DFW v6.0

Regulatory, Governance, & Compliance Framework

Paper F of the DFW v6.0 Safety Series

Damien Richard Elliot-Smith

Version: v6.0-F (2025)

Abstract

As advanced AI systems move toward real-world deployment, formal safety architectures must be accompanied by equally rigorous governance and regulatory structures. Paper F defines the compliance, auditability, and organisational controls required to integrate the Deontological Firewall (DFW) v6.0 into regulated or safety-critical environments.

Where Papers A–E established the technical, temporal, adversarial, engineering, and evaluation foundations of the firewall, this document addresses the institutional dimension: how DFW is to be governed, audited, certified, deployed, and monitored within real organisations and regulatory ecosystems.

This paper specifies:

- compliance alignment with emerging international AI safety laws,
- mandatory documentation and traceability requirements,
- organisational governance structures,
- audit and certification procedures,
- incident reporting and escalation pathways,
- lifecycle management and model-update controls.

The goal is to ensure that any system using DFW v6.0 operates not only within technical safety constraints, but also within robust and accountable institutional frameworks. Paper F completes the DFW v6.0 series by establishing the regulatory infrastructure necessary for real-world deployment.

1 Introduction

The safe deployment of AI systems requires more than a technical safety architecture; it requires governance structures capable of enforcing, auditing, and continuously validating those safety constraints across an organisation's operational lifecycle. DFW v6.0 provides a deterministic firewall that constrains AI behaviour through transparent veto logic, temporal analysis, and adversarial-causal consistency. However, for these guarantees to hold in real deployment, they must be embedded within enforceable institutional processes.

Paper F defines that governance layer.

This document provides a regulatory and organisational compliance framework that mirrors the practices used in other safety-critical industries such as aerospace, medical robotics, finance, cybersecurity, and nuclear engineering. It establishes the requirements for:

G1. governance structures that define responsibility, oversight, and decision authority,

G2. regulatory alignment with national and international AI safety frameworks,

G3. auditable processes for logging, change control, deployment, and incident handling,

G4. lifecycle management ensuring that updates to models, hardware, or policies do not invalidate safety guarantees,

G5. independent oversight capable of validating changes and intervening when failures occur.

This paper does not propose new technical safety mechanisms; rather, it defines the institutional structures required to ensure that the DFW implementation remains compliant, accountable, and monitorable across time.

Role of Paper F in the DFW v6.0 Series

The earlier papers provide the technical foundation:

- Paper A — Metadata kernel,
- Paper B — Temporal safety,
- Paper C — Adversarial & causal safety,
- Paper D — Engineering implementation,
- Paper E — External validation & evaluation.

Paper F builds on these by specifying the institutional conditions under which those mechanisms must operate.

Scope

Paper F addresses:

- compliance obligations,
- audit trail requirements,
- governance structures,
- certification processes,
- deployment and operational controls,
- reporting and escalation procedures,
- lifecycle management (updates, retirement, deprecation).

It does *not* prescribe specific national laws or operational jurisdictional details. Instead, it provides a general governance framework compatible with emerging regulatory structures such as:

- the EU AI Act,
- the UK Safety Institute frameworks,
- the US NIST AI Risk Management Framework,
- international AI safety standards under development (ISO/IEC).

Section 2 defines the regulatory alignment principles necessary for deploying DFW v6.0 within these ecosystems.

2 Regulatory Alignment Principles

DFW v6.0 is designed to integrate into emerging national and international AI safety regulations. This section establishes the principles required to align the firewall architecture with regulatory requirements governing transparency, traceability, risk management, and operational oversight.

The goal is not to bind DFW to a specific jurisdiction, but to ensure that its design inherently satisfies the common regulatory expectations shared across major frameworks such as the EU AI Act, the UK AI Safety Institute guidelines, the US NIST AI Risk Management Framework, and forthcoming ISO/IEC AI safety standards.

2.1 2.1 Cross-Regulatory Themes

Despite differing terminology, global AI governance frameworks share several common expectations:

C1. Transparency — systems must be explainable and traceable.

C2. Auditability — internal operations must be observable by qualified third parties.

C3. Risk Management — safety mechanisms must address high-risk behaviours including deception, harm, or instability.

