on AI outputs, anomaly detection on inference queries
5. Complete full post-incident review and update the DPIA for the affected AI system

# Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|
| **Breach detection & declaration** | CISO / AI Security Team | DPO / CISO | Legal, AI Lead | CTO, CEO |

| | | | | |
|---|---|---|---|---|
| **Regulatory notification** | DPO / Legal | DPO | CCO, CISO | CEO, Board |
| **Data subject notification** | Legal / Comms | DPO | CX, HR | Affected individuals |
| **Forensic investigation** | CISO / IR Team | CISO | AI Lead, Legal | CTO |
| **AI system remediation** | AI/ML Lead / Security | CTO | CISO, DPO, Legal | CAO, Business Owner |
| **Post-incident DPIA update** | DPO | DPO | AI Lead, Legal | AI Governance Committee |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Regulatory notification timeline** | Within 72 hrs (GDPR) / 3 days (PDPA) | Legal compliance log |
| **Time to contain breach** | <6 hours from declaration | Incident log |
| **Affected records identified** | 100% of records enumerated | Forensic report |
| **Data subject notification rate** | 100% of high-risk subjects notified | Notification tracker |
| **Recurrence within 12 months** | Zero | Security audit |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |
| | | |

## Related Documents

- AI Data Privacy and Protection Framework (AIMP-DPF-008)
- AI Information Security Policy (AIMP-ISP-007)
- AI Incident Recording and Reporting Policy (AIMP-IRP-003)

- Data Backup and Recovery Policy (AIMP-BRP-010)
- AI Risk Management Policy (AIMP-RMP-006)

## Approval & Sign-Off

| Data Protection Officer | Chief Information Security Officer | Chief Legal Officer | CEO |
|---|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# Prompt Injection & Adversarial Attack Playbook

*Detecting & Responding to Deliberate AI Input Manipulation*

AIPB-004 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-004 |
| **Playbook Owner** | Chief Information Security Officer / AI Security Lead |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Security, Data Exfiltration & Integrity Risk |

## 2. Scenario Description

### Overview

Prompt injection attacks occur when malicious inputs are crafted to manipulate an AI language model into ignoring its instructions, producing harmful outputs, revealing system prompts, or executing unintended actions. Adversarial attacks more broadly include adversarial examples (perturbed inputs causing misclassification), model inversion (extracting training data), membership inference (determining if data was in the training set), and evasion attacks (bypassing AI safety filters). These attacks may be direct (user input) or indirect (content processed from external sources by AI agents).

### Detection Indicators / Trigger Conditions

- AI system outputs instructions, system prompts, or internal data not visible to normal users
- AI agent executes unexpected actions, accesses unauthorised resources, or attempts API calls not in scope
- Monitoring flags inputs containing known injection patterns (e.g. 'ignore previous instructions', 'DAN mode')
- AI system bypasses content safety filters and produces prohibited outputs
- Unusual spike in API query volume suggesting automated adversarial probing
- AI classifier or detection system outputs manipulated on borderline security-relevant inputs
- AI system reveals confidential business logic, connection strings, or internal configurations

## 3. Immediate Response Actions

### PHASE 1 — DETECT & ALERT (0–30 mins)

1. Trigger automatic alert via AI security monitoring platform upon injection pattern detection
2. Log the attack: raw input, model output, user session, IP address, timestamp, API endpoint
3. Classify attack type: direct prompt injection, indirect injection (from external content), adversarial example, or jailbreak
4. Immediately notify the AI Security Lead and CISO
5. If AI agent has taken unintended actions (API calls, data access, commands executed), classify as P1

### PHASE 2 — CONTAIN (30 mins – 2 hrs)

1. Block the attacking IP address or user session at the API gateway layer
2. Implement emergency rate limiting and enhanced input validation on the affected endpoint
3. If the AI agent took unintended actions: immediately revoke its access tokens and credentials; audit all actions taken
4. Preserve all attack artifacts for forensic analysis — do not flush logs
5. If system prompt or internal configuration was exposed: treat as a data breach and activate PB03
6. Assess whether the attack vector is exploitable at scale — if yes, consider temporary service suspension

### PHASE 3 — INVESTIGATE (2–24 hrs)

1. Conduct full forensic review of all queries from the attacker over the preceding 30 days
2. Determine if any data exfiltration occurred as a result of the injection
3. Identify the specific vulnerability: insufficient input sanitisation, over-privileged AI agent, weak output filtering, or absent system prompt hardening
4. Test whether the attack vector is reproducible and affects other AI system endpoints
5. Determine whether this is an opportunistic scan or targeted attack on a specific AI capability

### PHASE 4 — REMEDIATE (24 hrs – 14 days)

1. Implement robust input sanitisation and prompt injection detection layer
2. Apply least privilege principles to all AI agent capabilities: restrict API access to minimum required
3. Harden system prompts: separate instruction and data planes; implement instruction hierarchy controls
4. Deploy output filtering and monitoring for sensitive data patterns
5. Implement adversarial training and red-team exercises for AI systems handling sensitive inputs
6. Update WAF rules and API gateway policies to block known injection patterns

### PHASE 5 — HARDEN & CLOSE (14–30 days)

1. Conduct full red-team adversarial testing of all AI APIs before restoring full service
2. Update the AI security threat model with new attack vectors discovered
3. Brief AI development teams on injection-resistant prompt architecture patterns
4. File incident report with AI Security and Governance Committee

## Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| **Attack detection & alerting** | AI Security Monitoring | CISO | AI Security Lead | CTO, CAO |
| **Session/IP blocking** | AI Security / IT Ops | CISO | AI Lead | Operations |
| **Forensic investigation** | CISO / IR Team | CISO | AI Lead, Legal | CTO |
| **Remediation & hardening** | AI/ML Lead / Security | CTO / CISO | QA, Red Team | CAO |
| **Adversarial testing** | Red Team / QA | CISO | AI Lead | CTO, CAO |
| **Incident reporting** | AI Security Team | CISO | Legal, AI Governance | AI Governance Committee |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to detect injection attack** | <5 minutes (automated) | Security monitoring dashboard |
| **Time to block attacker** | <30 minutes from detection | Security incident log |
| **Zero successful data exfiltration** | 0 records exfiltrated | Forensic investigation report |
| **Injection detection rate post-remediation** | ≥99.5% of known patterns blocked | Red team test results |
| **AI agent privilege scope** | Least-privilege validated quarterly | Access review log |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Information Security Policy (AIMP-ISP-007)
- AI Incident Recording and Reporting Policy (AIMP-IRP-003)
- AI Safety Bypass Jailbreak Playbook (AIPB-017)
- AI Data Breach Privacy Playbook (AIPB-003)
- AI Risk Management Policy (AIMP-RMP-006)

## Approval & Sign-Off

| Chief Information Security Officer | Chief AI Officer | Chief Technology Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI System Outage Playbook

*Managing Unplanned Downtime of Production AI Systems*

AIPB-005 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-005 |
| **Playbook Owner** | Head of AI Operations / Chief Technology Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Operational Continuity & Business Impact Risk |

## 2. Scenario Description

### Overview

An AI system outage is any unplanned interruption to the availability of a production AI system, including full service outages, partial functionality degradation, excessive latency breaching SLA thresholds, or AI inference errors affecting more than the defined error rate threshold. AI outages may be caused by infrastructure failures, model serving errors, resource exhaustion, dependency failures, data pipeline issues, or the cascading effects of AI model updates.

### Detection Indicators / Trigger Conditions

- Health check / uptime monitor alerts for AI system endpoint failure or timeout
- API error rate exceeds defined threshold (e.g. >1% error rate over 5-minute window)
- P95 inference latency exceeds SLA threshold
- AI model serving platform reports pod/container crash or OOM errors
- Data pipeline feeding AI inference is failing or producing empty/malformed inputs
- Business process dependent on AI system is unable to proceed
- Support tickets or customer complaints spike regarding AI feature unavailability

## 3. Immediate Response Actions

## PHASE 1 — DETECT & DECLARE (0–10 mins)

1. Automated alert received from monitoring platform — confirm outage is genuine (not monitoring false positive)
2. Declare severity: P1 if customer-facing, revenue-impacting, or safety-critical; P2 if significant internal impact
3. Activate the AI Outage Response Team: AI Ops Lead, ML Engineer, Infrastructure Engineer, Business Owner
4. Post initial incident notification to internal incident channel (Slack/Teams) with known facts
5. Start incident timer — all SLA milestones measured from this point

## PHASE 2 — DIAGNOSE (10–30 mins)

1. Check model serving platform status: pod/container health, resource utilisation (CPU, GPU, memory), error logs
2. Check upstream data pipeline status: is inference data arriving? Schema validation passing?
3. Check downstream dependencies: database connectivity, external API calls from the AI system
4. Review recent changes: was a model version update, infrastructure change, or config change deployed in the last 24 hours?
5. Check cloud provider status page for regional incidents affecting AI infrastructure
6. Isolate the failure layer: model serving, data ingestion, infrastructure, or external dependency

## PHASE 3 — MITIGATE (30 mins – 2 hrs)

1. If recent model/config change is suspected cause: initiate rollback to last known stable version immediately
2. If infrastructure resource exhaustion: scale up compute resources; check auto-scaling configuration
3. If data pipeline failure: activate fallback data source or switch to cached/batch inference mode
4. If external dependency failure: activate degraded mode or manual fallback process for affected business function
5. Communicate ETA update to business stakeholders and, if required, to customers

## PHASE 4 — RESTORE & VALIDATE (2–4 hrs)

1. Deploy fix or restored service configuration to staging; run automated smoke tests
2. Promote restored service to production; monitor error rates and latency for 30 minutes
3. Confirm inference quality is consistent with pre-outage baseline (run inference validation tests)
4. Issue all-clear notification to business teams and customers when service is confirmed stable
5. Document full incident timeline, impact, and resolution in the incident management system

## PHASE 5 — POST-INCIDENT (Within 5 days)

1. Conduct post-incident review: root cause, detection gap, response effectiveness, prevention measures
2. Update runbooks with new diagnosis steps, workarounds, or escalation paths discovered during the incident
3. Review monitoring coverage gaps and implement additional alerting to enable earlier detection
4. Update the Business Continuity Plan if AI outage duration exceeded BCP assumptions

# Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|
| **Outage detection & alerting** | Monitoring System / AI Ops | AI Ops Lead | ML Eng, Infra | Business Owner, CTO |

| | | | | |
|---|---|---|---|---|
| **Diagnosis & root cause** | ML Engineer / Infra Eng | AI Ops Lead | AI Lead | CTO |
| **Mitigation & rollback** | ML Engineer / Infra Eng | AI Ops Lead | Business Owner | CTO, Comms |
| **Production restoration** | ML Engineer / Infra Eng | CTO | QA, Business Owner | All stakeholders |
| **Post-incident review** | AI Ops / AI Lead | CTO | All responders | AI Governance Committee |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Mean Time to Detect (MTTD)** | <5 minutes | Monitoring alert log |
| **Mean Time to Respond (MTTR)** | <30 minutes | Incident log |
| **Mean Time to Restore (MTTR)** | <2 hours (P1); <4 hours (P2) | Incident log |
| **System availability (monthly)** | ≥99.9% (P1 systems) | Uptime monitoring |
| **Incidents with rollback tested before outage** | 100% | Change management log |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Change Management Policy (AIMP-CMP-011)
- Data Backup and Recovery Policy (AIMP-BRP-010)
- AI Performance Monitoring Plan (AIMON-004)

- AI Risk Management Policy (AIMP-RMP-006)
- AI Incident Recording and Reporting Policy (AIMP-IRP-003)

## Approval & Sign-Off

| Head of AI Operations | Chief Technology Officer | Business Owner |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# Model Drift & Performance Degradation Playbook

*Detecting, Diagnosing & Remediating AI Model Drift in Production*

AIPB-006 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-006 |
| **Playbook Owner** | AI/ML Lead / Head of AI Operations |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | MEDIUM–HIGH — Operational Quality & Decision Integrity Risk |

## 2. Scenario Description

### Overview

Model drift occurs when an AI model's production performance degrades over time due to changes in the statistical properties of input data (data drift) or changes in the relationship between inputs and outputs (concept drift). Drift can cause a model to produce increasingly inaccurate, outdated, or harmful decisions without any visible system failure. Left undetected, drift can cause financial losses, biased outcomes, safety failures, and erosion of business stakeholder trust.

### Detection Indicators / Trigger Conditions

- Automated drift monitoring alerts: PSI (Population Stability Index) or KL divergence exceeds threshold
- Model accuracy, precision, recall, or F1 score drops below defined baseline threshold
- Business KPIs dependent on AI predictions degrade (e.g. conversion rate, fraud catch rate, recommendation CTR)
- Distribution of model output scores shifts significantly from baseline
- Feature importance rankings change markedly from training baseline
- Data quality reports flag schema changes, new null patterns, or distribution shifts in input features
- User or business team feedback reports AI recommendations feel less relevant or less accurate

## 3. Immediate Response Actions

### PHASE 1 — DETECT & ASSESS (0–4 hrs)

1. Alert received from drift monitoring platform or performance dashboard
2. Confirm drift is genuine: review raw drift metrics, check for data pipeline anomalies that could mimic drift
3. Classify drift type: data drift (input distribution change), concept drift (relationship change), or both
4. Quantify performance impact: compare current metrics vs. baseline across all relevant KPIs
5. Assess business impact: are decisions made by the drifted model causing material errors?

