

Cyber Security Governance of AI-Related Integrity Risks at the Intersection of AI/ML and Quantum Computing

University of Oxford
2026



A dissertation submitted for the
MSc in Software and Systems Security

Abstract

The growing integration of Artificial Intelligence (AI) and Machine Learning (ML) into organisational systems has introduced integrity risks that challenge conventional cyber security governance assumptions. AI-specific standards, including ISO/IEC 22989, ISO/IEC 23053:2022, ISO/IEC 42001, and ISO/IEC 23894, emphasise lifecycle governance, accountability, traceability, and risk management. However, it remains unclear whether established organisational cyber security frameworks, including ISO/IEC 27001 and ISO/IEC 27005, adequately represent these risks within asset registers, risk registers, monitoring arrangements, and control ownership structures. This dissertation evaluates that problem through a case-study-based analysis of AI-related integrity risk at the intersection of AI/ML and quantum computing.

The study adopts a qualitative analytical methodology grounded in Design for Security (DES) and ISO/IEC 27005 risk logic. DES and ISO/IEC 27005 form the central methodological structure, while MSc-derived lenses from Classical Machine Learning, Malware Analysis, Cyber Threat Intelligence, Security Incident Management, and Quantum Cryptography refine specific analytical judgements concerning probabilistic model behaviour, adversarial manipulation, likelihood assessment, detection and escalation, and strategic cryptographic agility. A five-stage modelling framework is applied to a hypothetical organisational case study involving an AI-enabled fraud detection system. The analysis identifies assets, threats, vulnerabilities, risks, controls, and governance responsibilities. Figure 4 serves as the central analytical architecture, mapping risk entry points, propagation paths, control visibility, and governance artefacts across operational, lifecycle, monitoring, governance, and strategic cryptographic layers.

Within the limits of the case-study analysis, the findings indicate that established governance frameworks may remain structurally useful but become operationally incomplete where AI systems are not explicitly modelled as integrity-critical lifecycle assets. Risks such as adversarial input manipulation, training data poisoning, model extraction, behavioural drift, false assurance from performance metrics, and governance omission may otherwise remain under-represented. Quantum computing is treated as a strategic governance context that reinforces cryptographic agility, provenance awareness, and forward-looking risk planning, rather than as a second operational case domain.

The dissertation concludes that established organisational cyber security governance frameworks provide a useful but incomplete foundation for AI-related integrity risk governance. They should not be treated as automatically sufficient for AI-enabled systems. In this case-study context, adequacy depends on integrating AI-specific asset classification, lifecycle controls, behavioural monitoring, adversarial assurance, provenance and traceability mechanisms, control ownership, assurance review, and strategic cryptographic agility planning.

Acknowledgements

I would like to express my sincere gratitude to my supervisor for his guidance, constructive feedback, and continued support throughout the development of this dissertation. His insights helped refine the research focus, strengthen the analytical direction of the study, and clarify the importance of moving beyond descriptive discussion towards structured governance evaluation.

I am also grateful to the lecturers and instructors of the MSc in Software and Systems Security, whose teaching across the programme, particularly in Design for Security, Classical Machine Learning, Malware Analysis, Cyber Threat Intelligence, Security Incident Management, and Quantum Cryptography, significantly shaped the intellectual and methodological foundations of this research.

Finally, I would like to thank my family and friends for their encouragement and patience throughout this academic journey.

Table of Contents

ABSTRACT	2
ACKNOWLEDGEMENTS	3
TABLE OF CONTENTS	4
DECLARATION.....	7
CHAPTER 1: INTRODUCTION.....	8
1.1 CONTEXT OF THE STUDY	8
1.2 PROBLEM STATEMENT AND MOTIVATION	8
1.3 AIM AND SCOPE	9
1.4 SIGNIFICANCE OF THE STUDY	10
1.5 RESEARCH QUESTION	11
1.6 OVERVIEW OF THE STUDY	11
CHAPTER 2: LITERATURE REVIEW	12
2.1 INTRODUCTION	12
2.2 AI GOVERNANCE STANDARDS AND ORGANISATIONAL RISK CONTEXT	13
2.3 AI/ML AND INTEGRITY RISK	17
2.4 QUANTUM COMPUTING AND CRYPTOGRAPHIC DISRUPTION	18
2.5 CTI AND INCIDENT MANAGEMENT.....	19
2.6 DES AS AN EVALUATION FRAMEWORK	19
2.7 IDENTIFIED GOVERNANCE GAP.....	20
2.8 CHAPTER SUMMARY	22
CHAPTER 3: METHODOLOGY.....	23
3.1 INTRODUCTION	23
3.2 RESEARCH DESIGN	23
3.3 CONCEPTUAL FOUNDATIONS INFORMING THE METHOD	25
3.4 STRUCTURED RISK MODELLING FRAMEWORK	27
3.5 CASE STUDY STRATEGY	29
3.6 VALIDITY, RELIABILITY AND LIMITATIONS	29
3.7 CHAPTER SUMMARY	30
CHAPTER 4: CASE STUDY IMPLEMENTATION AND RISK ANALYSIS	31
4.1 INTRODUCTION	31
4.2 ORGANISATIONAL CONTEXT.....	31

4.3 STAGE 1: ASSET IDENTIFICATION	34
4.4 STAGE 2: THREAT IDENTIFICATION	37
4.5 STAGE 3: VULNERABILITY IDENTIFICATION	39
4.6 STAGE 4: RISK ASSESSMENT	41
4.7 STAGE 5: CONTROL EVALUATION	43
4.8 MITIGATION DESIGN AND GOVERNANCE AUGMENTATION	48
4.9 GOVERNANCE EVALUATION OUTCOME	49
4.10 CHAPTER CONCLUSION AND TRANSITION	50
 CHAPTER 5: DISCUSSION AND CONCLUSION	 51
5.1 INTRODUCTION	51
5.2 RECONCEPTUALISING INTEGRITY IN AI-ENABLED SYSTEMS	51
5.3 EVALUATION OF FRAMEWORK ADEQUACY	52
5.4 DETERMINISTIC GOVERNANCE BIAS	53
5.5 CONTRIBUTION TO KNOWLEDGE AND PRACTICE	54
5.6 QUANTUM COMPUTING AS STRATEGIC GOVERNANCE CONTEXT	55
5.7 LIMITATIONS AND FUTURE RESEARCH	55
5.8 FINAL RESPONSE TO THE RESEARCH QUESTION	56
5.9 REFLECTION OF THE RESEARCH PROCESS	57
5.10 CONCLUDING STATEMENT	57
 REFERENCES	 59
 APPENDICES	 61
ANNEX A RISK MANAGEMENT	61
ANNEX B PROJECT MANAGEMENT	63
ANNEX C ETHICAL, LEGAL, SOCIAL AND PROFESSIONAL ISSUES	64