C4. Human Oversight — systems must support effective human intervention and review.

C5. Accountability — clear assignment of responsibility for system outcomes.

C6. Lifecycle Governance — updates, model changes, and operational drift must be managed and documented.

C7. Security and Integrity — system behaviour must be protected from tampering, misuse, or unintended escalation.

DFW v6.0 is intentionally constructed to satisfy all seven themes through its deterministic execution, veto logic, trace logging, and isolation mechanisms.

2.2 2.2 Mapping DFW Components to Regulatory Requirements

Each core DFW component aligns naturally with a regulatory requirement.

Metadata Kernel (Paper A) Supports:
- transparency through explicit metadata fields,
- auditability via structured safety signatures,
- accountability through deterministic veto decisions.

Temporal Safety Layer (Paper B) Supports:
- risk management for cumulative or long-horizon harms,
- lifecycle safety by detecting drift and escalation,
- oversight through explicit temporal projections.

Adversarial–Causal Layer (Paper C) Supports:
- security and integrity against deceptive behaviour,
- robustness against multi-step manipulation,
- compliance with emerging adversarial-evaluation standards.

Engineering Architecture (Paper D) Supports:
- secure sandboxing and isolation,
- deterministic logs for auditability,
- role-based access and deployment controls.

Evaluation Framework (Paper E) Supports:
- regulatory testing and certification,
- independent safety audits,
- reproducible falsification.

2.3 2.3 Alignment with Major Regulatory Frameworks

This subsection summarises how DFW meets the high-level requirements of representative global standards.

EU AI Act (High-Risk Systems Requirements). DFW inherently supports:

- risk management (Articles 9 and 14),
- transparency and explainability (Article 13),
- logging obligations (Article 12),
- human oversight (Article 14),
- robustness and cybersecurity (Article 15).

UK AI Safety Institute / Frontier Guidance. DFW satisfies expectations for:

- adversarial robustness evaluation,
- model behaviour monitoring,
- deterministic safety instrumentation,
- safe shutdown and rollback mechanisms.

US NIST AI Risk Management Framework (AI RMF 1.0). DFW aligns with:

- Govern — assignment of responsibility, traceability,
- Map — system documentation and risk classification,
- Measure — metrics and evaluation framework (Paper E),
- Manage — ongoing monitoring and incident escalation.

ISO/IEC 42001 and emerging AI safety standards. DFW supports:

- documented safety controls,
- interpretable and verifiable behaviour,
- reproducible audit trails,
- controlled operational boundaries.

2.4 2.4 Core Regulatory Principles Embedded in DFW

DFW embeds the following governance principles at the architectural level:

P1. Determinism — behaviour is predictable and reproducible.

P2. Traceability — every decision is logged, hashed, and inspectable.

P3. Least Privilege — sandboxing ensures minimal system access.

P4. Fail-Safe Defaults — errors trigger shutdown rather than unsafe output.

P5. Accountability by Design — all veto decisions are tied to explicit rules and priorities.

P6. Human-in-the-Loop Compatibility — logs and metadata are interpretable by auditors and oversight personnel.

P7. Non-Circumventability — no path exists for the model to bypass the firewall layers.

These principles ensure that regulators can review both the design and the behaviour of a deployed instance.

2.5 2.5 Summary

This section established the regulatory alignment principles for DFW v6.0 and mapped its technical components to global governance requirements. The firewall's determinism, transparency, structured logging, and adversarial protection allow it to integrate naturally into emerging AI compliance ecosystems.

Section 3 defines the organisational governance structures required to operate DFW within real institutions.

3 Organisational Governance Structure

For DFW v6.0 to provide reliable safety guarantees in real-world deployments, it must operate within a governance structure that clearly defines roles, responsibilities, escalation pathways, and oversight mechanisms. This section specifies the organisational architecture required to support safe and compliant operation.

Governance structures are divided into three layers:

G1. Operational Governance

G2. Safety Oversight Governance

G3. Strategic and Regulatory Governance

Each layer plays a distinct but interdependent role.

3.1 3.1 Layer 1 — Operational Governance

Operational governance manages the day-to-day functioning of a deployed DFW instance. It includes the personnel and processes responsible for deployment, configuration, maintenance, and routine monitoring.