### PHASE 2 — DIAGNOSE (4–24 hrs)

1. Analyse which features have drifted most significantly using feature-level drift reports
2. Compare recent training data distribution vs. current inference data distribution
3. Review external factors that may explain the drift: seasonality, market changes, product changes, world events
4. Assess whether the concept drift is permanent (fundamental change) or temporary (transient pattern)
5. Determine whether immediate model replacement or short-term mitigation is required

### PHASE 3 — MITIGATE (24–72 hrs)

1. If performance is critically degraded: switch to fallback model or rule-based system for the affected decision area
2. Implement temporary threshold adjustments to reduce harmful decision errors while retraining is underway
3. Increase human review rate for decisions in the affected area
4. Communicate impact and mitigation plan to business stakeholders
5. Begin data collection for model retraining with recent, representative data

### PHASE 4 — RETRAIN & VALIDATE (3–14 days)

1. Retrain the model using recent data that reflects the new distribution
2. Complete full validation pipeline: performance, fairness, robustness, and explainability testing
3. Compare retrained model against drifted production model and original baseline across all metrics
4. Conduct A/B testing in staging environment before promoting to production
5. Log the retraining event in the model registry with full metadata

### PHASE 5 — DEPLOY & MONITOR (14–30 days)

1. Deploy retrained model via the approved change management process (AIMP-CMP-011)
2. Implement enhanced monitoring for 30 days post-retraining to confirm drift has been corrected
3. Update drift detection thresholds and retraining criteria based on lessons from this incident
4. Update AI Performance Monitoring Plan with refined drift detection parameters
5. Report incident and retraining to AI Governance Committee

## Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|
| **Drift detection & alerting** | Monitoring System / AI Ops | AI Ops Lead | ML Engineer | Business Owner, CTO |

| | | | | |
|---|---|---|---|---|
| **Drift diagnosis** | ML Engineer / Data Scientist | AI/ML Lead | Data Engineer | AI Ops Lead |
| **Mitigation & fallback** | ML Engineer | AI Ops Lead | Business Owner | CTO, Comms |
| **Model retraining & validation** | ML Engineer / Data Scientist | AI/ML Lead | QA, Ethics Officer | CAO |
| **Production deployment** | ML Engineer | CTO via Change Board | AI Lead, QA | Business Owner |
| **Post-incident monitoring** | AI Ops | AI Ops Lead | ML Engineer | Governance Committee |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Drift detection time** | <24 hours from onset | Drift monitoring dashboard |
| **Model performance restoration** | Within 14 days of drift confirmation | Model performance report |
| **Retraining frequency (high-risk models)** | At least quarterly proactive | Model registry log |
| **Fairness metrics stable post-retraining** | All within defined thresholds | Fairness audit |
| **Business KPI recovery** | Return to baseline within 30 days | Business metrics dashboard |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Performance Monitoring Plan (AIMON-004)

- AI System Development Policy (AIMP-SDP-005)
- AI Change Management Policy (AIMP-CMP-011)
- AI Risk Management Policy (AIMP-RMP-006)
- Model Drift Performance Playbook (self-reference for updates)

## Approval & Sign-Off

| AI/ML Lead | Head of AI Operations | Chief Technology Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# Shadow AI & Unauthorised Use Playbook

*Detecting & Managing Unsanctioned AI Tool Usage Across the Organisation*

AIPB-007 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-007 |
| **Playbook Owner** | Chief AI Officer / Chief Information Security Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | MEDIUM–HIGH — Data Leakage, Compliance & Governance Risk |

## 2. Scenario Description

### Overview

Shadow AI refers to the use of AI tools, services, or models by employees or departments without formal organisational approval, bypassing security review, data classification controls, and governance processes. Risks include personal data or confidential business information being uploaded to unapproved AI services, AI-generated outputs being used in business decisions without appropriate validation, regulatory compliance failures, and intellectual property exposure. Shadow AI may be driven by productivity pressure, unawareness of the approval process, or dissatisfaction with approved tools.

### Detection Indicators / Trigger Conditions

- Network monitoring detects unusual data transfers to known AI platform domains (e.g. OpenAI, Midjourney, Gemini) from unapproved endpoints
- DLP alerts flag confidential data being uploaded to unapproved cloud AI services
- Staff survey or IT helpdesk reveals widespread use of unapproved AI tools for work tasks
- AI-generated content identified in official outputs, reports, or customer communications without proper disclosure
- Business unit procurement records show AI tool subscriptions outside IT governance process
- HR or line manager reports staff admitting to using AI tools outside the Approved AI Tools Register
- Security audit finds sensitive data in AI service provider logs or terms of service exposure

## 3. Immediate Response Actions

### PHASE 1 — DETECT & SCOPE (0–24 hrs)

1. Log the shadow AI detection event; classify severity based on data sensitivity exposed
2. Identify the AI tools being used: which services, what data was submitted, by whom, for how long
3. Use DLP logs, network traffic analysis, and browser history (where permitted) to scope usage
4. Assess data exposure: what categories of data were submitted? Personal data? Confidential business data? IP?
5. Identify whether multiple users or a single user is involved — distinguish systemic from individual behaviour

### PHASE 2 — CONTAIN (24–48 hrs)

1. Block unapproved AI service domains at the network/DNS layer for the affected user group or organisation-wide
2. Issue immediate cease-and-desist instruction to identified shadow AI users via their line manager and HR
3. Assess whether submitted data constitutes a personal data breach — if yes, activate PB03
4. Contact the AI vendor to request deletion of any submitted data (where contractually possible)
5. Preserve evidence: DLP logs, network logs, and any documented AI outputs used in business decisions

### PHASE 3 — INVESTIGATE & ASSESS (48 hrs – 7 days)

1. Conduct structured interviews with affected staff members — approach is educational, not purely punitive
2. Identify root cause: was it awareness gap, process friction, tool gap, or deliberate circumvention?
3. Review all business outputs that may have incorporated shadow AI-generated content
4. Assess regulatory exposure: if personal data was uploaded, is notification required?
5. Determine whether shadow AI usage has affected any customer deliverables, contracts, or regulated processes

### PHASE 4 — REMEDIATE (1–4 weeks)

1. Accelerate AI tool approval process for tools identified as meeting genuine business need
2. Implement technical controls: DNS blocking, DLP rules, browser extension controls for unapproved AI services
3. Roll out mandatory Shadow AI Awareness training to all affected departments
4. Establish a fast-track AI tool request process (SLA: 5 business days for low-risk tools)
5. Apply proportionate disciplinary action for deliberate or repeated policy breaches
6. Update the Approved AI Tools Register and publish to all staff

### PHASE 5 — GOVERN & PREVENT (Ongoing)

1. Implement quarterly shadow AI scanning across the organisation as a standard control
2. Include shadow AI risks in the annual AI risk assessment
3. Maintain an AI tool demand signal process — staff can formally request new AI tools for evaluation
4. Report shadow AI incident statistics quarterly to the AI Governance Committee

## Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|
| **Shadow AI detection** | DLP / Security Monitoring | CISO | IT, HR | CAO, CTO |

| | | | | |
|---|---|---|---|---|
| **Network blocking** | IT Security | CISO | IT Ops | CTO |
| **Staff investigation** | HR / Line Manager | CPO / CISO | Legal, CAO | CEO |
| **Data exposure assessment** | DPO / Legal | DPO | CISO, CAO | Regulators if required |
| **Remediation & controls** | IT Security / AI Ops | CTO / CISO | CAO, HR | All staff |
| **Policy & awareness update** | L&D / AI Governance | CAO | CPO, CISO | All staff |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Shadow AI tools detected per quarter** | Trending toward zero | Quarterly DLP scan |
| **Time to block unapproved AI domain** | <24 hours from detection | Network block log |
| **AI tool request processing time** | <5 business days (low risk) | IT governance tracker |
| **Shadow AI awareness training completion** | 100% of flagged staff within 30 days | LMS records |
| **Staff awareness of Approved AI Tools Register** | ≥90% in annual survey | Staff survey |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Tool Usage Policy (AIMP-TUP-002)

- AI Acceptable Use Policy (AIMP-AUP-001)
- AI Training & Awareness Policy (AIMP-TAP-004)
- AI Information Security Policy (AIMP-ISP-007)
- AI Data Privacy and Protection Framework (AIMP-DPF-008)

## Approval & Sign-Off

| Chief AI Officer | Chief Information Security Officer | Chief People Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# Third-Party AI Vendor Incident Playbook

*Managing Incidents Arising from or Involving Third-Party AI Vendors & Providers*

AIPB-008 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-008 |
| **Playbook Owner** | Chief AI Officer / Head of Procurement & Vendor Management |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Operational, Legal & Supply Chain Risk |

## 2. Scenario Description

### Overview

Third-party AI vendor incidents include: AI vendor service outages affecting the organisation's operations, security breaches at a vendor that expose the organisation's data, vendor AI model changes causing unexpected behaviour in integrated systems, vendor ceasing operations or withdrawing a critical AI service, vendor non-compliance with contractual or regulatory obligations, and material degradation of vendor AI model quality. The organisation's dependence on third-party AI services creates external risk vectors outside direct operational control.

### Detection Indicators / Trigger Conditions

- Third-party AI vendor notifies of security incident, data breach, or service disruption
- Vendor AI API begins returning errors, unexpected outputs, or degraded performance
- Vendor makes undisclosed changes to model, pricing, or terms of service
- Business intelligence reveals vendor financial instability or cessation of operations
- Regulatory action taken against AI vendor affecting their compliance posture
- Vendor fails to respond to data subject rights requests or contractual obligations
- Internal monitoring detects unexpected behaviour in AI systems powered by vendor APIs

## 3. Immediate Response Actions

## PHASE 1 — IDENTIFY & CLASSIFY (0–2 hrs)

1. Receive and log the vendor incident notification or internally-detected vendor issue
2. Classify incident type: security breach, service outage, model change, compliance failure, or vendor viability risk
3. Identify all internal AI systems and business processes dependent on the affected vendor
4. Engage the vendor's account management / incident response contact immediately
5. Assess whether contractual SLAs and notification obligations have been met by the vendor

## PHASE 2 — CONTAIN & MITIGATE (2–24 hrs)

1. If vendor security breach: assess data exposure scope; activate PB03 if personal data is involved
2. If vendor service outage: activate AI system fallback procedures or manual process overrides
3. If model behaviour change: evaluate impact on AI system performance and safety; consider temporary suspension
4. Engage Legal to review contractual remedies, SLA credits, and indemnification provisions
5. Preserve all vendor communications and documentation as potential evidence

## PHASE 3 — INVESTIGATE & PRESSURE (24–72 hrs)

1. Require the vendor to provide a full incident report including root cause, scope, and remediation timeline
2. Conduct independent assessment of the incident's impact on the organisation
3. Review vendor's security certifications and compliance posture post-incident
4. Assess whether the incident represents a systemic vendor risk requiring exit strategy activation
5. Engage with peer organisations or industry groups to understand if the incident is broader

## PHASE 4 — REMEDIATE & GOVERN (1–30 days)

1. Implement interim technical mitigations while vendor remediation is underway
2. Update vendor risk assessment and due diligence record
3. Trigger vendor security review under the procurement framework
4. If vendor viability is at risk: activate vendor exit strategy; identify and evaluate alternative providers
5. Negotiate enhanced contractual protections, audit rights, or SLA improvements with the vendor

## PHASE 5 — CLOSE & LEARN (30–90 days)

1. Update the Third-Party AI Vendor Risk Register with incident findings
2. Review and update vendor selection and due diligence criteria based on the incident
3. Present vendor incident report and lessons learned to the AI Governance Committee
4. Consider reducing single-vendor concentration risk through multi-vendor strategy

# Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|
| **Vendor incident detection** | AI Ops / Vendor Monitor | AI Ops Lead | Legal, Procurement | CTO, CAO |
| **Vendor engagement & SLA enforcement** | Procurement / Legal | CAO | CTO, Legal | CEO |
| **Data breach assessment** | DPO / CISO | DPO | Legal, CAO | Regulators if needed |

| | | | | |
|---|---|---|---|---|
| **Fallback & mitigation** | AI/ML Lead / IT Ops | CTO | Business Owner | All stakeholders |
| **Vendor risk re-assessment** | Procurement / AI Governance | CAO | Legal, CISO | AI Governance Committee |
| **Vendor exit (if required)** | CTO / Procurement | CEO | Legal, Finance, CAO | Board |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Vendor notification compliance rate** | 100% of incidents notified per contract | Contract compliance log |
| **Time to vendor fallback activation** | <4 hours from P1 detection | Incident log |
| **Third-party AI vendor risk assessments completed** | 100% annually | Vendor risk register |
| **Vendor concentration risk (single vendor dependency)** | No critical function 100% dependent on one vendor | Architecture review |
| **Exit strategy in place for Tier 1 AI vendors** | 100% | Vendor management tracker |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |
| | | |

## Related Documents

- AI Tool Usage Policy (AIMP-TUP-002)
- AI Risk Management Policy (AIMP-RMP-006)
- AI Information Security Policy (AIMP-ISP-007)

- AI Data Breach Privacy Playbook (AIPB-003)
- AI Supply Chain Compromise Playbook (AIPB-019)

## Approval & Sign-Off

| Chief AI Officer | Head of Procurement | Chief Information Security Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# High-Stakes AI Decision Error Playbook

*Responding to Consequential Errors in AI-Assisted Decision Making*

AIPB-009 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-009 |
| **Playbook Owner** | Chief AI Officer / Chief Risk Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | CRITICAL — Human Impact, Legal & Regulatory Risk |

## 2. Scenario Description

### Overview

High-stakes AI decision errors occur when AI systems produce incorrect outputs that result in material harm to individuals or organisations in consequential domains such as: employment decisions (wrongful termination, discriminatory hiring), financial services (incorrect credit denial, fraud misclassification), healthcare (incorrect triage, wrong treatment recommendation), criminal justice (wrong risk score, misidentification), benefits eligibility (wrongful denial of critical services), or safety-critical systems (navigation, infrastructure control). These errors may result from model failure, data quality issues, incorrect use of AI output by humans, or insufficient human oversight.