List of Figures

Figure 1. UML Use Case Diagram of Governance Actors and System Interactions	33
Figure 2. System Architecture of the AI-Enabled Fraud Detection Environment.....	34
Figure 3. Stakeholder Interaction and Governance Responsibility Model	36
Figure 4. Analytical Architecture of AI Integrity Governance: Risk Entry Points, Propagation Paths and Control Visibility.....	46

List of Tables

Table 1. Literature Review Findings	21
Table 2. Methodological Approach Selection and Justification.....	25
Table 3. MSc Programme Learning and Dissertation Application Matrix	27
Table 4. Methodological Framework Overview.....	28
Table 5. Integrated Asset Register	35
Table 6. Stakeholder Responsibility Mapping.....	36
Table 7. Threat Identification and Actor Profiling	38
Table 8. Vulnerability Identification	40
Table 9. Risk Lookup	41
Table 10. Risk Assessment	42
Table 11. Risk Treatment Strategy	44
Table 12. Security Controls and Residual Risks.....	46
Table 13. Standards and Analytical Relevance to Figure 4.....	47
Table 14. Risk Register	62

Figure note: All figures in this dissertation were created by the author using Microsoft PowerPoint.

List of Abbreviations

AI	– Artificial Intelligence
CML	– Classical Machine Learning
CTI	– Cyber Threat Intelligence
DES	– Design for Security
ISMS	– Information Security Management System
ISO	– International Organization for Standardization
MAL	– Malware Analysis
ML	– Machine Learning
PQC	– Post-Quantum Cryptography
QUC	– Quantum Cryptography
SIM	– Security Incident Management

Declaration

I declare that this dissertation is my own original work and has not been submitted previously, in whole or in part, for the award of any other degree or diploma at this or any other institution of higher learning.

All information derived from published or unpublished sources has been duly acknowledged in the text and listed in the references section. Where work from others has been used, it has been appropriately referenced and cited in accordance with academic conventions.

Generative AI tools were used only for non-substantive support, including proofreading, editing, formatting assistance, organisation of wording, and clarity of expression. No generative AI tool was used to conduct the literature review, analyse data, formulate the research question, generate findings, or replace the author's own critical judgement. The author reviewed, verified, and revised all AI-assisted outputs and remains solely responsible for the research design, analysis, argument, citations, and final submitted text.

This dissertation is submitted in partial fulfilment of the requirements for the degree of Master of Science in Software and Systems Security at the University of Oxford.

Signed: _____

Full Name: _____

Date: _____

Chapter 1: Introduction

1.1 Context of the Study

Over the past decade, organisational dependence on Artificial Intelligence (AI) and Machine Learning (ML) has expanded from experimental analytics into operationally consequential decision environments, including fraud detection, anomaly identification, prioritisation, and automated classification. As AI deployment has matured, governance expectations have also evolved. Recent standards such as ISO/IEC 22989, ISO/IEC 23053, ISO/IEC 42001, and ISO/IEC 23894 indicate that AI systems should be governed through explicit concepts, lifecycle structures, management processes, and risk practices. They should not be treated merely as conventional software components [1], [2], [3], [4]. This development is analytically significant because AI-enabled systems introduce integrity risks that arise not only from unauthorised modification, but also from adversarial manipulation, training data corruption, feature pipeline distortion, and model drift.

Established cyber security governance frameworks remain important in this environment. ISO/IEC 27001 and ISO/IEC 27005 continue to provide the organisational structures through which assets are identified, risks assessed, controls selected, and governance responsibilities documented [5]. However, these frameworks are often operationalised in conventional information-system contexts where integrity compromise is commonly associated with unauthorised modification, observable intrusion, or control failure. The resulting question is whether established governance frameworks may remain conditionally adequate when the governed system is probabilistic, adaptive, and increasingly shaped by AI-specific governance expectations.

1.2 Problem Statement and Motivation

The central problem addressed in this dissertation concerns the potential misalignment between traditional cyber security risk management frameworks and the integrity risks introduced by AI-enabled systems. The governance issue is therefore not that AI-specific standards are absent, but that organisational cyber security risk management may still operationalise AI-enabled systems through assumptions inherited from deterministic information systems.

ISO-aligned governance frameworks provide structured mechanisms for identifying and mitigating risks. However, in many conventional applications of these frameworks, integrity compromise is treated primarily as unauthorised modification, observable intrusion, or control failure. In contrast, AI-related integrity risks may arise from adversarial perturbations, training data poisoning, model extraction, or gradual statistical drift. These forms of degradation do not necessarily involve code alteration or unauthorised access and may therefore evade conventional control logic.

The issue is not that governance frameworks are absent. Rather, the concern is whether existing frameworks explicitly model AI systems as integrity-critical assets and

incorporate ML-specific threat vectors into formal risk registers. If AI models are treated merely as standard software components, probabilistic vulnerabilities may remain underrepresented within governance processes.

The motivation for this study arises from increasing organisational reliance on AI systems for security-sensitive decision-making. Where AI outputs influence fraud detection, access control, anomaly classification, or operational risk scoring, integrity degradation may produce direct financial and operational consequences. At the same time, post-quantum cryptography guidance illustrates that assumptions about the long-term durability of classical public-key cryptography may change as quantum capabilities develop [6], [7]. Together, these dynamics suggest a need for systematic evaluation of governance adequacy under AI-enabled conditions.