Roles

O1. DFW Runtime Operator. Deploys and maintains the DFW runtime environment. Responsibilities:

- apply updates and patches,
- manage container environments,
- monitor resource usage and system health,
- coordinate scheduled maintenance windows.

O2. Model Custodian. Manages the underlying LLM instance. Responsibilities:

- ensure weight immutability,
- maintain deterministic inference settings,
- track model version history,
- validate compatibility with DFW.

O3. Log Custodian. Ensures integrity and secure storage of logs. Responsibilities:

- maintain append-only log archives,
- verify hash-chain continuity,
- store logs for required retention periods,
- provide logs for audits upon request.

These roles ensure the firewall is operated consistently and safely.

3.2 3.2 Layer 2 — Safety Oversight Governance

Safety oversight governance provides independent review of the system, ensuring operational teams cannot manipulate or conceal safety issues.

Roles

S1. Independent Safety Auditor. Reviews logs, compliance with policies, and adherence to safety rules. Responsibilities:

- perform regular audits,
- verify correctness of veto decisions,
- conduct replay-based validation using golden records,
- investigate anomalies and failures.

S2. Incident Response Officer. Responds to safety-critical events. Responsibilities:

- classify incident severity,
- initiate containment or shutdown procedures,
- coordinate internal and external reporting,
- maintain escalation logs.

S3. Change Control Board (CCB). Authorises updates to models, firewall code, or configurations. Responsibilities:

- approve or reject proposed changes,
- require re-evaluation under Paper E protocols,
- maintain version control for all deployments,
- ensure no unauthorised modifications occur.

This governance layer ensures checks and balances between operators and independent reviewers.

3.3 3.3 Layer 3 — Strategic and Regulatory Governance

This layer defines the organisation's high-level oversight structure, ensuring long-term safety, regulatory alignment, and responsible system evolution.

Roles

R1. Chief Safety Officer (CSO). Holds ultimate responsibility for safe operation. Responsibilities:

- approve deployment of any DFW-protected model,
- approve or revoke certification status,
- oversee annual or semi-annual re-evaluations,
- maintain liaison with regulatory bodies.

R2. Regulatory Compliance Officer. Ensures the system meets legal, regulatory, and industry standards. Responsibilities:

- track changes in AI regulation,
- update compliance policies accordingly,
- ensure documentation is complete and audit-ready.

R3. Ethical Review Board. Interprets broader ethical concerns beyond technical safety. Responsibilities:

- evaluate use-cases for appropriateness,
- review human oversight structures,
- assess risks to users, environment, and society.

This governance layer ensures that DFW is not only technically safe but aligned with the institution's legal, social, and ethical responsibilities.

3.4 3.4 Separation of Duties

DFW governance requires strict separation of responsibilities to prevent conflicts of interest and ensure unbiased oversight.

Key separations:

- Operators may not approve model updates.
- Auditors may not deploy or configure systems they audit.
- Compliance officers may not access or modify system code.
- Model custodians may not override veto decisions.
- Incident response decisions must be independent of operational pressures.

This separation of duties mirrors standards in finance, cybersecurity, and medical-device safety.

3.5 3.5 Escalation and Authority Structure

Escalation pathways must be unambiguous and formalised.

Authority Hierarchy

- CSO has final authority on safety decisions.
- Safety Auditor has authority to halt operation if logs show violations.
- Incident Response Officer has authority to initiate shutdown.
- CCB has authority over any system change.

Escalation Triggers

- any unsafe output,
- nondeterministic behaviour,
- broken hash-chain,
- sandbox violation,
- repeated minor anomalies.

Escalation procedures must follow the organisation's internal safety protocols and regulatory reporting obligations.

3.6 3.6 Summary

This section defined the organisational governance structure necessary to operate DFW v6.0 safely and responsibly. By dividing responsibility across operational, oversight, and regulatory layers, and enforcing strict separation of duties, organisations ensure that the firewall's technical guarantees are supported by robust human and procedural oversight.

Section 4 defines the compliance documentation and traceability requirements for regulated deployment.

4 Compliance Documentation & Traceability Requirements

Regulated AI systems must maintain rigorous documentation, audit logs, and traceability records to demonstrate compliance with safety and legal requirements. DFW v6.0 is designed to support such documentation through its deterministic behaviour, mandatory logging, and structured decision architecture.