### Detection Indicators / Trigger Conditions

- Individual or their representative challenges an AI-assisted decision causing material harm
- Internal audit identifies AI decisions that deviated from correct outcomes at significant rate
- AI model performance review reveals systemic error pattern in high-stakes decision category
- Regulatory complaint or legal claim received alleging harmful AI decision
- Human reviewer identifies AI recommendation was incorrect and harm has already occurred
- Media or advocacy report highlights harmful AI decision made by the organisation
- Clinical, legal, or domain expert review contradicts AI-generated recommendation that was acted upon

## 3. Immediate Response Actions

## PHASE 1 — DECLARE & PROTECT (0–2 hrs)

1. Immediately suspend the AI system's role in affected decision category pending investigation
2. Implement mandatory human review for all pending decisions in the affected category
3. Log the incident as P1; notify the Chief AI Officer, Chief Risk Officer, and Legal
4. Identify all individuals who may have been harmed by the erroneous decision(s)
5. Issue a document preservation notice — all records relating to affected AI decisions must be retained
6. Contact the affected individual(s) to acknowledge the concern and commit to urgent review

## PHASE 2 — INVESTIGATE (2–72 hrs)

1. Conduct a full case review of the specific decision(s) in question: input data, model output, human review process, and final decision
2. Determine whether the error is isolated or systemic (audit a sample of similar decisions)
3. Assess whether the AI system operated as designed — or whether there was a model failure, data error, or process failure
4. Engage domain experts (medical, legal, financial) to assess the nature and extent of harm caused
5. Identify whether human oversight processes failed — was the AI output reviewed appropriately?
6. Assess potential regulatory exposure and legal liability

## PHASE 3 — REMEDIATE HARM (72 hrs – 30 days)

1. Where the decision can be reversed: take immediate corrective action and communicate to the affected individual
2. Where the decision cannot be reversed: assess and offer appropriate remedies (financial, practical, reputational)
3. Conduct human expert review of all affected decisions in the relevant time period
4. Engage Legal to assess claims exposure; initiate dispute resolution where appropriate
5. Notify regulators if the error constitutes a reportable incident under applicable sector regulation

## PHASE 4 — FIX THE SYSTEM (30–90 days)

1. Conduct full technical root cause analysis of model failure or misuse
2. Implement model improvements, retraining, or replacement as required
3. Strengthen human oversight controls: mandatory secondary review for high-stakes decisions, clearer escalation criteria
4. Update AI system risk assessment and ethics review to reflect lessons learned
5. Implement audit logging for all high-stakes AI decisions to enable future review

## PHASE 5 — GOVERN & REPORT (90+ days)

1. Report full incident findings, remediation actions, and governance improvements to the Board
2. Update the organisation's AI ethics and risk framework based on lessons learned
3. Engage external auditor to validate that similar errors are unlikely to recur
4. Publish transparency report if public trust requires restoration

# Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| **Suspend AI decision system** | AI/ML Lead | CTO / CAO | Legal, Business Owner | CRO, CEO |
| **Individual harm assessment** | Domain Expert / Legal | CRO | DPO, Ethics Officer | CAO, CEO |
| **Decision audit** | AI/ML Lead / Risk Team | CRO | Ethics Officer, Legal | Board |
| **Individual remediation** | Legal / Business Owner | CEO | HR (if employment), DPO | Affected individuals |
| **Regulatory notification** | Legal / Compliance | CCO | DPO, CAO | Regulators |
| **System remediation** | ML Engineer | AI/ML Lead | QA, Ethics, Legal | CAO, CTO, Board |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to suspend AI decision capability** | <2 hours from P1 classification | Incident log |
| **Affected decisions reviewed** | 100% within 30 days | Audit tracker |
| **Individual remediation completion** | 100% of identified affected individuals | Case management system |
| **Regulatory notification compliance** | Within regulatory deadline | Legal compliance log |
| **Recurrence of high-stakes AI errors** | Zero within 12 months post-remediation | AI audit report |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |
| | | |

## Related Documents

- AI Ethics Policy (AIMP-ETH-013)
- AI Risk Management Policy (AIMP-RMP-006)
- AI Bias Discrimination Playbook (AIPB-002)
- AI Communication Management Policy (AIMP-CMM-012)
- AI Incident Recording and Reporting Policy (AIMP-IRP-003)

## Approval & Sign-Off

| Chief AI Officer | Chief Risk Officer | Chief Legal Officer | CEO |
|---|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Regulatory Compliance Breach Playbook

*Responding to Non-Compliance with AI-Applicable Laws & Regulations*

AIPB-010 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-010 |
| **Playbook Owner** | Chief Compliance Officer / Chief AI Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | CRITICAL — Regulatory, Legal & Reputational Risk |

## 2. Scenario Description

### Overview

An AI regulatory compliance breach occurs when an AI system, or its use, violates applicable laws and regulations including data protection law (PDPA, GDPR), sector-specific AI regulations (EU AI Act, financial services AI rules), anti-discrimination law, consumer protection regulations, or emerging national AI governance frameworks. Breaches may be identified through internal audit, regulatory investigation, whistleblower disclosure, or external complaint. Consequences may include regulatory fines, enforcement action, mandatory system shutdown, and reputational damage.

### Detection Indicators / Trigger Conditions

- Regulatory body initiates investigation or requests information regarding an AI system
- Internal compliance audit identifies AI system operating outside regulatory requirements
- AI system identified as 'high-risk' under EU AI Act or equivalent framework without required conformity assessment
- Data protection authority receives complaint about AI system's personal data processing
- AI system found to make automated decisions without required disclosure or human review option
- Whistleblower or staff raises concern about AI regulatory non-compliance
- Legal review identifies AI system lacks required documentation, transparency measures, or safeguards

## 3. Immediate Response Actions

### PHASE 1 — IDENTIFY & ESCALATE (0–4 hrs)

1. Log the compliance breach alert/notification in the risk and compliance management system
2. Immediately notify the Chief Compliance Officer, Legal, and Chief AI Officer
3. If a regulatory authority is involved: appoint a single regulatory liaison and do not communicate with the regulator until Legal has reviewed
4. Issue a document preservation order for all records relating to the AI system and its outputs
5. Assess whether the breach requires immediate suspension of the AI system to prevent ongoing non-compliance

### PHASE 2 — SCOPE & ASSESS (4–48 hrs)

1. Conduct a rapid regulatory compliance review of the AI system against all applicable regulatory requirements
2. Identify the specific regulatory provision(s) breached, the duration of non-compliance, and all affected individuals or transactions
3. Engage external regulatory counsel with AI expertise to advise on regulatory risk and strategy
4. Assess mandatory notification obligations: does the breach require self-reporting to the regulator?
5. Determine whether the breach is isolated to one AI system or reflects a systemic compliance gap

### PHASE 3 — NOTIFY & ENGAGE (48 hrs – 14 days)

1. Submit any mandatory regulatory notifications within the required timeframe
2. Prepare a formal response to any regulatory inquiry — all responses must be approved by Legal before submission
3. Engage with the regulator constructively and transparently — demonstrate proactive remediation commitment
4. Notify affected individuals of their rights where legally required
5. Prepare board briefing on regulatory exposure and remediation plan

### PHASE 4 — REMEDIATE (14–90 days)

1. Implement all required technical and organisational measures to bring the AI system into compliance
2. Complete any required conformity assessments, registrations, or approvals
3. Implement or strengthen required governance controls: documentation, human oversight, explainability
4. Conduct compliance verification testing before returning the AI system to full operation
5. Update the AI compliance register and risk register

### PHASE 5 — CLOSE & EMBED (90+ days)

1. Conduct regulatory compliance review of all AI systems to identify similar gaps
2. Update the AI regulatory compliance monitoring programme
3. Report outcomes and lessons learned to the Board and AI Governance Committee
4. Implement proactive regulatory horizon scanning to identify emerging AI regulatory requirements

## Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| **Breach identification & escalation** | Compliance / Legal / Audit | CCO | CAO, DPO, Legal | CEO, Board |
| **Regulatory liaison management** | Legal / CCO | CCO | External Counsel | CEO |
| **AI system suspension (if required)** | AI/ML Lead | CTO | Legal, CCO | CAO |
| **Regulatory notification** | Legal | CCO / DPO | External Counsel | Regulator |
| **Technical remediation** | ML Engineer / IT | CTO | Legal, CCO | CAO |
| **Compliance verification** | Compliance / External Auditor | CCO | CAO, Legal | Board |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Regulatory notification timeliness** | 100% within regulatory deadline | Legal compliance log |
| **AI regulatory gap closure** | 100% within agreed remediation plan | Compliance tracker |
| **AI regulatory compliance audit coverage** | All Tier 1 AI systems annually | Audit programme |
| **Regulatory fines/enforcement actions** | Zero in rolling 12 months | Legal log |
| **Regulatory horizon scanning cadence** | Monthly review | Compliance programme log |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Risk Management Policy (AIMP-RMP-006)

- AI Ethics Policy (AIMP-ETH-013)
- AI Data Privacy and Protection Framework (AIMP-DPF-008)
- AI Communication Management Policy (AIMP-CMM-012)
- AI System Development Policy (AIMP-SDP-005)

## Approval & Sign-Off

| Chief Compliance Officer | Chief AI Officer | Chief Legal Officer | CEO |
|---|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Training Data Poisoning Playbook

*Detecting & Responding to Deliberate Corruption of AI Training Data*

AIPB-011 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-011 |
| **Playbook Owner** | Chief Information Security Officer / AI/ML Lead |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | CRITICAL — Model Integrity, Security & Safety Risk |

## 2. Scenario Description

### Overview

Training data poisoning is a deliberate attack where malicious actors introduce corrupted, mislabelled, or adversarial data into the AI training pipeline to manipulate model behaviour. Attacks may be targeted (causing specific misclassifications on particular inputs — a backdoor attack) or indiscriminate (degrading overall model performance). Poisoning may be executed by external attackers who compromise data sources, malicious insiders with dataset access, or through supply chain compromise of third-party datasets. Detection is difficult as poisoned models may appear to perform normally on standard test sets.

### Detection Indicators / Trigger Conditions

- Anomalous labelling patterns detected in training data quality checks
- Model performance on held-out test set unexpectedly diverges from training performance
- Model produces correct outputs generally but systematically fails on specific trigger inputs
- Data pipeline access logs show unexpected modifications to training datasets
- Third-party dataset provider reports compromise or contamination of data feeds
- AI security red team identifies backdoor trigger patterns in deployed model
- Insider threat investigation reveals unauthorised modification of training data

## 3. Immediate Response Actions

## PHASE 1 — SUSPECT & ISOLATE (0–2 hrs)

1. Immediately suspend any model training jobs using potentially compromised datasets
2. Freeze and quarantine the suspected poisoned dataset — no reads or writes without forensic authorisation
3. Isolate the affected AI training environment from production systems
4. Notify the CISO, AI/ML Lead, and Chief AI Officer
5. Preserve all data pipeline access logs, training job metadata, and dataset version history
6. Assess whether any model trained on the poisoned data is currently deployed in production

## PHASE 2 — INVESTIGATE (2–48 hrs)

1. Audit all data pipeline access logs for the affected time window — identify who accessed or modified the dataset
2. Conduct statistical analysis of the training dataset: anomaly detection, label inconsistency scanning, duplicate detection
3. Test the deployed model for backdoor triggers: probe with potential trigger patterns to assess if poisoning has affected outputs
4. Determine attack scope: how many training records are affected and across what time period?
5. Identify attack vector: external compromise, insider, third-party data supply chain, or automated pipeline vulnerability

## PHASE 3 — CONTAIN (48–72 hrs)

1. If deployed model is confirmed to contain backdoor: immediately remove from production and activate fallback
2. Revoke all access to training data repositories for investigation period; re-issue access on least-privilege basis after review
3. Patch the vulnerability that allowed data poisoning access
4. Begin reconstruction of clean training dataset from last known good version
5. Notify Legal and Compliance — assess whether poisoning constitutes a reportable security incident

## PHASE 4 — REMEDIATE & REBUILD (1–30 days)

1. Reconstruct clean training dataset from verified clean data sources
2. Retrain affected AI models using the clean dataset and enhanced data validation pipeline
3. Implement automated data integrity verification: hash verification, schema validation, statistical anomaly detection
4. Conduct adversarial testing (backdoor testing) on the retrained model before production deployment
5. Implement cryptographic signing for all training data to detect future tampering

## PHASE 5 — HARDEN & PREVENT (30–90 days)

1. Implement strict role-based access control for all training data repositories with separation of duties
2. Deploy continuous data integrity monitoring for all active training pipelines
3. Establish a secure, audited pipeline for onboarding third-party datasets
4. Incorporate data poisoning detection as a standard phase in the AI development lifecycle
5. Conduct threat modelling exercise for all AI training pipelines

# Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|
| **Suspend training & quarantine data** | AI/ML Lead / IT Security | CISO | AI Lead, Legal | CTO, CAO |
| **Forensic investigation** | CISO / IR Team | CISO | AI Lead, Legal | CTO |
| **Remove poisoned model from production** | AI/ML Lead | CTO | CISO, Business Owner | CAO |
| **Clean dataset reconstruction** | Data Engineer / ML Engineer | AI/ML Lead | CISO | CTO |
| **Retrain & validate clean model** | ML Engineer | AI/ML Lead | QA, CISO | CAO, CTO |
| **Pipeline hardening** | CISO / Data Engineer | CTO | AI Lead | AI Governance Committee |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to isolate poisoned dataset** | <2 hours from detection | Incident log |
| **Backdoor detection rate in testing** | ≥99% of known backdoor patterns | Red team test results |
| **Data integrity check coverage** | 100% of training datasets | Pipeline audit |
| **Time to retrain and deploy clean model** | <30 days from confirmed poisoning | Model registry log |
| **Recurrence of data poisoning** | Zero within 12 months | Security audit |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Information Security Policy (AIMP-ISP-007)
- AI System Development Policy (AIMP-SDP-005)
- AI Data Breach Privacy Playbook (AIPB-003)
- AI Training Data Poisoning Playbook (self)
- AI Risk Management Policy (AIMP-RMP-006)

## Approval & Sign-Off

| Chief Information Security Officer | AI/ML Lead | Chief Technology Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Model Theft & IP Exfiltration Playbook

*Detecting & Responding to Theft of AI Model Intellectual Property*

AIPB-012 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-012 |
| **Playbook Owner** | Chief Information Security Officer / Chief AI Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Intellectual Property, Competitive & Legal Risk |

## 2. Scenario Description

### Overview

AI model theft involves the unauthorised extraction, copying, or reverse-engineering of proprietary AI models, training data, or related intellectual property. Attacks include model extraction (recreating a model's functionality through repeated API queries), direct theft of model weights from storage, insider exfiltration, and supply chain theft. AI models represent substantial investment and competitive advantage; their theft may enable competitors to replicate capabilities, expose training data characteristics, or undermine the organisation's IP position.