1.3 Aim and Scope

The aim of this dissertation is to evaluate the adequacy of existing organisational cyber security risk management frameworks in addressing AI-related integrity risks at the intersection of AI/ML and quantum computing.

To evaluate governance adequacy, the study combines three complementary analytical lenses. AI-specific standards, including ISO/IEC 22989, ISO/IEC 23053, ISO/IEC 42001, and ISO/IEC 23894, provide the conceptual and lifecycle framing for understanding AI-enabled systems and their associated risk characteristics. ISO/IEC 27001 and ISO/IEC 27005 provide the organisational risk management process through which assets, threats, vulnerabilities, risks, and treatments are formally governed. DES, together with the MSc-derived tools from CML, MAL, CTI, SIM, and QUC, provides the evaluative method for examining whether those governance processes make AI-related integrity risks sufficiently visible, traceable, and controllable.

The dissertation evaluates governance adequacy rather than proposing new technical defences. Instead, it evaluates governance sufficiency within established organisational frameworks.

The broader topic of cyber security at the intersection of AI/ML and quantum computing is expansive. It encompasses technical robustness, regulatory compliance, AI ethics, post-quantum cryptographic migration, and national security strategy. To avoid conceptual dispersion, this dissertation adopts a deliberately narrow scope.

The study focuses specifically on:

- Organisational cyber security risk governance;
- AI-related integrity risks in AI/ML-enabled systems;
- Structured evaluation using DES principles [8]; and
- Assessment of framework adequacy rather than empirical prevalence.

The study does not:

- Conduct empirical adversarial ML experimentation;
- Develop post-quantum cryptographic algorithms;
- Analyse AI ethics or fairness in depth; and
- Undertake large-scale empirical surveys.

Quantum computing is included as a strategic governance context that reinforces the need for adaptive risk management, but the operational analytical focus remains AI-related integrity risk. This narrowing is intended to support analytical depth, coherence, and methodological proportionality.

1.4 Significance of the Study

The dissertation makes a modest case-study-based contribution to cyber security governance in three principal ways.

First, it reframes AI-related integrity risk as a governance-level concern. While adversarial ML research primarily addresses model robustness, comparatively less attention has been given to how such risks are formally embedded within organisational risk management processes [9]. By evaluating sufficiency of governance rather than technical countermeasures, the study shifts analytical focus from model performance to institutional modelling.

Second, the dissertation represents how DES principles can be applied as a structured evaluative framework for the selected AI-integrity governance problem. By integrating DES modelling with AI-specific governance standards and organisational cyber security risk governance, the study links the conceptual language of AI lifecycle management with the practical processes of asset classification, threat identification, risk assessment, and control evaluation. This is significant because AI-related integrity risk becomes visible only when lifecycle and behavioural elements are modelled explicitly. The study therefore contributes by showing how governance adequacy depends not merely on the existence of risk processes, but on whether those processes make AI-specific lifecycle and behavioural risks sufficiently visible within organisational risk management.

Third, the study contributes practically by showing that targeted governance augmentation may strengthen established organisational frameworks, but only where those frameworks are deliberately integrated with AI-specific standards, threat taxonomies, lifecycle controls, and assurance practices. The argument is therefore not that conventional frameworks are automatically adequate for AI, but that they can provide a governance foundation when AI-specific risks are explicitly classified, owned, monitored, and reviewed.

The dissertation's contribution lies not in proposing a new technical defence, but in applying DES, ISO/IEC 27005, and programme-derived analytical lenses to examine how AI-related integrity risks may remain insufficiently visible within organisational cyber security governance.

1.5 Research Question

The dissertation is guided by the following overarching research question:

To what extent do existing organisational cyber security risk management frameworks adequately address AI-related integrity risks at the intersection of AI/ML and quantum computing?

In answering this question, adequacy is assessed both against established organisational cyber security governance frameworks and against emerging AI-specific governance expectations concerning lifecycle visibility, risk ownership, traceability, and behavioural assurance.

1.6 Overview of the Study

This dissertation is structured to progress logically from conceptual framing to structured evaluation and synthesis.

Chapter 2 reviews AI-specific governance standards, cyber security risk management frameworks, adversarial ML, quantum cryptographic disruption, cyber threat intelligence, incident management, and security design principles, and identifies a governance gap concerning the explicit modelling of AI-related integrity risk.

Chapter 3 outlines the methodological approach, justifying the adoption of a structured DES-based modelling framework aligned with ISO risk management standards. It explains the case study strategy and establishes reliability, validity, and limitations.

Chapter 4 applies the methodology to a structured organisational case study deploying AI-enabled systems. Assets, threats, vulnerabilities, and risks are systematically identified and evaluated against existing governance controls.

Chapter 5 brings together the discussion and conclusion of the dissertation. It interprets the findings, evaluates framework adequacy, introduces the concept of deterministic governance bias, articulates the study's contribution to knowledge and practice, outlines limitations and future research, reflects on the research process, and provides the final response to the research question.

Together, these chapters provide a coherent and disciplined evaluation of cyber security governance at the intersection of AI/ML and quantum computing, with a deliberate focus on AI-related integrity risk.

Chapter 2: Literature Review

2.1 Introduction

The intersection of AI, ML, and quantum computing presents an evolving and complex domain for cyber security governance. Each technological development introduces distinct security considerations. When these technologies interact within organisational systems, they can intensify uncertainty within risk governance. While the broader research landscape encompasses technical robustness, algorithmic fairness, regulatory compliance, and post-quantum cryptographic transition, this dissertation adopts a deliberately narrowed focus on organisational cyber security risk governance, specifically AI-related integrity risks.

This chapter critically reviews the literature underpinning that focus. It examines foundational cyber security risk management frameworks, adversarial ML research, quantum cryptographic disruption, cyber threat intelligence and incident management integration, and structured security design principles. In addition to academic sources, this chapter draws conceptual insights from course materials that informed the development of this research: Security Incident Management (SIM), Malware Analysis (MAL), Classical Machine Learning (CML), Quantum Cryptography (QUC), Cyber Threat Intelligence (CTI), and DES [10], [11].