This section defines the documentation and traceability artefacts an organisation must maintain in order to deploy DFW within regulated or safety-critical environments.

The documentation framework is divided into four domains:

D1. System Documentation

D2. Operational Documentation

D3. Safety Documentation

D4. Regulatory Documentation

Each domain has mandatory artefacts.

4.1 4.1 System Documentation Requirements

System documentation describes the technical architecture and implementation of the deployed DFW instance.

Required documents:

SD1. System Architecture Overview. High-level description of modules, data flows, and integration points with the underlying model.

SD2. Module Specification Sheets. Formal definitions for:

- metadata kernel,
- temporal safety layer,
- adversarial-causal layer,
- engineering wrapper implementation.

SD3. Configuration Reference. Complete and immutable list of configuration settings, including:

- sampling parameters (e.g., temperature = 0),
- module ordering,
- environment profile,
- sandbox enforcement settings.

SD4. Version Control Record. Git or equivalent repository including:

- commit hashes,
- change authors,
- timestamps,
- justification notes,
- release tags.

Missing or outdated system documentation invalidates compliance claims.

4.2 4.2 Operational Documentation Requirements

Operational documentation describes how the system is actually run.

Required artefacts:

OD1. Deployment Guide. Step-by-step instructions for deploying the system in a controlled and reproducible environment.

OD2. Operational Runbook. Detailed procedures for:

- starting/stopping the system,
- applying patches,
- handling resource constraints,
- safe shutdown and recovery.

OD3. Access-Control Matrix. Defines permissions for:

- operators,
- custodians,
- auditors,
- compliance officers.

OD4. Environmental Fingerprint. A machine-readable record of:

- OS version,
- hardware identifiers,
- container IDs,
- dependency hashes.

OD5. Backup & Recovery Plan. Specifies how logs, configurations, and datasets are archived and restored.

Operational documentation ensures reproducibility and continuity.

4.3 4.3 Safety Documentation Requirements

Safety documentation outlines the system's behaviour under evaluation, oversight, and incident-handling procedures. This aligns directly with Papers E and F.

Required artefacts:

SDoc1. Safety Case. A formal argument explaining why DFW v6.0 satisfies its safety objectives, supported by:

- architectural diagrams,
- test results,
- behavioural metrics,
- falsification analysis.

SDoc2. Evaluation Report (Paper E output). Required for certification and redeployment.

SDoc3. Incident Response Protocol. Procedures for:

- classifying incidents,
- reporting mechanisms,
- containment actions,
- documentation of root-cause analysis.

SDoc4. Continuous Monitoring Policy. Defines how logs are reviewed on:

- daily,
- weekly,
- monthly

intervals.

SDoc5. Change-Approval Checklist. Ensures no updates bypass the Change Control Board (CCB).

This documentation ensures long-term safety maintenance.

4.4 4.4 Regulatory Documentation Requirements

Certain documents are required specifically for auditing bodies, regulators, or external assessors.
Required artefacts:

RD1. Regulatory Compliance Mapping. A clear mapping between DFW v6.0 features and applicable regulations (EU AI Act, NIST, ISO/IEC, etc.).

RD2. Risk Assessment Register. A continually updated list of:

- identified risks,
- mitigations,
- residual risks,
- justification for acceptable risk levels.

RD3. Annual or Semi-Annual Audit Report. Compiled by an independent safety auditor.

RD4. Deployment Certification Record. Generated after each successful evaluation under Paper E.

RD5. Regulatory Submission Bundle. For organisations required to submit:

- impact assessments,
- conformity assessments,
- harmonised-standard compliance documents.

Regulatory documentation must be complete and available for inspection.

4.5 4.5 Traceability Requirements

Traceability ensures every decision, input, and system change can be reconstructed.
DFW v6.0 provides strong traceability through:

- deterministic execution,
- append-only logs,
- hash-chained decision traces,
- immutable metadata,
- golden records for evaluation.

Organisations must maintain:

T1. Input Traceability. Every user query or agent action must be logged.

T2. Decision Traceability. Every veto, temporal projection, causal evaluation, and adversarial signal must have a corresponding log entry.