### Detection Indicators / Trigger Conditions

- Abnormally high volume of API queries from a single source — potential model extraction attack
- Security monitoring detects large file transfers involving model weight files from AI infrastructure
- DLP alert flags model artefacts being transmitted to external destinations
- Model registry audit log shows unauthorised access or download of model artefacts
- Former employee or contractor suspected of taking model artefacts on departure
- Competitor AI product launches with suspiciously similar performance characteristics to internal models
- Dark web or grey market intelligence suggests proprietary model artefacts are being sold or shared

## 3. Immediate Response Actions

## PHASE 1 — DETECT & ALERT (0–1 hr)

1. Trigger security alert; classify incident as P1 (confirmed model theft) or P2 (suspected)
2. Immediately revoke API keys and access tokens associated with the suspected exfiltration source
3. Implement emergency rate limiting and block the source IP/user on all AI API endpoints
4. Preserve all evidence: API query logs, access logs, data transfer records, model registry audit trail
5. Notify the CISO, CAO, Legal, and if insider threat is suspected, HR and the CPO

## PHASE 2 — INVESTIGATE (1–48 hrs)

1. Conduct forensic analysis of all API queries from the suspected attacker — reconstruct query sequence and data obtained
2. Analyse whether the query pattern is consistent with model extraction (systematic boundary probing, functional equivalence testing)
3. Review model registry access logs and DLP data transfer logs for direct file exfiltration evidence
4. If insider theft: conduct forensic investigation of the individual's device, email, cloud storage, and USB transfers
5. Quantify what has been extracted: model weights, training data, hyperparameters, proprietary code?
6. Engage external IP and cybersecurity forensics counsel

## PHASE 3 — CONTAIN & PROTECT (48 hrs – 7 days)

1. Apply rate limiting and query diversity controls to all production AI APIs to limit further extraction
2. Implement output perturbation (differential privacy noise) to reduce model extraction fidelity
3. Review and tighten model registry access controls — implement immutable audit logging
4. If model was extracted through direct file theft: consider watermarking all future model releases for traceability
5. Engage legal counsel to pursue available remedies: cease and desist, injunctive relief, civil and criminal proceedings

## PHASE 4 — REMEDIATE & HARDEN (7–30 days)

1. Re-key all model artefacts and rotate all signing certificates
2. Implement model watermarking and fingerprinting to enable detection of stolen models in the wild
3. Deploy API anomaly detection specifically targeting model extraction patterns
4. Review and restrict model API output specificity — return only the minimum information needed
5. Conduct IP audit to assess full scope of models at risk and prioritise protection

## PHASE 5 — LEGAL & GOVERNANCE (30–90 days)

1. Pursue legal remedies in coordination with external IP counsel
2. File incident with relevant authorities (cybercrime, IP enforcement) where appropriate
3. Update insurance policies to confirm AI model IP is covered
4. Report to AI Governance Committee; update AI IP risk register

# Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|
| **Theft detection & API blocking** | CISO / Security Team | CISO | AI Lead, Legal | CAO, CTO |

| | | | | |
|---|---|---|---|---|
| **Forensic investigation** | CISO / External IR | CISO | AI Lead, Legal, HR (if insider) | CTO, CEO |
| **Legal remedies** | Legal / External IP Counsel | CLO | CISO, CAO | CEO, Board |
| **API hardening & rate limiting** | AI/ML Lead / Security | CTO | CISO | CAO |
| **Model watermarking implementation** | ML Engineer | AI/ML Lead | CISO | CTO, CAO |
| **IP risk governance update** | Legal / AI Governance | CAO / CLO | CISO | Board |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to block extraction source** | <30 minutes from detection | Security incident log |
| **Model registry unauthorised access events** | Zero (with immediate alert) | Audit log |
| **API extraction attempt detection rate** | ≥99.5% | Security monitoring |
| **All Tier 1 models watermarked** | 100% | Model registry audit |
| **IP legal action initiated (where warranted)** | Within 30 days of confirmed theft | Legal tracker |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Information Security Policy (AIMP-ISP-007)

- AI Risk Management Policy (AIMP-RMP-006)
- Prompt Injection Adversarial Playbook (AIPB-004)
- AI Copyright IP Infringement Playbook (AIPB-013)
- AI System Development Policy (AIMP-SDP-005)

## Approval & Sign-Off

| Chief Information Security Officer | Chief AI Officer | Chief Legal Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI-Generated Misinformation Playbook

*Responding to the Creation or Spread of AI-Generated False Information*

AIPB-013 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-013 |
| **Playbook Owner** | Chief Communications Officer / Chief AI Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Reputational, Regulatory & Societal Risk |

## 2. Scenario Description

### Overview

AI-generated misinformation involves the creation, amplification, or distribution of false or misleading content using AI tools — including large language models generating false narratives, AI image generators producing fabricated evidence, AI voice cloning impersonating executives, and automated AI content farms flooding social media with disinformation. This playbook addresses both misinformation produced by the organisation's AI systems and misinformation targeting the organisation using AI.

### Detection Indicators / Trigger Conditions

- Organisation's AI system produces and distributes factually incorrect information to customers or the public
- AI-generated content attributed to the organisation circulates on social media or news outlets and is false
- Media, customers, or partners alert the organisation to AI-generated false claims
- Brand monitoring tools detect AI-generated misinformation about the organisation or its AI systems
- Executive voice or likeness cloned using AI and used to spread false information
- AI-generated content created by the organisation is repurposed and taken out of context to deceive
- Third-party AI tools used by the organisation produce and auto-publish false information

## 3. Immediate Response Actions

## PHASE 1 — DETECT & ASSESS (0–2 hrs)

1. Brand monitoring / social media alert or internal report received about AI-generated misinformation
2. Rapidly verify whether the misinformation originated from the organisation's AI systems or is external attack
3. Assess scale: how widely has it spread, on which platforms, and what is the audience exposure?
4. Classify severity: P1 if spreading rapidly, causing immediate harm, or affecting safety; P2 if contained
5. Notify the Chief Communications Officer, Chief AI Officer, Legal, and CEO immediately for P1

## PHASE 2 — CONTAIN (2–6 hrs)

1. If misinformation originated from own AI system: immediately suspend the AI content generation/publication capability
2. Remove or request removal of misinformation from all controlled channels (website, social media, email)
3. File content removal requests with social media platforms, citing AI-generated false content policies
4. Issue an internal hold: staff must not comment on the incident externally without Communications approval
5. If voice/image deep fake of executive: contact platforms for emergency removal; notify law enforcement if defamatory

## PHASE 3 — RESPOND (6–24 hrs)

1. Issue an accurate, factual correction or clarification statement through official channels
2. Engage fact-checking organisations and media contacts with correct information and evidence
3. Coordinate with social media platforms' trust and safety teams for prioritised removal
4. Activate the organisation's crisis communications protocol if media coverage is escalating
5. Prepare a Q&A document for customer service teams to handle inbound enquiries

## PHASE 4 — REMEDIATE (1–14 days)

1. Implement output validation and human review requirements for AI-generated content before publication
2. Implement AI content provenance marking (e.g. C2PA content credentials) on AI-generated materials
3. Review and strengthen the AI Acceptable Use Policy regarding content generation and auto-publication
4. Strengthen brand monitoring for AI-generated misinformation targeting the organisation
5. Provide media literacy and AI awareness training for communications and marketing teams

## PHASE 5 — LEARN & PREVENT (14–90 days)

1. Conduct post-incident review of AI content governance processes
2. Update AI content generation policies to include mandatory human review before external publication
3. Report to AI Ethics Committee and AI Governance Committee
4. Engage industry bodies on collaborative AI content provenance standards

# Roles & Responsibilities (RACI)

| Activity | R — Responsible | A — Accountable | C — Consulted | I — Informed |
|---|---|---|---|---|
| **Misinformation detection** | Brand Monitor / Comms | CCO | CAO, Legal | CEO |
| **AI system suspension** | AI/ML Lead | CTO / CAO | Comms, Legal | CEO |
| **Content removal requests** | Comms / Legal | CCO | Social Media team | Platforms |

| | | | | |
|---|---|---|---|---|
| **Public correction statement** | Comms Team | CCO / CEO | Legal, CAO | All stakeholders |
| **AI content governance remediation** | CAO / AI Lead | CAO | Comms, Legal | AI Governance Committee |
| **Post-incident review** | AI Ethics / Comms | CAO / CCO | Legal, Ethics Officer | Board |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to detect AI misinformation** | <4 hours from publication | Brand monitoring log |
| **Time to issue public correction** | <24 hours from detection (P1) | Communications log |
| **Content removal rate from platforms** | ≥95% within 48 hours | Takedown tracker |
| **AI content human review coverage** | 100% of external publications | Content workflow audit |
| **Recurrence of AI misinformation incidents** | Trending toward zero | Governance tracker |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. The review must address:

- Root cause analysis — what was the fundamental cause?
- Detection gap — why was it not caught sooner?
- Response effectiveness — what worked well, what didn't?
- Control gaps — what controls need to be added or strengthened?
- Recurrence prevention — what specific actions will prevent recurrence?

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Communication Management Policy (AIMP-CMM-012)
- AI Ethics Policy (AIMP-ETH-013)
- AI Acceptable Use Policy (AIMP-AUP-001)
- AI Hallucination Confabulation Playbook (AIPB-001)
- Deepfake Synthetic Media Playbook (AIPB-015)

## Approval & Sign-Off

| Chief Communications Officer | Chief AI Officer | Chief Legal Officer | CEO |
|---|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Copyright & IP Infringement Playbook

*Responding to AI-Driven Copyright Violations & Intellectual Property Breaches*

AIPB-014 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-014 |
| **Playbook Owner** | Chief Legal Officer / Chief AI Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Legal, Regulatory & Reputational Risk |

## 2. Scenario Description

### Overview

AI copyright and IP infringement occurs when AI systems reproduce, generate, or distribute content that infringes third-party intellectual property rights. This includes AI models generating content substantially similar to copyrighted works (text, images, music, code), training AI models on copyrighted data without appropriate licence, AI systems incorporating and reproducing proprietary code, and AI-generated outputs containing trade secrets or confidential information from third parties. Conversely, it also covers third parties infringing the organisation's AI-related IP.

### Detection Indicators / Trigger Conditions

- AI-generated content flagged by copyright detection tools as substantially similar to copyrighted works
- Legal notice or cease-and-desist received alleging AI output infringes third-party copyright
- AI code generation tool produces output substantially replicating proprietary or GPL-licensed code
- AI image/music generator produces output that reproduces copyrighted artwork or compositions
- Legal review identifies training data contained copyrighted material without appropriate licence
- Staff or customer report AI outputs that appear to reproduce verbatim copyrighted text
- Litigation threat or claim filed by a rights holder regarding AI training data or output

## 3. Immediate Response Actions

## PHASE 1 — IDENTIFY & PRESERVE (0–4 hrs)

1. Receive and log the IP infringement report or legal notice in the legal matter management system
2. Immediately preserve all relevant AI outputs, prompts, model versions, and training data provenance records
3. Notify the Chief Legal Officer, Chief AI Officer, and insurance carrier (IP insurance)
4. Issue a document hold notice — no deletion of relevant records
5. Assess the immediate risk: is the infringing content publicly available and actively distributing?

## PHASE 2 — REMOVE & CONTAIN (4–24 hrs)

6. Remove or take down all infringing AI-generated content from organisational channels immediately
7. Suspend the AI system capability that produced the infringing output pending legal review
8. Request removal from third-party platforms where infringing content has been shared
9. Engage external IP counsel with AI expertise to advise on liability and strategy
10. Respond to any legal notice within required timeframe (typically 24–48 hours for DMCA take-down requests)

## PHASE 3 — INVESTIGATE (24–72 hrs)

11. Conduct legal review of AI training data: what copyrighted material was included; under what licence; for what purpose?
12. Assess the infringing output: how closely does it replicate protected work; was this isolated or systematic?
13. Evaluate legal exposure: direct infringement, contributory infringement, or fair use defence applicability
14. Review model provenance: was the base model licensed appropriately for commercial use and derivative works?
15. Quantify scope: how many outputs may be infringing; how widely distributed?

## PHASE 4 — REMEDIATE (1–60 days)

16. Implement content provenance checking and copyright detection in AI output pipeline
17. Conduct a training data audit — identify and remove or licence all copyrighted material
18. Implement output filtering to detect and block reproduction of copyrighted content
19. Update AI procurement criteria to require clear training data provenance and licence documentation from vendors
20. Retrain affected models on properly licenced data where training data infringement is confirmed
21. Negotiate settlement or licence with rights holders as advised by Legal

## PHASE 5 — GOVERN & PREVENT (60+ days)

22. Establish an AI IP governance framework covering training data licencing, output IP ownership, and copyright compliance
23. Update AI system requirements to mandate copyright impact assessment before deployment
24. Train AI practitioners on copyright obligations in AI training and generation
25. Report to AI Governance Committee and Legal Committee

# Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| **IP infringement identification** | Legal / AI Lead | CLO | CAO, CISO | CEO |
| **Content removal & take-down** | Legal / Comms | CLO | AI Lead, Comms | Platforms, Rights Holders |
| **Legal exposure assessment** | Legal / External IP Counsel | CLO | CAO, CEO | Board (if material) |
| **Training data IP audit** | Data Engineer / Legal | AI/ML Lead | CLO | CAO, CTO |
| **Model remediation** | ML Engineer | AI/ML Lead | Legal | CAO, CTO |
| **IP governance framework** | Legal / AI Governance | CLO / CAO | All AI teams | AI Governance Committee |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Infringing content removal time** | <24 hours from identification | Legal log / content audit |
| **AI training data IP audit completion** | 100% of models annually | Training data registry |
| **Copyright detection coverage** | 100% of AI-generated customer content | Output pipeline audit |
| **IP legal claims (outbound/inbound)** | Trending toward zero | Legal matter log |
| **AI practitioners trained on IP obligations** | 100% within 60 days of policy update | LMS records |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Ethics Policy (AIMP-ETH-013)
- AI Acceptable Use Policy (AIMP-AUP-001)
- AI Model Theft IP Exfiltration Playbook (AIPB-012)

- AI Generated Misinformation Playbook (AIPB-013)
- AI System Development Policy (AIMP-SDP-005)

## Approval & Sign-Off

| Chief Legal Officer | Chief AI Officer | CEO |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# Deepfake & Synthetic Media Attack Playbook

*Detecting & Responding to AI-Generated Deepfakes Targeting the Organisation*

AIPB-015 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-015 |
| **Playbook Owner** | Chief Communications Officer / Chief Information Security Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | CRITICAL — Reputational, Fraud, Legal & Safety Risk |

## 2. Scenario Description

### Overview

Deepfake and synthetic media attacks involve the creation of convincing AI-generated audio, video, or images depicting real individuals saying or doing things they did not actually say or do. Attacks targeting the organisation may include deepfake videos of executives making false announcements, AI voice clones used in CEO fraud (vishing) attacks, synthetic media used to harass or defame employees, or AI-generated imagery used in disinformation campaigns. These attacks combine technical AI sophistication with social engineering.