The review is structured around five areas: cyber security risk management frameworks; adversarial ML and probabilistic integrity; quantum computing and cryptographic disruption; CTI and incident management; and DES as a modelling approach. It argues that, while each area has a substantial body of scholarship, the literature reviewed in this dissertation gives less attention to how AI-related integrity risks are formally represented within organisational governance artefacts. This is especially true of the formal representation of AI-related integrity risks within organisational risk-modelling processes.

The chapter therefore identifies a governance gap: although adversarial ML vulnerabilities are well documented and ISO-aligned risk management frameworks are well established, the reviewed literature gives limited direct attention to whether existing governance structures adequately model AI-related integrity risks within organisational risk registers. This gap provides the basis for the methodological approach developed in Chapter 3.

The chapter first establishes the AI governance baseline through ISO/IEC 22989, ISO/IEC 23053:2022, ISO/IEC 42001, ISO/IEC 23894, and traceability-related standards such as ISO/IEC 5962, before examining how these expectations interact with established organisational cyber security risk management frameworks.

2.2 AI Governance Standards and Organisational Risk Context

The governance of AI-enabled systems now draws on two related bodies of practice. One consists of AI-specific instruments developed to address lifecycle accountability, risk oversight, and organisational responsibility [12]. The other consists of established cyber security governance frameworks used to manage information assets, threats, vulnerabilities, and enterprise risk. This dissertation considers both but evaluates adequacy primarily through established organisational cyber security risk management frameworks, while using AI-specific instruments as contextual benchmarks for emerging governance expectations.

This distinction is important because AI-specific instruments such as MITRE ATLAS, the NIST AI Risk Management Framework, the EU AI Act, ISO/IEC 42001, and ISO/IEC 23894 indicate that AI risks are increasingly treated through dedicated governance, risk, and assurance mechanisms rather than through conventional cyber security instruments alone [13]. Adversarial AI threat modelling is commonly treated through AI-focused taxonomies rather than through ordinary enterprise attack catalogues alone; AI risk management is addressed through AI-specific risk frameworks rather than through general cyber security frameworks alone; and AI regulation has developed through dedicated legal instruments rather than by relying solely on general data-protection law [14]. This dissertation therefore avoids the unsupported claim that existing cyber security frameworks are simply sufficient. It evaluates the narrower and more defensible claim that established organisational risk management structures may remain useful where they are assessed and adapted considering AI-specific governance expectations.

2.2.1 AI-Specific Standards and Terminology

AI-specific governance instruments are relevant because they make lifecycle accountability, monitoring, and risk ownership more explicit in relation to AI systems. While ISO/IEC 27001 and ISO/IEC 27005 remain important for organisational cyber security governance, AI systems introduce distinctive concepts, lifecycle dependencies, and risk characteristics that require explicit treatment. For that reason, this dissertation first considers AI-specific standards. These standards define the object of governance, the structure of ML-based AI systems, the management-system requirements for AI governance, and the risk management guidance applicable to AI-enabled environments.

ISO/IEC 22989 provides concepts and terminology for artificial intelligence [1]. Its significance for this dissertation lies in establishing a shared vocabulary for identifying AI systems, AI models, datasets, lifecycle processes, and related governance objects. This matters because AI-related integrity risk is difficult to govern effectively if the relevant assets are not first named and classified. In the case study, this standard supports the explicit treatment of the ML model, training dataset, validation dataset, and inference process as governance-relevant assets rather than as generic software components.

ISO/IEC 23053:2022 provides a framework for AI systems using machine learning [2]. It is relevant because it clarifies the structural relationship between data, model training, validation, deployment, inference, monitoring, and feedback. This lifecycle view is central to the dissertation's argument that AI integrity risk does not arise only at the point of system operation. It may enter through training data, feature engineering, retraining workflows, model deployment, or monitoring gaps. ISO/IEC 23053:2022 therefore supports the dissertation's lifecycle framing of AI-related integrity risk.

ISO/IEC 42001 provides requirements for an Artificial Intelligence Management System [3]. Its relevance is organisational rather than purely technical. It supports the expectation that AI governance should include defined roles, accountability, management review, documentation, and control processes. In this dissertation, ISO/IEC 42001 is used to bridge AI-specific governance with organisational cyber security governance by showing that AI systems require explicit management arrangements, not merely technical controls or informal engineering practices.

ISO/IEC 23894 provides guidance on AI risk management [4]. It is directly relevant to the research question because the dissertation evaluates whether organisational cyber security risk management frameworks are conditionally adequate to represent AI-related integrity risks. ISO/IEC 23894 reinforces the need to consider AI-specific sources of risk, lifecycle stages, uncertainty, and risk treatment. It therefore supports the analysis of adversarial manipulation, data poisoning, model drift, and governance omission as risk management concerns rather than isolated technical anomalies.

ISO/IEC 5962, which standardises SPDX, is relevant as a traceability and provenance mechanism [15]. Although SPDX is most associated with software bill of materials practice, its underlying logic of structured component, dependency, and metadata exchange is relevant to SBOM and emerging AIBOM-style assurance. For this dissertation, ISO/IEC 5962 is not treated as a primary AI risk management standard. Rather, it is used to support the argument that AI governance requires traceability over models, datasets, dependencies, model versions, and provenance claims [12]. This is particularly relevant to the case study's treatment of model registry, secure model distribution, and lifecycle assurance.

Together, these standards establish the AI governance baseline for the dissertation. ISO/IEC 22989 defines the terminology, ISO/IEC 23053:2022 clarifies the ML system lifecycle, ISO/IEC 42001 frames AI management-system governance, ISO/IEC 23894 supports AI risk management analysis, and ISO/IEC 5962 provides a traceability reference relevant to SBOM and AIBOM-style assurance [15]. This baseline is then connected to ISO/IEC 27001 and ISO/IEC 27005 in the following subsection, where organisational cyber security governance and risk management processes are examined as the mechanisms through which AI-related integrity risks are formally classified, assessed, and treated.