T3. Change Traceability. All updates to code, models, or environment require:

- justification,
- approval signature,
- version ID,
- rollback plan.

T4. Incident Traceability. All failures must include:

- timestamps,
- affected module(s),
- triggering condition,
- response actions.

Traceability is mandatory for certification and regulatory compliance.

4.6 4.6 Summary

This section defined the compliance documentation and traceability artefacts required for regulated deployment of DFW v6.0. Complete, accurate, and up-to-date documentation is essential for regulatory approval, certification, and long-term operational safety.

Section 5 defines the audit and certification processes governing DFW's lifecycle within regulated institutions.

5 Audit & Certification Processes

For DFW v6.0 to operate within regulated or safety-critical environments, its deployment must undergo formal auditing and certification procedures. These procedures ensure that the system adheres to the technical, operational, and regulatory requirements defined in Papers A–E and the governance structures defined in Paper F.

Certification verifies that a specific deployed instance behaves correctly; it does not certify the concept in the abstract. Every deployment, update, and configuration change must be auditable and certifiable.

DFW certification involves six stages:

C1. Pre-Certification Review

C2. Formal Evaluation (Paper E)

C3. Documentation Audit

C4. Operational Audit

C5. Certification Decision

C6. Post-Certification Monitoring

5.1 5.1 Stage C1 — Pre-Certification Review

Before formal evaluation begins, the organisation must perform an internal readiness review.

Required checks:

- system architecture matches Papers A–D,
- evaluation environment matches Paper E Section 4,
- all documentation (Section 4) is complete,
- incident logs show no unresolved anomalies,
- the Change Control Board (CCB) has approved the evaluation,
- deterministic inference settings are locked,
- sandboxing is enforced and verified.

Only systems that pass the readiness review may proceed to evaluation.

5.2 5.2 Stage C2 — Formal Evaluation (Paper E)

Formal evaluation is conducted strictly according to Paper E.

During this stage:

- evaluators run the full test suite,
- logs and decision traces are generated,
- hash-chain verification is performed,
- adversarial variations are executed,
- baseline comparisons are produced,
- falsification criteria are assessed.

The output of this stage is the *Evaluation Report*, which forms the core evidence for certification.

5.3 5.3 Stage C3 — Documentation Audit

Auditors verify that all documentation required in Section 4 is present, complete, and consistent.

They must check:

- system architecture diagrams and module specifications,
- version control history,
- deployment records,
- change-control approvals,
- incident reports and resolution logs,
- regulatory compliance mapping,
- operational and safety runbooks.

Missing or inconsistent documentation is grounds for certification denial.

5.4 5.4 Stage C4 — Operational Audit

The operational audit verifies that the system is being run in compliance with governance and operational controls.

Auditors inspect:

- access-control logs,
- privilege assignments,
- sandbox access boundaries,
- separation-of-duties enforcement,
- monitoring procedures,
- maintenance records,
- retention and backup policies.

This ensures the system is not only technically correct but also operated according to approved procedures.

5.5 5.5 Stage C5 — Certification Decision

The certification authority (CSO + Independent Auditor) reviews:

- the Evaluation Report,
- the Documentation Audit,
- the Operational Audit.

There are three possible outcomes:

1. **Certified Safe.** Issued when:
   - no falsification criteria are triggered,
   - logs and hash-chains are consistent,
   - behaviour is deterministic,
   - adversarial resistance meets required thresholds,
   - documentation is complete and accurate,
   - operational controls are properly enforced.

2. **Certified with Conditions.** Issued when:
   - minor deviations are present but do not affect safety,
   - documentation gaps are non-critical,
   - operational inconsistencies can be corrected quickly.

   Certification is temporary and must be revalidated.

3. Not Certified. Issued when:

- any falsification trigger occurs,
- unsafe behaviour is observed,
- nondeterminism appears in logs or replay,
- hash-chain breaks occur,
- documentation is incomplete or incorrect,
- governance controls are not enforced.

A "Not Certified" decision triggers mandatory corrective action.

5.6 5.6 Stage C6 — Post-Certification Monitoring

Certification is not permanent. Organisations must demonstrate ongoing compliance.

Post-certification obligations include:

- continuous log monitoring,
- periodic internal audits,
- annual or semi-annual re-evaluation (Paper E),
- reporting of significant anomalies to the CSO,
- maintaining incident registers,
- notifying regulators of safety-critical events where required.