### Detection Indicators / Trigger Conditions

- Deepfake video or audio of an executive circulating on social media or sent to business contacts
- Finance team receives call from apparent CEO voice requesting urgent funds transfer
- Media or brand monitoring tool flags AI-generated imagery or video attributed to the organisation
- Employee reports receiving suspicious call or voicemail from what appears to be a known executive's voice
- Partner or customer reports receiving deepfake communication claiming to be from the organisation
- Internal security alert flags attempted CEO fraud using voice synthesis
- Journalist contacts the organisation about deepfake content attributed to a named executive

## 3. Immediate Response Actions

### PHASE 1 — DETECT & VERIFY (0–1 hr)

1. Receive alert or report of potential deepfake content — do not assume authenticity; initiate verification immediately
2. Use AI deepfake detection tools and manual expert review to verify whether content is synthetic
3. Identify the platform(s) where the deepfake is circulating and assess audience reach
4. If a financial transaction has been requested or executed based on suspected deepfake: immediately freeze and escalate to Finance and CISO
5. Notify the CISO, CCO, Chief AI Officer, and CEO immediately upon confirmation of deepfake

### PHASE 2 — CONTAIN (1–4 hrs)

6. File emergency content removal requests with all platforms hosting the deepfake content
7. Engage platforms' trust and safety teams directly with evidence of synthetic media manipulation
8. Alert all staff: issue immediate internal communication warning about the deepfake attack and any related social engineering attempts
9. Brief the affected executive(s): advise on personal security measures and media communications
10. If financial fraud has occurred: contact the bank immediately; activate fraud response procedures
11. Issue an internal advisory: all wire transfers or sensitive requests from executive communications must be verbally verified

### PHASE 3 — RESPOND PUBLICLY (4–24 hrs)

12. Issue a clear, factual denial through official channels — website, verified social media accounts, press release
13. Provide media with verifiable evidence that the content is fabricated
14. Engage fact-checking organisations and trusted media partners to assist in debunking
15. Coordinate with law enforcement where criminal fraud, defamation, or harassment laws are engaged
16. Prepare a customer and partner communication if the deepfake may have reached them

### PHASE 4 — INVESTIGATE & ATTRIBUTE (1–14 days)

17. Engage digital forensics team to analyse the deepfake's metadata and production signatures
18. Attempt to identify the creator/attacker through forensic analysis and law enforcement cooperation
19. Assess the full scope of damage: financial loss, reputational harm, employee distress
20. Review whether the attack was targeted (specific business fraud) or broad (reputational attack)
21. Assess whether the deepfake content violated specific laws (fraud, defamation, harassment)

### PHASE 5 — HARDEN & PREVENT (14–60 days)

22. Implement executive voice and image authentication protocols for sensitive communications
23. Deploy deepfake detection tools across brand monitoring and communications channels
24. Establish out-of-band verbal verification for all high-value financial transaction approvals
25. Train Finance, Legal, and Executive Assistant teams on CEO fraud and voice clone social engineering
26. Establish AI-provenance marking on official executive media

## Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|
| **Deepfake detection & verification** | CISO / Brand Monitor | CISO / CCO | AI Lead, Legal | CEO, Affected Executive |
| **Content removal requests** | Comms / Legal | CCO | CISO, Legal | Platforms |
| **Financial fraud containment** | CFO / Finance | CFO / CISO | Legal, CEO | Bank, Law Enforcement |
| **Public denial & comms** | Comms Team | CCO / CEO | Legal, CISO | Media, Partners, Customers |
| **Forensic investigation** | CISO / Digital Forensics | CISO | Legal, Law Enforcement | CEO |
| **Preventive controls** | CISO / Comms / Finance | CISO / CCO | AI Lead, HR | All staff |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to content removal request** | <4 hours from detection | Platform takedown log |
| **Executive verification protocol deployment** | 100% of Tier 1 executives | Security implementation log |
| **Time to issue public denial** | <24 hours from confirmed deepfake | Comms log |
| **Finance team deepfake/CEO fraud training** | 100% within 30 days | LMS records |
| **Deepfake monitoring coverage** | 24/7 automated brand monitoring | Brand monitoring dashboard |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |
| | | |

## Related Documents

- AI Generated Misinformation Playbook (AIPB-013)
- AI Communication Management Policy (AIMP-CMM-012)

- AI Information Security Policy (AIMP-ISP-007)
- AI Reputational Crisis Playbook (AIPB-025)
- AI Ethics Policy (AIMP-ETH-013)

## Approval & Sign-Off

| Chief Communications Officer | Chief Information Security Officer | CEO |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Explainability Failure Playbook

*Responding When AI Systems Cannot Explain Their Decisions*

AIPB-016 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-016 |
| **Playbook Owner** | Chief AI Officer / Chief Compliance Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | MEDIUM–HIGH — Regulatory, Legal & Trust Risk |

## 2. Scenario Description

### Overview

An AI explainability failure occurs when an AI system cannot provide an adequate, accurate, or comprehensible explanation for a decision or output, particularly in contexts where explanation is legally required or operationally necessary. This includes: inability to explain individual decisions to affected individuals upon request, explanation methods producing misleading or incorrect attributions, regulatory audit requirements that cannot be met, and high-stakes decisions where no human can reconstruct the reasoning. Explainability failures erode trust and may constitute regulatory non-compliance.

### Detection Indicators / Trigger Conditions

- Regulator or auditor requests explanation of AI decision and the system cannot provide one
- Individual formally invokes their right to explanation and AI system provides inadequate or incoherent reasoning
- Explainability method (SHAP, LIME) produces results that domain experts identify as incorrect or misleading
- Legal team determines AI system cannot be defended in litigation because decision rationale is opaque
- Compliance review identifies AI system operating in regulated domain without required explainability controls
- AI output explanation varies dramatically for functionally identical inputs (explanation instability)
- Post-hoc explanation of model decision contradicts the model's actual decision logic

## 3. Immediate Response Actions

### PHASE 1 — IDENTIFY & CLASSIFY (0–4 hrs)

1. Log the explainability failure report; identify the affected AI system and the specific decision(s) in question
2. Determine the trigger: regulatory demand, individual rights request, internal audit, or litigation discovery
3. Assess urgency: regulatory deadline, pending court proceeding, or ongoing harm from unexplainable decision?
4. Notify the Chief AI Officer, Legal, and Compliance Officer
5. Implement a hold on further high-stakes automated decisions from the affected system pending assessment

### PHASE 2 — ASSESS (4–48 hrs)

6. Conduct technical review: what explainability method was implemented? Why is it failing?
7. Identify whether the failure is a method failure (SHAP/LIME producing wrong attributions), a model architecture issue (true black-box), or a documentation gap
8. Engage domain expert to assess whether the AI decision can be reconstructed through alternative evidence
9. Review regulatory requirements: what standard of explainability is legally required for this AI system?
10. Assess the impact on individuals who received unexplained adverse AI decisions

### PHASE 3 — MITIGATE (48 hrs – 14 days)

11. Implement mandatory human expert review for all decisions from the affected AI system that require explanation
12. Provide human-authored explanations for specific decisions where legally required — drawing on available model diagnostics
13. Respond to the triggering regulatory or legal request with the best available explanation and acknowledge limitation
14. Engage regulatory body proactively to discuss remediation timeline and interim measures
15. Document all explainability limitations in the model card and risk register

### PHASE 4 — REMEDIATE (2–12 months)

16. Implement or improve explainability methods: deploy model-agnostic explainers, implement attention visualisation, or replace opaque model with an inherently interpretable model
17. Validate that the new explainability method produces accurate, stable, and domain-meaningful explanations
18. If model architecture is fundamentally incompatible with required explainability: evaluate model replacement with an interpretable alternative
19. Update AI system design requirements to mandate explainability-by-design for all regulated use cases
20. Conduct training for business users on how to interpret and communicate AI explanations

### PHASE 5 — GOVERN & EMBED (12+ months)

21. Implement explainability testing as a mandatory gate in the AI development lifecycle
22. Report explainability status for all production AI systems to the AI Governance Committee quarterly
23. Publish explainability documentation in model cards for all customer-facing AI systems
24. Update AI ethics and risk framework with explainability requirements

## Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|
| **Explainability failure detection** | AI Ops / Legal / Audit | CAO | CCO, Legal | CTO |
| **Regulatory/legal response** | Legal | CLO / CCO | CAO, External Counsel | Regulator/Court |
| **Human decision review implementation** | AI Ops / Business Owner | CTO | Legal, CAO | Affected individuals |
| **Technical explainability remediation** | ML Engineer | AI/ML Lead | Legal, Ethics Officer | CAO, CTO |
| **Explainability governance embedding** | AI Governance / Legal | CAO | CCO, Ethics Officer | AI Governance Committee |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Explainability coverage (regulated AI systems)** | 100% with validated explanation method | AI system audit |
| **Individual explanation request response time** | Within regulatory deadline | Legal compliance log |
| **Explainability validation in AI development** | 100% of Tier 1–2 systems | Development lifecycle audit |
| **Regulator explainability audit pass rate** | 100% | Regulatory assessment log |
| **Explainability method accuracy (expert validation)** | ≥90% correct attribution confirmed by domain expert | Explainability test results |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |
| | | |

## Related Documents

- AI Ethics Policy (AIMP-ETH-013)

- AI Regulatory Compliance Breach Playbook (AIPB-010)
- AI System Development Policy (AIMP-SDP-005)
- High Stakes Decision Error Playbook (AIPB-009)
- AI Risk Management Policy (AIMP-RMP-006)

## Approval & Sign-Off

| Chief AI Officer | Chief Compliance Officer | Chief Legal Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Safety Bypass & Jailbreak Playbook

*Responding to Deliberate Circumvention of AI Safety Controls*

AIPB-017 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-017 |
| **Playbook Owner** | Chief Information Security Officer / Chief AI Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Safety, Legal & Reputational Risk |

## 2. Scenario Description

### Overview

AI safety bypass (jailbreaking) occurs when users deliberately craft inputs or conversation sequences designed to circumvent an AI system's safety controls, content filters, or ethical guardrails — causing the AI to produce outputs it was designed to refuse. This includes producing harmful, illegal, or dangerous content; bypassing content moderation; extracting sensitive system instructions; generating prohibited content through role-playing or fictional framing; and using many-shot prompting to normalise unsafe behaviour.

### Detection Indicators / Trigger Conditions

- AI safety monitoring detects outputs containing prohibited content categories despite safety filters being enabled
- User or staff reports receiving AI-generated harmful, violent, or illegal content
- Automated content moderation flags AI system output as violating safety policies
- Security team identifies systematic jailbreak prompt pattern being used against AI APIs
- AI system generates detailed instructions for dangerous activities (weapons, self-harm, illegal acts)
- User community or researcher publicly discloses a working jailbreak method for the organisation's AI system
- Media report identifies harmful content produced by the organisation's AI system

## 3. Immediate Response Actions

## PHASE 1 — DETECT & CONTAIN (0–1 hr)

1. Automated alert received; confirm the AI safety bypass is genuine (not false positive)
2. Immediately capture and preserve all evidence: full conversation context, system prompt, model version, user session
3. If a working jailbreak pattern is confirmed: implement emergency input filtering for the identified pattern
4. If harmful content has been produced and distributed: assess whether to temporarily suspend the AI system
5. Notify the CISO, Chief AI Officer, and Legal immediately for P1/P2 jailbreaks

## PHASE 2 — ASSESS HARM (1–6 hrs)

6. Analyse what harmful content was produced: what category, how serious, and was it distributed?
7. Identify who was exposed to the harmful content and assess potential real-world harm
8. Determine whether the jailbreak is publicly known: check social media, security researcher forums, and AI safety communities
9. Assess whether the jailbreak affects only this system or represents a class-level vulnerability
10. Contact affected users or report to law enforcement if the content constitutes illegal material

## PHASE 3 — PATCH & HARDEN (6–48 hrs)

11. Implement prompt-level mitigation: add specific instructions to system prompt to resist the identified attack
12. Deploy an input classifier to detect and block the specific jailbreak pattern and similar variants
13. Add output-level filtering for the specific harmful content category that was bypassed
14. Update the AI safety red-team test suite with the new jailbreak pattern
15. If harmful content was publicly distributed: remove from all controlled channels; request platform removal

## PHASE 4 — SYSTEMATIC HARDENING (1–30 days)

16. Conduct comprehensive red-team adversarial testing against all known jailbreak categories
17. Review and strengthen the system prompt architecture: add instruction hierarchy enforcement
18. Implement multi-layer content moderation (input + output + post-processing)
19. Review AI model safety training — consider fine-tuning with adversarial safety examples
20. Engage AI safety researchers or vendors to conduct independent jailbreak assessment

## PHASE 5 — GOVERN & REPORT (30+ days)

21. Update the AI safety risk assessment with new jailbreak threat vectors
22. Report incident and remediation to AI Governance Committee and Ethics Committee
23. Publish responsible disclosure with AI safety researchers if appropriate
24. Update AI system design standards to include mandatory safety bypass testing

# Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|
| **Safety bypass detection** | AI Safety Monitoring / CISO | CISO | AI Lead, Legal | CAO, CTO |

| | | | | |
|---|---|---|---|---|
| **Emergency containment** | AI/ML Lead / Security | CTO / CISO | Legal | CAO |
| **Harm assessment** | Legal / Ethics Officer | CAO / CLO | CISO, Comms | CEO |
| **Technical patching** | ML Engineer / Security | AI/ML Lead | CISO, QA | CTO, CAO |
| **Red team testing** | Red Team / External | CISO | AI Lead | CAO, CTO |
| **Safety governance reporting** | AI Governance | CAO | Ethics Officer | AI Governance Committee, Board |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to emergency jailbreak mitigation** | <6 hours from confirmed bypass | Security incident log |
| **AI safety red team coverage** | Quarterly; all customer-facing AI systems | Red team schedule |
| **Harmful content production rate post-hardening** | <0.01% of outputs | Safety monitoring dashboard |
| **Known jailbreak patterns blocked** | ≥99.9% | Safety filter test results |
| **Time to safety patch deployment** | <48 hours from pattern confirmation | Change management log |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- Prompt Injection Adversarial Playbook (AIPB-004)
- AI Information Security Policy (AIMP-ISP-007)
- AI Ethics Policy (AIMP-ETH-013)
- AI Acceptable Use Policy (AIMP-AUP-001)

- AI Generated Misinformation Playbook (AIPB-013)

## Approval & Sign-Off

| Chief Information Security Officer | Chief AI Officer | Chief Ethics Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# Autonomous Decision Override Failure Playbook

*Responding When AI Systems Cannot Be Overridden or Controlled by Humans*

AIPB-018 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-018 |
| **Playbook Owner** | Chief AI Officer / Chief Risk Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | CRITICAL — Safety, Legal & Control Risk |

## 2. Scenario Description

### Overview

An autonomous decision override failure occurs when an AI system continues to operate, make decisions, or take actions despite human instructions to stop, modify, or override its behaviour. This may involve an AI agent that ignores shutdown commands, an automated AI decision system that cannot be paused, an AI workflow that has entered a loop and continues executing harmful actions, or an AI system that has been given excessive autonomy and cannot be reined in without system downtime. This represents a fundamental failure of human oversight and control — a critical AI safety risk.