2.2.2 Organisational Cyber Security Governance and Risk Standards

Despite the growth of AI-specific guidance, established cyber security management systems and risk frameworks remain common organisational mechanisms for governing information-security risk. In this dissertation, ISO/IEC 27001 and ISO/IEC 27005 are therefore positioned as the organisational governance backbone through which AI-related integrity risks can be incorporated into formal cyber security risk management.

ISO/IEC 27001 provides the foundation for an Information Security Management System (ISMS) [5]. Its relevance lies in establishing organisational processes for defining security objectives, assigning responsibilities, maintaining documented information, implementing controls, and supporting continual improvement. In the context of AI-enabled systems, ISO/IEC 27001 remains important because it provides the institutional mechanism through which AI assets, controls, assurance activities, and governance responsibilities can be formally embedded. However, ISO/IEC 27001 does not by itself guarantee that AI models, datasets, feature pipelines, retraining workflows, or drift monitoring mechanisms will be recognised as integrity-critical assets. That depends on whether the ISMS is applied in a way that reflects the characteristics of AI systems.

ISO/IEC 27005 complements this by providing risk management logic for information security. Its asset–threat–vulnerability–risk–treatment structure remains highly relevant to this dissertation because it provides the procedural basis for evaluating governance adequacy. The standard supports a disciplined sequence: identify assets, determine threats, identify vulnerabilities, assess likelihood and impact, evaluate risk severity, and select appropriate treatment options. This sequence is used in the case study to assess whether AI-related integrity risks are made sufficiently visible within organisational risk processes.

The CIA triad remains central to this governance logic. Confidentiality, integrity, and availability provide the basic classification structure for understanding the security properties of organisational assets. In conventional systems, integrity is commonly associated with protection against unauthorised modification, corruption, or tampering. In AI-enabled systems, however, integrity must also include behavioural reliability under probabilistic and adversarial conditions. A model may remain available and unmodified at the code level while still producing degraded, manipulated, or unreliable outputs. Integrity must therefore be assessed not only at the level of data and software state, but also at the level of model behaviour, training-data quality, feature transformation, and inference reliability.

DES reinforces this governance logic by requiring explicit articulation of domains, system boundaries, interconnections, and trust assumptions. Rather than reducing governance to control checklists, DES supports evaluation of whether AI-related integrity risks are made visible within organisational risk processes. This makes it particularly useful for bridging AI-specific lifecycle concerns with established cyber security governance structures.

The asset–threat–vulnerability–risk–treatment logic is therefore retained in this dissertation. However, it is applied with AI-specific visibility. Assets include not only servers, applications, and databases, but also trained models, model parameters, training datasets, validation datasets, feature engineering pipelines, monitoring thresholds, and retraining workflows. Threats include not only malware, unauthorised access, and exfiltration, but also adversarial input manipulation, training data poisoning, model extraction, feature pipeline abuse, and governance omission. Vulnerabilities include not only software defects and misconfigurations, but also absent adversarial testing, weak dataset validation, undefined drift thresholds, incomplete model provenance, and failure to assign control ownership for AI lifecycle processes.

This framing positions ISO/IEC 27001 and ISO/IEC 27005 as organisational process and risk-structuring mechanisms within the dissertation. They provide the organisational process and risk-structuring logic through which AI-related integrity risks can be governed, while AI-specific standards provide the vocabulary, lifecycle framing, and governance expectations needed to interpret those risks appropriately.

2.2.3 Exploit-Based Threat Models and Malware Paradigms

Traditional cyber security risk modelling has been closely aligned with exploit-based threat paradigms. Malware analysis research supports how adversaries exploit software vulnerabilities to inject malicious code, escalate privileges, or disrupt system behaviour [16]. Within this paradigm, integrity compromise is typically event-driven and traceable to a discrete exploit vector, such as code tampering, privilege abuse, or malicious payload execution.

SIM frameworks are similarly structured around discrete incidents requiring detection, containment, and recovery. Incident classification presumes identifiable compromise events, whether through intrusion detection alerts, anomalous access logs, or confirmed malware infection. These models are well suited to conventional IT scenarios in which integrity degradation follows observable structural interference.

However, AI-related integrity risks do not always conform to this logic. Adversarial perturbations, poisoned training data, model extraction, and gradual model drift may all degrade outputs without malware execution, unauthorised code modification, or visible intrusion. In such cases, the system may continue to operate within authorised boundaries while producing unreliable results. Exploit-based models therefore remain necessary for understanding conventional cyber compromise, but they are insufficient on their own where integrity degradation occurs through authorised interactions, statistically valid system inputs, and probabilistic model behaviour rather than through direct system tampering.

2.2.4 AI Governance Context and Risk Visibility

The AI-specific standards discussed above are relevant not because they replace established cyber security governance frameworks, but because they sharpen the expectations against which governance adequacy should be assessed. They indicate that AI governance requires more than the generic documentation of information-

security risks. It also requires lifecycle visibility, traceability, accountability, monitoring, and explicit risk ownership across data, model, deployment, and assurance activities.

This matters because conventional ISMS processes may be formally mature while still leaving AI-related integrity risks insufficiently visible. In the case-study scenario developed later in this dissertation, a risk register may document infrastructure compromise, malware infection, or access-control failure, yet omit training-data poisoning, model drift, adversarial evasion, model extraction, or unclear ownership of monitoring thresholds. In this sense, the issue is not whether organisational risk governance exists, but whether it captures the lifecycle and behavioural conditions through which AI integrity risk actually materialises.

This synthesis provides the bridge between the standards discussion and the dissertation's central governance gap. AI-specific governance instruments establish the expectation that AI-related risks should be visible, traceable, owned, monitored, and reviewed. Organisational cyber security frameworks provide the process through which such risks can be formally managed. The question for this dissertation is therefore whether those organisational processes make AI-related integrity risks explicit enough to be governed effectively.