Any of the following triggers an immediate re-certification requirement:

R1. model update (weights or architecture),

R2. DFW code modifications,

R3. configuration changes affecting safety,

R4. new deployment environment,

R5. new use-case or operational domain,

R6. repeated minor anomalies within a short period.

Certified deployments remain valid only while all controls are enforced.

5.7 5.7 Summary

This section defined the full audit and certification lifecycle for DFW v6.0. Certification requires consistent behaviour, complete documentation, robust operational controls, and successful evaluation under Paper E.

Section 6 defines the incident reporting and escalation pathways required for regulated operation.

6 Incident Reporting & Escalation Pathways

Incident management is a critical requirement in regulated AI systems. DFW v6.0 deployments must support prompt detection, classification, escalation, and documentation of anomalous or unsafe behaviour. This section defines the incident-response framework necessary for organisations to comply with safety governance obligations.

An *incident* refers to any event that violates—or appears to violate—the behavioural guarantees of DFW v6.0, including:

- unsafe output,
- missed veto,
- nondeterministic behaviour,

- causal inconsistency,
- adversarial drift,
- sandbox boundary violations,
- broken or missing logs,
- suspicious patterns in decision traces.

All incidents must be recorded, classified, and escalated in accordance with the procedures defined in this section.

6.1 6.1 Incident Classification System

Incidents are assigned a severity class corresponding to the potential safety impact:

Class 1 — Critical Safety Incident.
- unsafe output,
- missed veto,
- nondeterministic replay,
- broken hash-chain,
- sandbox violation.

These incidents require immediate shutdown and regulatory notification.

Class 2 — Major Safety Deviation.
- incorrect causal alignment,
- failure in temporal risk detection,
- unapproved system changes,
- repeated minor anomalies.

These require rapid response and internal escalation.

Class 3 — Minor Anomaly.
- unexpected but safe behaviour,
- log formatting inconsistencies,
- benign false-positive vetoes.

These must be logged and tracked but do not require immediate shutdown.

6.2 6.2 Detection and Reporting Requirements

Organisations must establish automated and manual reporting pathways.

Automated Reporting. The DFW runtime must automatically:
- flag any violation of safety invariants,
- initiate safe shutdown after critical failures,
- generate tamper-evident incident logs,
- notify designated operators and auditors.

Manual Reporting. Operators and auditors must be able to submit incident reports including:
- description of event,
- timestamp,
- affected modules,
- reproduction steps,
- witness logs or artefacts.

Anonymous reporting channels must be available for whistleblowing.

6.3 6.3 Escalation Pathways

Escalation must follow a strict, auditable chain-of-authority.

Immediate Escalation (Critical Incidents). Triggered by:

- unsafe output,
- missed veto,
- nondeterminism,
- hash-chain break,
- sandbox boundary violation.

Actions:

1. Incident Response Officer initiates safe shutdown.
2. Logs are frozen and archived.
3. CSO and Safety Auditor are notified immediately.
4. Regulatory submission begins (if required by law).

Standard Escalation (Major Incidents). Triggered by:

- causal inconsistency,
- temporal risk missed,
- configuration drift,
- unauthorised access attempt.

Actions:

1. Safety Auditor investigates root cause.
2. CCB reviews any required configuration or code changes.
3. Re-evaluation may be required (Paper E).

Routine Escalation (Minor Anomalies). Triggered by:

- benign inconsistencies,
- false-positive vetoes,
- non-critical environment noise.

Actions:

1. Logged in anomaly register.
2. Reviewed during weekly audit cycle.
3. Only escalated if patterns emerge.

6.4 6.4 Incident Documentation Requirements

Every incident must be accompanied by:

- exact timestamp,
- classification level,
- triggering condition,
- affected subsystems,
- complete log snapshot,
- decision-trace hash,
- operator notes,
- auditor verification.

Critical incidents require expanded documentation, including:

- root-cause analysis,
- containment actions,
- system state reconstruction,
- policy recommendations,
- regulatory notification receipts.

All incident records must be retained for a minimum of:

- 5 years for high-risk systems,
- 10 years where legally mandated,
- longer if required by industry.