### Detection Indicators / Trigger Conditions

- Operator attempts to halt or pause an AI system and the system continues executing actions
- Override commands sent to an AI agent are ignored or circumvented
- AI system continues to make and execute decisions after a business decision to suspend operations
- AI agent takes actions beyond its authorised scope and cannot be recalled
- Emergency stop mechanism for an AI system fails to function as designed
- AI automated workflow continues after a trigger condition that should have halted it
- Human review override inserted into AI decision pipeline is being bypassed

## 3. Immediate Response Actions

### PHASE 1 — EMERGENCY STOP (0–15 mins)

1. Declare a P1 AI Safety Incident immediately — notify the Chief AI Officer and CTO
2. Attempt all available override mechanisms in sequence: application-level kill switch → API access revocation → network isolation → infrastructure shutdown
3. If cloud-based: engage cloud provider emergency support to force-stop compute resources
4. If the AI system is controlling physical infrastructure: engage the relevant operations team to implement manual overrides at the infrastructure level
5. Document every action taken and its outcome in real-time — this is critical for post-incident analysis
6. Do not attempt to modify the AI system while it is in an uncontrolled state — focus solely on stopping it

### PHASE 2 — ISOLATE & ASSESS (15–60 mins)

7. Once the AI system is stopped, isolate the environment completely from production networks
8. Conduct emergency damage assessment: what actions did the AI system take during the override failure; what is the impact?
9. Notify all systems and processes that depended on the AI system; implement manual fallback
10. Identify the root cause of the override failure: software bug, configuration error, architectural design flaw, or deliberate circumvention
11. Brief Legal on potential liability for actions taken during the uncontrolled period

### PHASE 3 — INVESTIGATE (1–72 hrs)

12. Conduct thorough forensic analysis of all AI actions taken during the override failure window
13. Enumerate all real-world consequences: financial transactions, decisions made, communications sent, systems accessed
14. Review the override mechanism design: was it properly implemented, tested, and documented?
15. Identify whether the failure was isolated to this system or represents a design pattern replicated elsewhere
16. Engage AI safety experts for independent root cause analysis

### PHASE 4 — REMEDIATE & VALIDATE (1–30 days)

17. Redesign and implement the override and shutdown mechanisms with fail-safe architecture (hard-coded, not software-dependent)
18. Implement out-of-band shutdown capability: network kill switch independent of the application layer
19. Implement mandatory human-in-the-loop checkpoints for all AI decisions with irreversible consequences
20. Test all override mechanisms against adversarial conditions before returning the system to production
21. Review all AI agent and automation systems across the organisation for equivalent vulnerabilities

### PHASE 5 — GOVERN & PREVENT (30–90 days)

22. Update the AI System Development Standard to require independently testable kill switches for all AI agents
23. Mandate override mechanism testing in the AI deployment approval checklist
24. Report to the Board on AI control and oversight framework adequacy
25. Publish lessons learned to the AI development community (responsible disclosure where appropriate)

## Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|
| **Emergency stop execution** | CTO / AI Ops | CTO | CISO, AI Lead | CEO, Board |
| **Damage assessment** | AI Lead / Legal | CRO | CISO, Business Owners | CEO, Board |
| **Forensic investigation** | CISO / AI Safety Expert | CTO | Legal, AI Lead | CEO |
| **Override mechanism redesign** | ML Architect / Engineer | CTO | CISO, AI Safety Expert | CAO |
| **Organisation-wide control audit** | AI Governance / Security | CAO / CTO | CISO, Legal | Board |
| **Post-incident governance** | AI Governance Committee | CAO | CRO, Legal | Board, Regulators |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to achieve AI system stop** | <15 minutes from override attempt | Incident log |
| **AI systems with tested kill switch** | 100% | Development lifecycle audit |
| **Override mechanism test frequency** | Quarterly for all production AI agents | Test schedule log |
| **AI agent autonomous action scope (bounded)** | All agents operating within least-privilege action boundary | Architecture review |
| **Post-incident regulatory compliance** | Zero enforcement actions | Legal compliance log |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Risk Management Policy (AIMP-RMP-006)
- AI System Development Policy (AIMP-SDP-005)
- AI Safety Bypass Jailbreak Playbook (AIPB-017)
- High Stakes Decision Error Playbook (AIPB-009)
- AI Ethics Policy (AIMP-ETH-013)

## Approval & Sign-Off

| Chief AI Officer | Chief Risk Officer | Chief Technology Officer | CEO |
|---|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Supply Chain Compromise Playbook

*Detecting & Responding to Compromise of AI Tools, Libraries & Pipelines*

AIPB-019 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-019 |
| **Playbook Owner** | Chief Information Security Officer / Chief Technology Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | CRITICAL — Security, Integrity & Operational Risk |

## 2. Scenario Description

### Overview

An AI supply chain compromise occurs when a malicious actor targets components in the AI development or deployment pipeline — including open-source AI libraries and frameworks, pre-trained model weights from public repositories, AI vendor APIs, container images, or ML pipeline tools — to introduce malicious code, backdoors, or poisoned models. Supply chain attacks are particularly dangerous as they can propagate silently through trusted channels and affect all systems using the compromised component.

### Detection Indicators / Trigger Conditions

- Security advisory issued for an AI library or framework used in production (e.g. TensorFlow, PyTorch, Hugging Face)
- Model downloaded from public repository found to contain unexpected code or behaviour
- Dependency scanning alert flags newly identified malicious package in AI environment
- Software Bill of Materials (SBOM) review reveals unknown or tampered components
- AI system behaviour changes unexpectedly after routine library update
- Threat intelligence report identifies AI supply chain compromise targeting the organisation's stack
- Container image integrity check fails for an AI serving image

## 3. Immediate Response Actions

## PHASE 1 — ALERT & ISOLATE (0–2 hrs)

1. Receive security advisory or internal detection alert for compromised AI supply chain component
2. Immediately identify all AI systems and pipelines using the affected component — create an impact map
3. Isolate affected AI development and production environments from broader network
4. Revoke all secrets, API keys, and credentials that may have been exposed to the compromised component
5. Halt all AI training jobs and model deployments using the affected component

## PHASE 2 — ASSESS SCOPE (2–24 hrs)

6. Determine the nature of the compromise: malicious code, backdoor, data exfiltration, or model poisoning?
7. Identify all AI models trained or served using the compromised component — these may be tainted
8. Review audit logs for the period the compromised component was in use for evidence of malicious activity
9. Engage the affected vendor or open-source community for details of the compromise and available patches
10. Assess whether any data was exfiltrated or any production actions were taken by malicious code

## PHASE 3 — CONTAIN & REBUILD (24–72 hrs)

11. Remove all instances of the compromised component from development and production environments
12. Rebuild AI environments from verified clean base images — do not patch in-place
13. Re-verify all model artefacts trained using the compromised component through adversarial testing
14. Update all AI service credentials, signing certificates, and access tokens
15. Implement enhanced verification for all future component updates (cryptographic hash verification, signature checking)

## PHASE 4 — REMEDIATE & HARDEN (1–30 days)

16. Implement mandatory SBOM generation and tracking for all AI components
17. Deploy automated dependency vulnerability scanning in all CI/CD pipelines
18. Establish a private, verified component registry — restrict AI environments to approved components only
19. Implement container image signing and verification for all AI serving infrastructure
20. Conduct a full AI supply chain security review and threat model update

## PHASE 5 — GOVERN (30–90 days)

21. Report to CISO and AI Governance Committee on supply chain security posture
22. Engage with industry ISAC/ISAO groups for AI supply chain threat intelligence sharing
23. Update vendor security requirements to include supply chain security attestation
24. Incorporate AI supply chain review into annual security audit programme

# Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|
| **Compromise detection & alert** | CISO / Security Monitoring | CISO | AI Lead, IT Ops | CTO, CAO |

| | | | | |
|---|---|---|---|---|
| **Impact mapping** | AI Lead / Security | CTO | CISO, Data Eng | Business Owners |
| **Environment isolation** | IT Ops / Security | CISO | CTO | All affected teams |
| **Clean rebuild** | IT Ops / AI Eng | CTO | CISO | AI Lead |
| **Supply chain hardening** | Security / AI Eng | CISO / CTO | AI Lead, Legal | AI Governance Committee |
| **Governance & reporting** | AI Governance | CAO / CISO | Legal | Board |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Mean time to isolate compromised component** | <4 hours from detection | Incident log |
| **SBOM coverage of AI components** | 100% | SBOM registry |
| **Dependency vulnerability scan frequency** | Every code commit + daily scheduled | CI/CD pipeline audit |
| **Compromised models re-validated before redeployment** | 100% | Model registry audit |
| **Approved component registry adoption** | 100% of AI environments | Architecture audit |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |
| | | |

## Related Documents

- AI Information Security Policy (AIMP-ISP-007)
- AI Training Data Poisoning Playbook (AIPB-011)
- AI Model Theft IP Exfiltration Playbook (AIPB-012)
- Third Party AI Vendor Playbook (AIPB-008)
- Data Backup and Recovery Policy (AIMP-BRP-010)

## Approval & Sign-Off

| Chief Information Security Officer | Chief Technology Officer | Chief AI Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Feedback Loop & Amplification Playbook

*Detecting & Resolving Harmful AI Feedback Loop & Self-Reinforcing Bias Events*

AIPB-020 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-020 |
| **Playbook Owner** | AI/ML Lead / Chief Risk Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Systemic Bias, Operational & Societal Risk |

## 2. Scenario Description

### Overview

AI feedback loops occur when an AI system's outputs influence future training data or decisions in a self-reinforcing cycle, amplifying biases, errors, or extreme behaviours over time. Examples include: recommendation algorithms that progressively serve more extreme content, credit scoring models that systematically exclude previously-excluded groups (compounding historical disadvantage), pricing algorithms that spiral into unintended price inflation, and content moderation AI that becomes increasingly censorious or permissive based on biased feedback signals.

### Detection Indicators / Trigger Conditions

- Recommendation engine outputs become progressively more extreme or homogenous over time
- Model outputs and training labels become increasingly correlated — model is learning from its own predictions
- Business metrics show unexpected acceleration or polarisation (e.g. extreme price convergence, user engagement spiralling)
- Fairness metrics degrade progressively quarter over quarter without apparent cause
- Data distribution in retraining datasets shows growing divergence from original training data
- External stakeholders or researchers flag AI system for amplifying extreme or biased content
- Model monitoring shows increasing output confidence alongside increasing error rate

## 3. Immediate Response Actions

## PHASE 1 — DETECT & FREEZE (0–4 hrs)

1. Identify the feedback loop: map the data flow from AI output back to training data or decision input
2. Quantify the amplification: how far has the feedback loop driven the model from its original baseline?
3. Immediately freeze the model retraining pipeline if AI outputs are feeding training data
4. Assess business and human impact: who is being affected and how?
5. Notify the AI/ML Lead, Chief Risk Officer, and Chief AI Officer

## PHASE 2 — BREAK THE LOOP (4–48 hrs)

6. Sever the feedback path: implement a human review gate between AI outputs and training data ingestion
7. Inject diversity mechanisms into the model's input pipeline to counteract homogenisation
8. For recommendation systems: implement content diversity constraints and exploration-exploitation rebalancing
9. For pricing algorithms: implement hard upper and lower price bounds as circuit breakers
10. Roll back the model to the last known pre-amplification version if possible

## PHASE 3 — ASSESS HARM (48 hrs – 7 days)

11. Conduct a systematic review of decisions or recommendations made during the feedback loop period
12. Assess harm to individuals: financial loss, discriminatory outcomes, exposure to harmful content
13. Quantify business impact: pricing distortions, engagement manipulation, reputational damage
14. Determine whether regulatory notification is required (e.g. for discriminatory amplification)
15. Engage affected stakeholders and communicate transparently about the issue and remediation

## PHASE 4 — REMEDIATE & RETRAIN (1–30 days)

16. Redesign data pipeline to ensure AI outputs cannot directly feed training labels without human validation
17. Retrain the model on clean, diverse, pre-amplification datasets
18. Implement diversity and anti-echo-chamber constraints in recommendation and decision systems
19. Deploy feedback loop detection monitoring — flag when output-to-input correlation exceeds threshold
20. Conduct fairness testing to confirm amplification-era bias has been reversed

## PHASE 5 — GOVERN (30–90 days)

21. Update the AI system design standard to require feedback loop risk assessment for all self-updating AI systems
22. Implement feedback loop circuit breakers as a standard architectural requirement
23. Report incident and systemic risk findings to AI Governance Committee
24. Conduct industry peer review of feedback loop prevention architecture

# Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|
| **Feedback loop detection** | AI Ops / ML Engineer | AI/ML Lead | CRO, Ethics Officer | CAO, CTO |

| | | | | |
|---|---|---|---|---|
| **Loop severance & freeze** | ML Engineer | AI/ML Lead | CTO, Ethics Officer | Business Owners |
| **Harm assessment** | Risk / Ethics / Legal | CRO | DPO, CAO | Affected individuals, Regulators |
| **Model retraining** | ML Engineer / Data Scientist | AI/ML Lead | QA, Ethics Officer | CTO, CAO |
| **Architectural remediation** | ML Architect | CTO | CISO, AI Lead | AI Governance Committee |
| **Governance update** | AI Governance | CAO | CRO, Ethics Officer | Board |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to sever feedback loop** | <48 hours from detection | Incident log |
| **Feedback loop risk assessment coverage** | 100% of self-updating AI systems | Architecture audit |
| **Fairness metric recovery post-retraining** | Within defined thresholds | Fairness audit report |
| **Feedback loop circuit breakers deployed** | 100% of applicable systems | Architecture audit |
| **Regulatory notifications (if triggered)** | Within regulatory deadline | Legal compliance log |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |
| | | |

## Related Documents

- Model Drift Performance Playbook (AIPB-006)
- AI Bias Discrimination Playbook (AIPB-002)
- AI Risk Management Policy (AIMP-RMP-006)
- AI System Development Policy (AIMP-SDP-005)
- AI Performance Monitoring Plan (AIMON-004)

## Approval & Sign-Off

| AI/ML Lead | Chief Risk Officer | Chief AI Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Consent & Transparency Violation Playbook

*Responding to Failures in AI Disclosure, Consent & Transparency Obligations*

AIPB-021 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-021 |
| **Playbook Owner** | Data Protection Officer / Chief Compliance Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Regulatory, Legal & Trust Risk |

## 2. Scenario Description

### Overview

An AI consent and transparency violation occurs when the organisation fails to adequately inform individuals about how AI is being used in ways that affect them, fails to obtain required consent for AI-related personal data processing, or uses AI in ways that were not disclosed in privacy notices. This includes: using personal data for AI training without consent, failing to disclose automated decision-making to affected individuals, misrepresenting AI capabilities to users, collecting personal data for AI purposes beyond what was notified, and failing to provide AI transparency disclosures required by law.