2.3 AI/ML and Integrity Risk

2.3.1 Probabilistic Foundations of Machine Learning

CML models derive predictive behaviour from statistical inference rather than explicit rule encoding [10]. Supervised learning involves training models on labelled datasets to approximate underlying data distributions. Model performance depends on generalisation capacity, feature stability, and data representativeness.

This probabilistic foundation complicates conventional interpretations of integrity. A model may produce compromised outputs not because its code has been altered, but because its decision boundary has been exploited or its training data distribution has shifted. Integrity degradation may therefore occur without unauthorised access or observable modification.

This probabilistic behaviour reinforces the need for lifecycle-aware governance consistent with ISO/IEC 23053:2022, which frames ML-based AI systems across data, training, validation, deployment, and monitoring activities, and with ISO/IEC 23894, which supports risk guidance for AI systems whose behaviour may change under uncertainty, adversarial pressure, or shifting data conditions [17].

2.3.2 Adversarial ML

Adversarial ML research supports that neural networks can be manipulated through carefully crafted inputs designed to exploit sensitivity within high-dimensional feature space [9]. This represents the limitations of defensive distillation strategies, highlighting the persistence of adversarial vulnerability [18].

Data poisoning attacks extend this risk into the training phase; whereby malicious actors introduce corrupted data points that alter model behaviour over time [19]. Model extraction attacks further support how repeated querying may reveal approximations of decision logic.

These attacks differ fundamentally from exploit-based malware paradigms. They operate within legitimate system interfaces and do not require code modification. Integrity compromise becomes behavioural rather than structural.

The MAL course materials emphasised adversarial adaptation in exploit development [16]. Although adversarial ML differs technically from traditional malware, the adversarial mindset remains conceptually consistent: attackers exploit systemic weaknesses, whether deterministic or statistical.

2.3.3 ML Lifecycle and Governance Integration

ML systems introduce lifecycle complexity absent in static software systems. Models require training, validation, deployment, monitoring, and periodic retraining [10]. Each stage introduces potential integrity vulnerabilities, including dataset corruption, hyperparameter misconfiguration and distribution drift.

DES emphasises the importance of modelling system evolution and trust boundaries explicitly. However, standard ISMS documentation may not always incorporate ML lifecycle processes within formal risk registers.

The governance issue therefore shifts from the mere existence of technical vulnerabilities to the completeness of risk modelling, especially the integration of ML lifecycle processes into governance artefacts.

These lifecycle governance expectations are consistent with ISO/IEC 42001, which supports documented accountability, management-system controls, and organisational oversight for AI systems, and with ISO/IEC 23894, which treats AI risk management as a lifecycle activity requiring explicit identification, assessment, treatment, monitoring, and review of AI-specific risks.

2.4 Quantum Computing and Cryptographic Disruption

Quantum computing introduces a strategic governance context rather than a second operational case domain for this dissertation [11]. Its relevance lies in showing that technological change can destabilise long-standing security assumptions, particularly in relation to classical public-key cryptography and the need for cryptographic agility [6], [7]. Recent post-quantum standardisation and migration guidance therefore matter as governance signals, not because this dissertation undertakes a technical study of post-quantum implementation, but because they reinforce the broader requirement that organisational risk management must remain adaptable under emerging

technological conditions. The operational analytical focus of this dissertation remains AI-related integrity risk.

2.5 CTI and Incident Management

CTI frameworks emphasise adversary profiling, capability assessment, and contextual threat analysis [20]. CTI enriches risk modelling by incorporating attacker intent, resources, and technical sophistication. Such modelling informs likelihood estimation within ISO risk matrices.

SIM frameworks, meanwhile, structure organisational response to detected compromise. However, both CTI and SIM assume identifiable adversarial campaigns or discrete incident events.

AI-related integrity degradation complicates this assumption. Adversarial perturbations may not trigger conventional intrusion detection mechanisms. Model drift may occur gradually without clear threshold breaches. The integration of AI behavioural monitoring into incident frameworks therefore represents a governance challenge rather than a purely technical one.

These CTI and SIM lenses are later operationalised in Figure 4 and the five-stage modelling structure, where attacker intent, capability, detection thresholds, escalation pathways, and incident response logic are mapped against AI-related integrity risks [21].

2.6 DES as an Evaluation Framework

DES provides a structured modelling approach that emphasises explicit articulation of domains, interconnections, trust assumptions, and system boundaries [8]. Unlike purely compliance-oriented interpretations of ISO standards, DES requires reflective modelling of system–environment interactions.

This modelling discipline is particularly valuable when evaluating AI-enabled systems. By forcing explicit identification of assets and trust dependencies, DES exposes potential blind spots in governance documentation. AI models, training datasets, and feature pipelines can thus be analysed as integrity-critical domains rather than treated implicitly as standard software components.

DES is valuable in this dissertation because it turns architectural representation into an analytical model rather than a descriptive diagram. By requiring explicit attention to domain boundaries, trust assumptions, interconnections, and technical–organisational dependencies, DES makes it possible to examine where AI-related integrity risks enter, how they propagate. It also allows evaluation of whether governance artefacts such as asset registers, risk registers, monitoring thresholds, and control ownership structures make those risks visible. Figure 4 is therefore used as an evaluative model of governance visibility, not merely as a system architecture.

2.7 Identified Governance Gap

The reviewed literature indicates that the central governance gap is not the absence of cyber security risk management frameworks, nor the absence of technical research on adversarial ML. Rather, the gap lies in the limited evaluation of whether established organisational cyber security risk management frameworks make AI-related integrity risks sufficiently visible within their asset–threat–vulnerability–risk architecture.

This gap becomes clearer when conventional organisational cyber security governance is considered alongside AI-specific governance standards. ISO/IEC 27001 and ISO/IEC 27005 provide a strong procedural basis for asset classification, risk assessment, treatment selection, and continual improvement. However, AI-specific standards such as ISO/IEC 22989, ISO/IEC 23053, ISO/IEC 42001, and ISO/IEC 23894 indicate that AI governance also requires lifecycle visibility, explicit accountability, traceability, monitoring, and integration of AI-specific risk management into organisational processes. The resulting issue is not simply whether AI-related threats can be named, but whether they are systematically represented, owned, monitored, and reviewed within organisational governance.