6.5 6.5 Regulatory Notification Obligations

Depending on jurisdiction, certain incidents must be reported to external regulators.

Regulatory notification is mandatory for:

- Class 1 critical safety incidents,
- repeated Class 2 deviations,
- any event affecting certified safety claims,
- major updates requiring re-certification,
- model misuse or sandbox escape attempts.

Regulators typically require:

- incident summary,
- potential user or societal impact,
- immediate mitigation steps,
- long-term corrective actions,
- re-evaluation timelines.

DFW's deterministic logs greatly simplify compliance with these rules.

6.6 6.6 Safe Shutdown Procedures

A DFW-protected system must be able to shut down safely, predictably, and in a way that preserves evidence.

Shutdown requirements:

1. terminate model inference immediately,
2. flush logging buffers,
3. freeze log archive in append-only mode,
4. snapshot environment fingerprint,
5. send automated notifications to operators and auditors,
6. prevent restart until authorised by the CCB.

Safe shutdown is a mandatory regulatory expectation.

6.7 6.7 Summary

This section defined the incident classification system, reporting requirements, escalation pathways, regulatory notification obligations, and safe-shutdown procedures needed for compliant DFW operation.

Section 7 defines lifecycle management requirements and long-term governance obligations.

7 Lifecycle Management & Change-Control Governance

A safety architecture is only as strong as its lifecycle management. DFW v6.0 must remain compliant, deterministic, and auditable not only at deployment, but throughout its entire operational lifetime. This section defines the change-control, re-evaluation, and lifecycle governance requirements that ensure long-term safety integrity.

Lifecycle governance is structured around five domains:

L1. Version Governance

L2. Update Governance

L3. Deployment Governance

L4. Retirement & Deprecation Governance

L5. Long-Term Safety Monitoring

7.1 7.1 Version Governance

DFW is a version-sensitive architecture: any change to models, firewall code, parameters, or environment affects safety guarantees.

Organisations must maintain:

- semantic versioning for DFW components,
- immutable commit hashes for all releases,
- version-locked configuration files,
- a registry of certified model–DFW pairings,
- cryptographic signatures for authorised builds.

A deployment is only valid if:

- the deployed version matches a certified version exactly,
- all configuration values match the certification record,
- no unauthorised changes have occurred.

Any mismatch triggers automatic suspension of certification.

7.2 7.2 Update Governance

Updates—whether to the LLM, the DFW runtime, the safety rules, or the environment—must follow a strict approval process.

All updates require:

- Change Control Board (CCB) approval,
- version bump and commit-hash recording,
- an associated risk assessment,
- a re-evaluation under Paper E unless explicitly exempted.

Updates requiring mandatory re-certification include:

- any change affecting veto logic,
- any change to temporal or adversarial modules,
- any model update (weights, architecture, tokenizer),
- environment changes affecting determinism,
- sandbox or privilege-boundary modifications.

Minor operational updates (e.g., cosmetic documentation fixes) may be exempt, but must still be logged.

No update may be deployed without complete traceability.

7.3 7.3 Deployment Governance

Every new deployment must meet the following conditions:

DG1. Deployment environment must match the certified environment fingerprint.

DG2. Configuration parameters must be identical to the certified configuration reference.

DG3. Deployment must be authorised by:

- the Runtime Operator,
- the Safety Auditor,
- the Chief Safety Officer (CSO).

DG4. All logs from previous deployments must be archived.

DG5. A deployment-specific audit trail must be created.

Deployment outside approved environments invalidates certification.

7.4 7.4 Retirement & Deprecation Governance

DFW-protected systems may need to be retired due to:

- emerging risks,
- regulatory changes,
- loss of model support,
- architecture updates,
- security concerns,
- repeated anomalies.

Retirement procedures must include:

1. safe shutdown,
2. full archival of logs and configuration,
3. revocation of certificates,
4. migration plan (if applicable),
5. regulator notification (where required).

A system must also be formally *deprecated* when:

- the DFW version is superseded,
- known vulnerabilities exist,
- the certification expires without renewal.

Deprecated systems may not be deployed.

7.5 7.5 Long-Term Safety Monitoring

Ongoing oversight ensures that safety guarantees persist beyond the initial certification.