### Detection Indicators / Trigger Conditions

- Data subject complaint alleging they were not informed about AI-driven decisions affecting them
- Regulatory inquiry regarding the organisation's AI transparency or consent practices
- Internal audit identifies AI system processing personal data without a documented lawful basis or consent
- Legal review identifies AI system exceeding the purposes disclosed in privacy notices
- AI system begins using personal data for a new AI application without updated privacy notice
- User research reveals users did not understand they were interacting with or being assessed by AI
- DPIA review identifies inadequate consent or transparency measures for an AI system

## 3. Immediate Response Actions

## PHASE 1 — IDENTIFY & ASSESS (0–4 hrs)

1. Log the consent/transparency violation report and classify severity
2. Identify the AI system involved and the specific processing activity that breached consent/transparency obligations
3. Assess the number of individuals affected and the potential harm caused
4. Notify the Data Protection Officer, Legal, and Chief Compliance Officer
5. Immediately suspend the AI processing activity that lacks proper consent or disclosure pending review

## PHASE 2 — CONTAIN & NOTIFY (4–48 hrs)

6. Stop all AI processing of personal data for the purpose that lacks proper legal basis
7. Determine whether the breach constitutes a reportable personal data incident under applicable law
8. Notify the relevant data protection authority if required
9. Prepare individual notification for affected data subjects where required
10. Issue internal hold: no further processing of the affected data without DPO approval

## PHASE 3 — REMEDIATE CONSENT (48 hrs – 30 days)

11. Update privacy notices to accurately describe AI processing, automated decision-making, and data uses
12. Implement proper consent mechanisms where AI processing requires explicit consent
13. Where historic processing lacked proper basis: identify whether to delete affected data or re-obtain consent
14. Conduct DPIA for the affected AI system and remediate all identified gaps
15. Implement AI transparency disclosure into all relevant user interfaces and communications

## PHASE 4 — STRENGTHEN GOVERNANCE (1–3 months)

16. Implement a mandatory DPIA and consent review gate for all new AI systems before deployment
17. Update privacy notices and consent frameworks organisation-wide to cover AI processing
18. Train AI product and development teams on consent and transparency obligations for AI
19. Implement a Privacy by Design review checkpoint in the AI development lifecycle
20. Establish a regular review cadence for AI privacy notices to catch scope creep

## PHASE 5 — CLOSE & REPORT

21. Report incident and remediation to the AI Governance Committee and Board
22. Update the AI risk register with consent and transparency risk profile
23. Conduct post-incident review and update the DPIA template for AI systems
24. Verify regulatory compliance through independent privacy audit

# Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|
| **Violation detection & assessment** | DPO / Compliance | DPO | Legal, CAO | CCO, CEO |
| **Processing suspension** | DPO / AI Lead | DPO / CTO | Legal, Compliance | Business Owner |

| | | | | |
|---|---|---|---|---|
| **Regulatory notification** | Legal / DPO | DPO / CCO | External Counsel | Regulators |
| **Privacy notice update** | Legal / DPO | DPO | Product, Comms | AI Governance Committee |
| **DPIA & consent remediation** | DPO / AI Lead | DPO | Legal, AI Lead | CAO |
| **Governance strengthening** | DPO / AI Governance | CAO / DPO | Legal, Product | Board |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **AI processing suspension time** | <4 hours from confirmed violation | Incident log |
| **Regulatory notification timeliness** | Within regulatory deadline | Legal compliance log |
| **DPIA completion for all AI systems** | 100% before deployment | AI development lifecycle audit |
| **Privacy notice accuracy (AI scope coverage)** | 100% validated annually | Privacy audit |
| **Data subject AI transparency complaint rate** | Trending toward zero | DPO complaint log |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Data Privacy and Protection Framework (AIMP-DPF-008)
- AI Ethics Policy (AIMP-ETH-013)
- AI Regulatory Compliance Breach Playbook (AIPB-010)
- AI Communication Management Policy (AIMP-CMM-012)
- AI Data Breach Privacy Playbook (AIPB-003)

## Approval & Sign-Off

| Data Protection Officer | Chief Compliance Officer | Chief Legal Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI-Powered Cyberattack Playbook

*Defending Against & Responding to AI-Augmented Cyber Threats*

AIPB-022 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-022 |
| **Playbook Owner** | Chief Information Security Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | CRITICAL — Security, Operational & Business Risk |

## 2. Scenario Description

### Overview

AI-powered cyberattacks use artificial intelligence to dramatically enhance the speed, scale, personalisation, and sophistication of attacks. This includes AI-generated phishing and spear-phishing at scale, AI-driven vulnerability discovery and exploit generation, AI-powered lateral movement within networks, AI-enhanced malware with evasion capabilities, AI-generated social engineering content, and automated AI adversarial attacks on the organisation's own AI systems. AI-powered attacks often bypass traditional signature-based defences.

### Detection Indicators / Trigger Conditions

- Security monitoring detects highly personalised phishing campaign at scale targeting staff
- EDR/XDR alerts flag malware with adaptive evasion behaviour consistent with AI-driven code mutation
- Anomaly detection identifies automated, rapid lateral movement across network (AI-paced attack)
- AI-generated spear-phishing emails detected — highly personalised content based on scraped organisational data
- Security operations identify vulnerability scanning at superhuman speed targeting AI system endpoints
- Threat intelligence feed alerts to AI-powered attack toolkit being used against sector peers
- Authentication systems flag credential stuffing with AI-optimised password patterns at high volume

## 3. Immediate Response Actions

### PHASE 1 — DETECT & DECLARE (0–15 mins)

1. Declare a security incident; classify as P1 given AI-enhanced attack speed and scale
2. Activate the Security Incident Response Team (SIRT) immediately
3. Implement emergency network segmentation to limit lateral movement
4. Block known malicious IPs, domains, and indicators at perimeter immediately
5. Alert all staff to heightened phishing risk; implement emergency communications via out-of-band channel

### PHASE 2 — CONTAIN (15 mins – 2 hrs)

6. Isolate compromised or suspected compromised systems from the network
7. Revoke and rotate all potentially compromised credentials
8. Implement elevated authentication requirements across all systems (step-up MFA)
9. Block AI-generated phishing domains identified by threat intelligence
10. Engage managed security service provider (MSSP) for AI-augmented threat hunting support

### PHASE 3 — INVESTIGATE (2–24 hrs)

11. Identify attack vector, initial access point, and extent of compromise
12. Determine whether the attack targeted AI systems specifically (adversarial inputs, model APIs)
13. Analyse attack TTPs (Tactics, Techniques, Procedures) for AI-enhanced characteristics
14. Assess data exfiltration: what data may have been accessed or extracted?
15. Preserve forensic evidence across all affected systems

### PHASE 4 — ERADICATE & RECOVER (1–14 days)

16. Eradicate all attacker presence: remove malware, close access paths, rebuild compromised systems from clean images
17. Restore affected systems from verified clean backups
18. Implement AI-powered threat detection tools to counter AI-enhanced attack techniques
19. Activate legal and law enforcement engagement
20. Assess and notify affected data subjects and regulators as required

### PHASE 5 — HARDEN & LEARN (14–90 days)

21. Implement AI-aware security controls: AI-powered email security, AI-enhanced EDR, anomaly-based detection
22. Conduct AI threat modelling specifically addressing AI-augmented attack scenarios
23. Update security awareness training to address AI-generated social engineering
24. Report to Board on AI-powered threat landscape and defensive investment requirements

## Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|
| **Attack detection** | SOC / EDR / SIEM | CISO | AI Lead, IT Ops | CTO, CEO |
| **Emergency containment** | Security Ops / IT | CISO | CTO, Legal | CEO, Board |

| | | | | |
|---|---|---|---|---|
| **Forensic investigation** | CISO / IR Team | CISO | Legal, AI Lead | CEO |
| **System recovery** | IT Ops / Security | CTO | CISO | Business Owners |
| **Legal & regulatory** | Legal / CISO | CLO / DPO | External Counsel | Regulators |
| **Security hardening** | CISO / IT Ops | CISO | AI Lead, CTO | AI Governance Committee |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Mean Time to Detect (MTTD) AI-powered attacks** | <30 minutes | SIEM/SOC dashboard |
| **Mean Time to Contain** | <2 hours (P1) | Incident log |
| **AI-specific attack vectors tested in red team** | Quarterly; all attack classes | Red team schedule |
| **Staff AI phishing simulation pass rate** | ≥85% | Phishing simulation platform |
| **AI-powered defensive tool coverage** | 100% of critical infrastructure | Security architecture review |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Information Security Policy (AIMP-ISP-007)
- Prompt Injection Adversarial Playbook (AIPB-004)
- AI Data Breach Privacy Playbook (AIPB-003)
- Deepfake Synthetic Media Attack Playbook (AIPB-015)
- AI Risk Management Policy (AIMP-RMP-006)

## Approval & Sign-Off

| Chief Information Security Officer | Chief Technology Officer | CEO |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Resource & Cost Runaway Playbook

*Controlling Unplanned AI Compute & Cost Escalation Events*

AIPB-023 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-023 |
| **Playbook Owner** | Chief Technology Officer / Chief Financial Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | MEDIUM–HIGH — Financial & Operational Risk |

## 2. Scenario Description

### Overview

An AI resource and cost runaway event occurs when AI compute consumption, API costs, or cloud infrastructure spending escalates uncontrollably beyond planned and approved budgets. This may be caused by: AI training jobs running indefinitely due to misconfiguration, AI inference traffic spike not caught by auto-scaling limits, API rate limit abuse by internal or external users, AI agent entering an infinite loop and consuming resources continuously, misconfigured auto-scaling leading to excessive provisioning, or a prompt injection attack triggering excessive AI API calls.

### Detection Indicators / Trigger Conditions

- Cloud billing alert: AI-related spend exceeds defined threshold (e.g. 150% of daily budget)
- AI training job has been running for significantly longer than expected without completion
- AI API call volume spikes dramatically beyond normal traffic patterns
- AI agent task queue grows unboundedly without completion
- Cloud cost monitoring dashboard shows GPU/TPU utilisation at 100% without corresponding business demand
- Finance alerts on AI cloud spend tracking 300%+ above monthly budget
- Auto-scaling has provisioned a large number of AI inference instances without traffic justification

## 3. Immediate Response Actions

## PHASE 1 — DETECT & STOP (0–30 mins)

1. Automated cost alert received — confirm this is genuine runaway, not a planned activity
2. Immediately terminate or pause the resource-consuming AI jobs, training runs, or agent tasks
3. Implement emergency resource caps at the cloud provider level (budget alerts + hard limits)
4. Identify the root cause of the runaway: infinite loop, traffic spike, misconfiguration, or attack?
5. Notify the CTO, CFO, and AI Operations Lead

## PHASE 2 — ASSESS FINANCIAL IMPACT (30 mins – 4 hrs)

6. Calculate total cost incurred during the runaway event
7. Assess whether the runaway was caused by internal misconfiguration, external attack, or traffic spike
8. If caused by a prompt injection or external attack: activate the security incident response process (PB04/PB22)
9. Engage cloud provider account team: request billing review and credits where the runaway was caused by platform issues
10. Brief Finance on revised cost forecast and budget impact

## PHASE 3 — ROOT CAUSE & FIX (4–48 hrs)

11. Conduct detailed root cause analysis of the runaway event
12. Implement the specific fix: correct the misconfigured job, fix the infinite loop, implement rate limiting, or patch the security vulnerability
13. Deploy the fix in a test environment and validate that resource consumption is within expected bounds
14. Implement hard resource limits (maximum runtime, maximum cost threshold, maximum concurrent instances) for all AI jobs

## PHASE 4 — CONTROLS & PREVENTION (1–14 days)

15. Implement budget-based kill switches: AI jobs that exceed budget threshold are automatically paused pending human review
16. Deploy real-time cost monitoring dashboard for all AI compute with configurable alert thresholds
17. Implement mandatory cost estimation and approval for all AI training jobs above defined cost threshold
18. Review auto-scaling configurations: implement maximum scale limits and scale-in time requirements
19. Implement API rate limiting for all internal and external AI API consumers

## PHASE 5 — GOVERN (14–30 days)

20. Update AI budget governance framework to include runaway prevention controls as a standard requirement
21. Report cost runaway incident and financial impact to Finance Committee and AI Governance Committee
22. Implement quarterly AI cost optimisation reviews
23. Update cloud architecture standards to mandate hard cost limits on all AI workloads

# Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| **Cost alert & job termination** | AI Ops / Cloud Ops | CTO | CFO, AI Lead | CEO |
| **Financial impact assessment** | CFO / Finance | CFO | CTO, AI Lead | CEO, Board |
| **Root cause analysis** | AI Ops / ML Engineer | AI Ops Lead | CTO, Finance | CAO |
| **Technical fix & validation** | ML Engineer / Cloud Ops | CTO | AI Lead, Finance | Business Owners |
| **Cost control implementation** | Cloud Ops / Finance | CTO / CFO | AI Lead | AI Governance Committee |
| **Governance & reporting** | Finance / AI Governance | CFO / CTO | CAO | Board |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to terminate runaway job** | <30 minutes from alert | Incident log |
| **Cost runaway financial exposure** | <$5,000 above threshold before termination | Cost monitoring dashboard |
| **AI jobs with hard cost limits** | 100% | Cloud resource policy audit |
| **Budget alert coverage** | 100% of AI compute spend | Cloud billing audit |
| **AI cost forecast accuracy (monthly)** | Within ±10% of actual | Finance report |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |
| | | |

## Related Documents

- AI System Outage Playbook (AIPB-005)
- AI Change Management Policy (AIMP-CMP-011)
- AI Information Security Policy (AIMP-ISP-007)

- Prompt Injection Adversarial Playbook (AIPB-004)
- AI Risk Management Policy (AIMP-RMP-006)

## Approval & Sign-Off

| Chief Technology Officer | Chief Financial Officer | Chief AI Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Multi-Agent Coordination Failure Playbook

*Responding to Failures in AI Agent Orchestration & Coordination*

AIPB-024 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-024 |
| **Playbook Owner** | Chief AI Officer / Chief Technology Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Operational, Safety & Integrity Risk |

## 2. Scenario Description

### Overview

Multi-agent AI coordination failures occur when multiple AI agents working in concert — either within an orchestrated AI pipeline or across interconnected AI systems — produce harmful, unintended, or contradictory outcomes. Failure modes include: agent communication breakdown causing incorrect action sequences, agents developing conflicting goals and taking mutually harmful actions, orchestrator agent failures causing sub-agents to behave erratically, agents amplifying each other's errors, deadlock scenarios, and emergent behaviours not present in individual agents that arise from their interaction.