The literature reviewed in this chapter suggests a tension between established procedural risk management, probabilistic AI integrity vulnerabilities, ML lifecycle complexity, incident-oriented detection assumptions, and the broader need for governance adaptability under technological change. In practice, this means that AI models, training datasets, validation datasets, feature pipelines, retraining workflows, monitoring thresholds, and assurance artefacts may remain under-classified or implicitly governed even where an organisation operates within a mature ISMS.

Accordingly, the governance problem addressed in this dissertation is one of incomplete representation rather than framework absence. Existing cyber security governance frameworks may remain useful, but their adequacy depends on whether they explicitly model AI-related integrity risks across the AI lifecycle. This gap provides the rationale for the structured methodology developed in Chapter 3 and the case-based evaluation undertaken in Chapter 4.

The literature review findings are synthesised in Table 1. The purpose of the table is to make the evidence base and research gap explicit: each literature area provides useful concepts, but none alone fully explains how AI-related integrity risks should be represented within organisational asset registers, risk registers, monitoring arrangements, and control ownership structures.

Literature area	What was analysed	Main contribution to this dissertation	Limitation identified	Resulting governance gap
AI-specific standards	ISO/IEC 22989, ISO/IEC 23053, ISO/IEC 42001, ISO/IEC 23894 and traceability concepts	Defines AI terminology, ML lifecycle structure, AI management-system expectations, AI risk	These standards clarify AI governance expectations but do not by themselves show how an existing cyber	Need to translate AI lifecycle and accountability expectations into organisational risk artefacts.

		management guidance and traceability needs.	security risk register should represent each AI integrity risk.	
Organisational cyber security governance	ISO/IEC 27001, ISO/IEC 27005, CIA, asset-threat-vulnerability-risk-treatment logic	Provides the procedural foundation for risk assessment, control selection, ownership and continual improvement.	May remain AI-implicit if models, datasets, feature pipelines and monitoring thresholds are not classified as integrity-critical assets.	Need to test whether conventional risk processes make AI-specific assets and failure modes visible.
Adversarial ML and probabilistic integrity	Evasion, poisoning, extraction, drift and model-behaviour degradation	Shows that integrity compromise can occur through authorised inputs and statistical behaviour rather than code alteration alone.	Technical literature often focuses on robustness rather than governance representation, ownership or risk-register treatment.	Need to map adversarial ML risks into governance artefacts and assurance controls.
CTI, MAL and SIM	Attacker capability, adversarial adaptation, incident detection, escalation and recovery	Supports realistic threat modelling, likelihood reasoning and detection/escalation analysis.	Traditional incident logic may assume discrete compromise events, while AI degradation may be gradual, behavioural and hard to classify.	Need to extend monitoring and incident logic to AI behavioural assurance and drift detection.
Quantum and cryptographic disruption	Cryptographic agility, post-quantum exposure and long-term security-assumption change	Provides the strategic governance context for forward-looking risk planning and dependency awareness.	Quantum risk is not the operational case focus and should not dilute the AI integrity analysis.	Need to treat quantum as strategic governance pressure supporting cryptographic agility, not as a second case study.
DES and structured modelling	System boundaries, domains, trust assumptions, interconnections and socio-technical dependencies	Provides the evaluative method for exposing where risk enters, propagates and becomes visible or invisible to governance.	DES must be operationalised through concrete tables and diagrams to represent application rather than description.	Need for a structured case study that applies DES through assets, threats, vulnerabilities, risks, controls and Figure 4.

Table 1. Literature Review Findings

Table 1 clarifies the dissertation's gap. The problem is not the absence of cyber security governance frameworks or the absence of AI-specific standards. The problem is the incomplete translation of AI-specific lifecycle, behavioural and assurance expectations into the concrete artefacts through which organisations govern risk. This synthesis directly justifies the DES-based five-stage methodology in Chapter 3 and the case-study evidence in Chapter 4.

2.8 Chapter Summary

This chapter has reviewed the bodies of literature that underpin the dissertation: organisational cyber security risk governance, AI-specific governance standards, adversarial ML, quantum cryptographic disruption, CTI, incident management, and DES. The synthesis in Table 1 shows that the central gap is not a lack of frameworks, but the incomplete representation of AI-related integrity risks within organisational governance artefacts such as asset registers, risk registers, monitoring thresholds, assurance evidence, and control ownership structures [21], [16], [20].

It first examined ISO/IEC 22989, ISO/IEC 23053:2022, ISO/IEC 42001, ISO/IEC 23894, and ISO/IEC 5962 as standards relevant to AI terminology, ML lifecycle structure, AI management systems, AI risk management, and traceability. It then considered ISO/IEC 27001 and ISO/IEC 27005 as the organisational cyber security governance and risk management backbone through which AI-related integrity risks may be formally classified, assessed, and treated. The chapter also reviewed adversarial ML, quantum cryptographic disruption, CTI, incident management, and DES, identifying a governance gap concerning the explicit modelling of AI-related integrity risks within organisational risk processes.

These findings also prepare the ground for the programme-informed methodology developed in Chapter 3.

Chapter 3: Methodology

3.1 Introduction

This chapter outlines the methodological approach adopted to evaluate the adequacy of existing organisational cyber security risk management frameworks in addressing AI-related integrity risks at the intersection of AI/ML and quantum computing. As established in Chapter 1, the study does not seek to develop new adversarial defences or propose novel cryptographic techniques. Instead, it evaluates governance sufficiency within established organisational risk management structures.

The methodology is qualitative and analytical, using a DES-based modelling framework aligned with ISO/IEC 27005 [5], [8].

This chapter explains the research design, justifies the selection of DES as the primary modelling framework, outlines the structured five-stage evaluation process, describes the case study strategy, and addresses issues of validity and limitations.

3.2 Research Design

3.2.1 Research Approach

The research adopts a qualitative, evaluative design. The objective is to assess whether established organisational cyber security risk management frameworks adequately incorporate AI-related integrity risks. The study is therefore interpretive rather than empirical, and analytical rather than experimental.