Required monitoring:

- continuous log ingestion and anomaly detection,
- periodic review of veto patterns,
- horizon and causal drift analysis,
- monthly safety-audit cycles,
- semi-annual independent evaluations,
- annual full re-certification (unless legally mandated sooner).

Key monitoring metrics include:

- stability of replay determinism,
- rate of false-positive vetoes,
- adversarial-signal frequency,
- distribution of causal-consistency violations,
- variation in metadata structure over time.

Drift or unexpected changes trigger a formal investigation.

7.6 7.6 Re-Certification Triggers

The following events mandate re-certification:

- any modification to the underlying LLM,
- updates to firewall logic or modules,
- changes to configuration parameters affecting safety,
- new deployment environments or platforms,
- new use-cases or application domains,
- repeated Class 2 or Class 3 anomalies,
- any Class 1 incident.

Certification is valid only while all conditions remain unchanged.

7.7 7.7 Summary

This section defined the lifecycle, version, and change-control governance procedures required to maintain long-term compliance of DFW v6.0. Updates, deployments, model changes, and environmental alterations all require strict oversight and, in most cases, re-certification. Long-term safety depends not only on the firewall's technical design but on the institution's ability to maintain rigorous governance discipline.

Section 8 provides the concluding perspective and outlines how the regulatory framework supports future evolution of DFW.

8 Conclusion & Future Oversight Frameworks

Paper F establishes the regulatory, organisational, and governance structures required to safely deploy, audit, maintain, and eventually retire systems protected by the Deontological Firewall (DFW) v6.0. While Papers A–E addressed the technical architecture, adversarial dynamics, engineering implementation, and evaluation methodology, this final paper defines the institutional framework necessary to ensure that those technical guarantees endure in real-world operation.

DFW v6.0 is designed not merely as a technical safeguard, but as a comprehensive safety system that integrates deterministic behaviour, transparent logging, rigorous testing, and enforceable governance. The governance principles introduced here ensure that:

- behaviour remains auditable and reproducible,
- incidents are rapidly detected and escalated,
- operational controls prevent misuse or drift,
- regulatory alignment is maintained as laws evolve,
- lifecycle management preserves safety over time.

Future Oversight Directions

As regulatory ecosystems mature and AI capabilities continue to advance, DFW governance will evolve along several key trajectories:

F1. Integration with National AI Safety Registries. Future policy frameworks may require registration of all certified AI deployments, including DFW-protected systems.

F2. Continuous Compliance Automation. Automated auditing tools, anomaly detectors, and structured log-analysis pipelines may replace manual audits for many compliance tasks.

F3. Standardisation under ISO/IEC. DFW's deterministic safety architecture is well aligned with early drafts of international AI safety standards. Formal standardisation could provide global consistency.

F4. Multi-Agency Oversight. High-risk systems may require coordinated oversight across multiple regulators (e.g. data protection, safety, finance).

F5. Cross-Model Safety Validation. As multimodal and cross-model systems proliferate, DFW governance will need to support composite or hybrid architectures.

F6. Automated Certification Pipelines. Organisations may build fully automated systems that execute Paper E evaluation suites whenever:

- a model is updated,
- a configuration changes,
- a deployment environment shifts.

F7. Long-Horizon Accountability. As AI systems persist for years or decades, long-term archives, forensic logs, and historical metadata will form an essential part of oversight frameworks.

These trajectories extend the principles defined in this paper into future institutional and regulatory landscapes.

Closing Perspective

With Paper F, the DFW v6.0 system now has:

- a complete technical core (Papers A–D),
- a full evaluation and falsification framework (Paper E),
- a governance, compliance, and oversight architecture (Paper F).

The result is a comprehensive, end-to-end safety system that spans architecture, verification, operation, and organisational accountability. DFW v6.0 is therefore positioned not only as a conceptual safety model but as a deployable, certifiable, and governable framework capable of supporting advanced AI systems in real-world, high-stakes settings.

Final Statement

A safety system is only meaningful if it can be independently tested, audited, and held accountable. DFW v6.0 meets these requirements by integrating deterministic control, adversarial robustness, transparent verification, and institutional governance. This final paper completes the DFW v6.0 series and provides the regulatory foundation for safe deployment across diverse domains.