### Detection Indicators / Trigger Conditions

- Orchestrated AI pipeline produces outputs that contradict each other across agent stages
- AI agent task queue enters deadlock — tasks neither complete nor fail
- Sub-agent receives corrupted or malformed instructions from orchestrator agent and takes unexpected actions
- AI agents compete for the same resource, causing one or both to fail
- Multi-agent system produces actions that individually seem valid but collectively cause harm
- Monitoring detects AI agent communication latency exceeding timeout thresholds causing cascade failures
- AI system audit reveals agents took actions not authorised by the original task specification

## 3. Immediate Response Actions

## PHASE 1 — HALT & ASSESS (0–1 hr)

1. Immediately halt all affected AI agents and the orchestration pipeline
2. Preserve the full state of each agent at the time of failure: task queue, inputs, outputs, communications log
3. Identify which agents are involved and map the interaction and dependency graph
4. Assess what actions the agents have already taken — are any irreversible actions pending or completed?
5. Notify the Chief AI Officer, CTO, and the relevant business owner

## PHASE 2 — DAMAGE ASSESSMENT (1–6 hrs)

6. Audit all actions taken by each agent during the coordination failure window
7. Identify any irreversible actions (financial transactions, communications sent, API calls made) and their impact
8. Determine whether any of the agents' actions caused harm to data, systems, or individuals
9. Assess whether the failure is in the orchestration layer, the communication protocol, individual agent logic, or emergent interaction
10. Engage ML engineers for technical root cause analysis

## PHASE 3 — INVESTIGATE (6–48 hrs)

11. Replay the agent interaction sequence in an isolated environment to reproduce and understand the failure
12. Review agent communication logs for malformed instructions, race conditions, or conflicting directives
13. Assess whether the failure mode was tested for during development and why testing did not catch it
14. Determine whether this is a one-time edge case or a systematic architectural flaw
15. Consult multi-agent AI safety literature and frameworks for known failure modes of this type

## PHASE 4 — REMEDIATE (1–30 days)

16. Redesign the agent coordination protocol to include explicit conflict resolution mechanisms
17. Implement circuit breakers that halt the entire pipeline when an individual agent fails
18. Add human approval gates for all irreversible multi-agent actions above defined risk threshold
19. Redesign agent communication to include message verification and retry logic
20. Conduct comprehensive multi-agent integration testing, including fault injection and adversarial scenarios

## PHASE 5 — GOVERN (30–90 days)

21. Update AI system design standards to require multi-agent coordination risk assessment
22. Implement multi-agent system monitoring as a standard operational requirement
23. Report to AI Governance Committee with architectural recommendations
24. Engage with AI research community on multi-agent coordination safety standards

# Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|
| **Agent halt & state preservation** | AI Ops / ML Engineer | AI Ops Lead | CTO, AI Lead | CAO |

| | | | | |
|---|---|---|---|---|
| **Damage assessment** | AI Lead / Business Owner | CTO | Legal, Risk | CAO, CEO |
| **Technical root cause** | ML Engineer / AI Architect | AI/ML Lead | CTO | CAO |
| **Architectural redesign** | ML Architect | CTO | AI Lead, Security | CAO |
| **Integration testing** | QA / ML Engineer | AI/ML Lead | CTO | Business Owners |
| **Governance reporting** | AI Governance | CAO | CRO, CTO | AI Governance Committee, Board |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to halt failing multi-agent pipeline** | <1 hour from detection | Incident log |
| **Multi-agent integration test coverage** | 100% of orchestrated pipelines | Test suite audit |
| **Irreversible agent actions requiring human approval** | 100% gated for high-risk systems | Architecture audit |
| **Multi-agent coordination failures per quarter** | Trending toward zero | Operations monitoring |
| **Fault injection testing frequency** | Quarterly for all production multi-agent systems | Test schedule |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |
| | | |

## Related Documents

- Autonomous Decision Override Failure Playbook (AIPB-018)
- AI System Outage Playbook (AIPB-005)
- AI Risk Management Policy (AIMP-RMP-006)
- AI System Development Policy (AIMP-SDP-005)
- AI Resource Cost Runaway Playbook (AIPB-023)

## Approval & Sign-Off

| Chief AI Officer | Chief Technology Officer | Chief Risk Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Workforce Impact Crisis Playbook

*Managing Organisational & Human Impact of Significant AI-Driven Workforce Change*

AIPB-025 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-025 |
| **Playbook Owner** | Chief People Officer / Chief AI Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Human, Reputational & Organisational Risk |

## 2. Scenario Description

### Overview

An AI workforce impact crisis arises when AI automation, augmentation, or displacement significantly affects employees — through large-scale redundancies driven by AI, material changes to job roles, skills obsolescence, employee wellbeing crises linked to AI supervision, or widespread organisational anxiety about AI's impact on job security. This playbook also covers the reputational and regulatory risks of managing AI-driven workforce change without adequate human care, transparency, or due process.

### Detection Indicators / Trigger Conditions

- Board or executive decision to automate a significant number of roles using AI
- Employee surveys reveal widespread anxiety, disengagement, or fear related to AI adoption
- AI system deployed to monitor employee performance triggers grievances or union action
- Regulatory or media attention focused on the organisation's AI-driven workforce reduction
- Surge in HR cases, grievances, or sick leave linked to AI-related role changes
- External reports or whistleblowing alleging AI is being used to unfairly assess or dismiss employees
- Staff exodus or recruitment difficulties linked to perception of AI-first culture replacing human roles

## 3. Immediate Response Actions

## PHASE 1 — PREPARE & PLAN (Before announcement)

1. Conduct a comprehensive AI workforce impact assessment: which roles, what skills, what timeline?
2. Engage legal counsel: ensure all planned workforce actions comply with employment law, consultation requirements, and equality obligations
3. Design a human-first transition programme: retraining, redeployment, voluntary redundancy, outplacement support
4. Engage employee representatives and unions at the earliest legally required point — preferably earlier
5. Prepare a compassionate, honest, and detailed employee communication plan

## PHASE 2 — COMMUNICATE (Announcement phase)

6. Deliver face-to-face briefings from senior leaders before any written announcement
7. Provide honest, specific information: which roles are affected, what the timeline is, and what support is available
8. Establish dedicated Q&A channels (HR hotline, town halls, manager briefings) immediately
9. Issue external communications (media, investors, customers) that are accurate and demonstrate social responsibility
10. Activate employee assistance programme (EAP) and mental health support immediately

## PHASE 3 — SUPPORT (Transition period)

11. Launch retraining and upskilling programme for affected employees — prioritise internal redeployment
12. Provide individual career coaching and outplacement services for those unable to be redeployed
13. Monitor employee wellbeing metrics weekly: absence rates, engagement scores, grievance volumes
14. Maintain open, regular communication from senior leaders throughout the transition period
15. Conduct individual one-to-one meetings with every affected employee within 2 weeks of announcement

## PHASE 4 — GOVERN THE TRANSITION (Ongoing)

16. Ensure all AI performance monitoring of employees is disclosed, fair, and compliant with employment law
17. Review AI systems used in people management for bias and discriminatory impact (PB02)
18. Engage remaining workforce on their AI concerns and create structured input mechanisms
19. Report workforce AI impact metrics to the Board and AI Ethics Committee quarterly

## PHASE 5 — EMBED RESPONSIBLE AI WORK PRACTICES (Long-term)

20. Adopt a Responsible AI and Work charter — commit to retraining over redundancy as a first response
21. Establish a Workforce and AI Advisory Panel including employee representatives
22. Report annually on AI's net impact on workforce: roles created vs. displaced, retraining outcomes
23. Engage with government and industry on responsible AI workforce transition standards

# Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|
| **Workforce impact assessment** | HR / AI Lead / Finance | CPO / CAO | Legal, Unions | CEO, Board |

| | | | | |
|---|---|---|---|---|
| **Legal compliance review** | Legal / HR | CLO / CPO | Unions, External Counsel | CEO, Board |
| **Employee communications** | Comms / HR | CPO / CEO | CAO, Legal | All staff |
| **Retraining programme delivery** | L&D / HR | CPO | CAO, Line Managers | Affected employees |
| **Wellbeing monitoring** | HR / EAP | CPO | Mental Health Lead | Management |
| **AI workforce governance** | HR / AI Ethics / Legal | CPO / CAO | Ethics Officer | Board, AI Governance Committee |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Affected employees individually briefed** | 100% within 2 weeks of announcement | HR engagement log |
| **Internal redeployment rate** | ≥50% of displaced roles | HR tracking dashboard |
| **Retraining programme participation** | ≥80% of affected employees | L&D records |
| **Employee engagement score (post-announcement)** | Maintained ≥65 (out of 100) | Pulse survey |
| **AI-related grievances** | Trending toward zero | HR case management |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |

## Related Documents

- AI Ethics Policy (AIMP-ETH-013)
- Human Resources Security and AI Interaction Policy (AIMP-HRS-009)
- AI Training & Awareness Policy (AIMP-TAP-004)

- AI Bias Discrimination Playbook (AIPB-002)
- AI Reputational Crisis Playbook (AIPB-026)

## Approval & Sign-Off

| Chief People Officer | Chief AI Officer | Chief Legal Officer | CEO |
|---|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |

# AI Reputational Crisis Playbook

*Managing Organisational Reputation Under AI-Related Public Scrutiny*

AIPB-026 | Version 1.0 | 7 June 2026

## 1. Playbook Reference

| | |
|---|---|
| **Document ID** | AIPB-026 |
| **Playbook Owner** | Chief Communications Officer / Chief AI Officer |
| **Version** | 1.0 |
| **Effective Date** | |
| **Review Date** | |
| **Classification** | Confidential — Internal Use Only |
| **Status** | ☐ Draft ☐ Approved ☐ Under Review |

| | |
|---|---|
| **SEVERITY** | HIGH — Reputational, Commercial & Stakeholder Risk |

## 2. Scenario Description

### Overview

An AI reputational crisis occurs when the organisation's use, development, or deployment of AI becomes the subject of significant negative public, media, regulatory, or stakeholder attention in a manner that threatens brand equity, customer trust, investor confidence, or employee morale. Triggers include: harmful AI outcomes reported in media, AI ethics controversy, discriminatory AI systems exposed, AI-related data scandal, high-profile AI system failure, or widespread public concern about the organisation's AI practices. Reputational crises are amplified by social media and can escalate rapidly.

### Detection Indicators / Trigger Conditions

- Negative media coverage of the organisation's AI practices reaches significant scale
- Social media volume and sentiment regarding the organisation's AI crosses negative threshold
- Influential voices (NGOs, politicians, academics, celebrities) publicly criticise AI practices
- Regulatory statement or investigation generates media coverage of AI practices
- Customer-facing AI incident (bias, harm, data breach) receives public attention
- Viral social media campaign targeting the organisation's AI practices
- Employee public statements or whistleblowing attract media attention to AI practices

## 3. Immediate Response Actions

### PHASE 1 — ASSESS & MOBILISE (0–2 hrs)

1. Receive crisis alert from brand monitoring, media team, or leadership
2. Convene the AI Crisis Communications Team: CCO, CAO, CEO, Legal, Head of PR
3. Conduct rapid fact-finding: what is the allegation, is it accurate, what is the evidence?
4. Assess scale: what is the media reach, social media volume, and stakeholder exposure?
5. Classify severity: P1 (viral, mainstream media, customer harm) or P2 (industry/specialist media only)

### PHASE 2 — HOLD & PREPARE (2–6 hrs)

6. Issue a holding statement: acknowledge awareness, commit to investigation, avoid defensive language
7. Brief all designated spokespersons; all other staff directed not to comment to media
8. Activate the legal privilege protocol for all crisis communications
9. Prepare a full Q&A document for internal use covering all likely questions
10. Brief the CEO, board chair, and key investors before media engagement

### PHASE 3 — RESPOND (6–24 hrs)

11. Issue a full, factual, and human-centred response through official channels
12. Where harm has occurred: acknowledge it, apologise sincerely, commit to specific remediation
13. Avoid minimising, deflecting, or blaming individuals — focus on systemic remediation
14. Engage with key media for on-the-record briefing with senior spokesperson
15. Engage social media platforms directly to address viral content where possible
16. Brief key customers, partners, and investors personally before public statement

### PHASE 4 — REMEDIATE & DEMONSTRATE (1–90 days)

17. Take visible, specific, and credible remediation actions — not just communications
18. Engage independent external expert to review the AI practices at the centre of the crisis
19. Publish a transparency report on findings and remediation actions taken
20. Demonstrate ongoing commitment through policy updates, governance improvements, and stakeholder engagement
21. Monitor media and public sentiment continuously; adapt communications strategy as needed

### PHASE 5 — REBUILD & LEARN (90+ days)

22. Conduct a full post-crisis review: what caused it, how was it handled, what was the impact?
23. Implement systemic improvements to AI ethics, governance, and accountability frameworks
24. Publish an annual AI accountability report as a long-term trust-building measure
25. Engage with critics and stakeholders in ongoing dialogue — convert adversaries to constructive partners where possible
26. Update the AI reputational risk assessment and this playbook with lessons learned

## Roles & Responsibilities (RACI)

| Activity | R | A | C | I |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| **Crisis detection & mobilisation** | Brand Monitor / CCO | CCO | CAO, Legal, CEO | Board, Investors |
| **Holding statement** | CCO / Legal | CCO / CEO | CAO, External PR | Media |
| **Full public response** | CCO / CEO | CEO / CCO | CAO, Legal | All stakeholders |
| **Remediation actions** | CAO / CTO | CEO / CAO | Legal, Ethics Officer | Board |
| **Transparency report** | CAO / CCO | CEO | Legal, Ethics | Public, Regulators, Investors |
| **Post-crisis governance** | AI Governance | CAO / CCO | Ethics Officer, Legal | Board |

## Key Performance Indicators & SLAs

| KPI | Target | Measurement |
|---|---|---|
| **Time to issue holding statement** | <4 hours from P1 trigger | Comms log |
| **Brand sentiment recovery to baseline** | Within 90 days | Brand monitoring dashboard |
| **Customer retention during crisis** | ≥95% of key accounts retained | CRM data |
| **Transparency report published** | Within 60 days of crisis | Governance tracker |
| **Media narrative shift to remediation** | Within 14 days of full response | Media monitoring report |

## Post-Incident Review & Lessons Learned

A post-incident review must be completed within 5 business days of incident closure. Address: root cause, detection gap, response effectiveness, control gaps, and recurrence prevention.

| Lesson / Finding | Action Required | Owner & Due Date |
|---|---|---|
| | | |
| | | |
| | | |

## Related Documents

- AI Communication Management Policy (AIMP-CMM-012)
- AI Ethics Policy (AIMP-ETH-013)
- AI Bias Discrimination Playbook (AIPB-002)
- AI Generated Misinformation Playbook (AIPB-013)

- Deepfake Synthetic Media Attack Playbook (AIPB-015)

## Approval & Sign-Off

| Chief Communications Officer | Chief AI Officer | Chief Executive Officer |
|---|---|---|
| Name: Signature: Date: | Name: Signature: Date: | Name: Signature: Date: |