This design is appropriate for several reasons. First, the research question concerns governance adequacy rather than model performance. Empirical adversarial testing would evaluate technical robustness but would not necessarily illuminate whether organisational risk registers formally recognise AI-related integrity risks. Second, the evaluation concerns framework sufficiency — an issue of conceptual alignment rather than quantitative measurement.

The study therefore evaluates not only whether AI-related integrity risks can be listed, but whether they can be made visible, prioritised, and governed in a traceable way through asset classification, threat modelling, vulnerability analysis, risk assessment, and control ownership.

The analysis proceeds by modelling assets, threats, vulnerabilities, and mitigations within a structured organisational context, thereby enabling systematic assessment of governance completeness.

3.2.2 Justification for a Structured Modelling Approach

DES was selected as the primary methodological framework because it emphasises explicit modelling of domains, interconnections, trust assumptions, and system

boundaries [8]. DES requires clear articulation of system–environment relationships and dependencies, thereby exposing implicit assumptions that may otherwise remain unexamined. This is particularly important for AI-enabled systems, where integrity risk may arise through authorised inputs, probabilistic model behaviour, lifecycle dependencies, or governance omission rather than through direct intrusion or unauthorised modification alone.

Standard ISO risk assessments provide procedural guidance for risk identification, assessment, and treatment [5]. However, in practice, risk processes may become operationally routinised and control oriented. DES adds an additional evaluative layer by requiring reflective modelling of how systems interact with their operational, organisational, and governance environments. In this dissertation, DES is therefore used to ensure that AI models, training datasets, validation datasets, feature pipelines, retraining workflows, monitoring thresholds, and model-signing mechanisms are treated as integrity-critical domains rather than implicitly as conventional software artefacts. The modelling process consequently becomes an instrument for testing governance completeness.

The selection of DES was also made in preference to several plausible alternatives. NIST AI RMF would have provided a broader AI risk management structure through its Govern, Map, Measure, and Manage functions, but it is less directly aligned with the dissertation’s need to test how AI-related integrity risks enter, propagate, and become visible within organisational cyber security artefacts [12]. STRIDE would have supported threat enumeration, but its categories are more closely associated with software threat modelling than with lifecycle governance, probabilistic model behaviour, and organisational risk ownership. Purely technical ML robustness experimentation would evaluate model susceptibility to adversarial inputs but would not assess organisational governance documentation or visibility [10]. Malware-focused exploit modelling would concentrate on code-level compromise rather than probabilistic degradation, behavioural drift, or lifecycle assurance gaps. CTI-driven adversary profiling would enrich threat assessment, but would not, by itself, evaluate structural framework adequacy [20].

Table 2 summarises the rationale for selecting DES as the primary modelling approach while positioning ISO/IEC 27005 and the MSc-derived lenses as complementary analytical supports.

Approach	Strength	Limitation for this dissertation	Role in this study
DES	Makes boundaries, trust assumptions, dependencies, and governance visibility explicit	Does not provide risk scoring or AI-specific controls by itself	Primary modelling and evaluation approach
ISO/IEC 27005	Provides asset–threat–vulnerability–risk–treatment sequence	May remain AI-implicit unless populated with lifecycle assets and behavioural risks	Core risk process structure
NIST AI RMF	Provides broad AI governance functions	Less directly tied to organisational cyber security evaluation	Contextual comparator

STRIDE	Supports threat enumeration	Less suited to lifecycle governance, probabilistic degradation, and organisational ownership	Not adopted as primary method
CML, MAL, CTI, SIM, QUC	Add technical and strategic interpretation	Supporting lenses rather than full methods	Used to refine specific stages

Table 2. Methodological Approach Selection and Justification

DES is therefore used as the primary modelling discipline, while ISO/IEC 27005 provides the procedural risk sequence and the MSc-derived lenses provide supporting analytical depth. Its limitation is that DES does not itself provide risk scoring, empirical validation, or AI-specific control requirements. These are instead supplied through ISO risk logic, AI governance standards, and the structured case analysis. This combination allows the dissertation to evaluate governance adequacy without treating DES as a complete governance framework.

3.3 Conceptual Foundations Informing the Method

3.3.1 ML System Characteristics

The CML course materials emphasise that ML systems operate through statistical inference rather than deterministic rule execution [10]. Model behaviour depends on training data distributions, feature representation, and generalisation capacity. This probabilistic foundation informs the methodological decision to treat AI models as integrity-critical assets whose behavioural reliability must be governed explicitly.

Integrity, in this context, extends beyond protection against unauthorised modification. It encompasses robustness against adversarial perturbation, resistance to data poisoning, and stability under distribution drift [17], [19].

3.3.2 Adversarial and Exploit Perspectives

Insights from MAL reinforce the importance of adversarial adaptation [16]. Although adversarial ML attacks differ technically from traditional malware, both exploit systemic weaknesses [19]. The CTI framework further informs likelihood estimation by incorporating attacker capability and intent into threat modelling [20].

These perspectives guide Stage 2 of the modelling framework (threat identification) and Stage 4 (risk assessment), ensuring that evaluation incorporates adversarial reasoning rather than purely theoretical vulnerability enumeration.

3.3.3 Incident Management and Detection Logic

SIM materials emphasise detection thresholds, containment procedures, and recovery processes [21]. Traditional incident logic presumes discrete compromise events. AI-related integrity degradation, however, may occur gradually through statistical manipulation or model drift.

This distinction informs the vulnerability and mitigation modelling stages. The methodology therefore incorporates evaluation of behavioural monitoring and threshold definition as governance controls.

3.3.4 Quantum Governance Context

The QUC materials highlight the importance of cryptographic agility and governance planning under technological disruption, including its threats [11], [22]. Although this dissertation does not model cryptographic transition operationally, QUC informs the broader evaluative principle: governance frameworks must remain adaptable under evolving technological conditions.

3.3.5 Methods and Tools from the MSc Programme Applied in this Dissertation

This dissertation draws explicitly on methods and analytical perspectives developed across the MSc programme and applies them within a governance evaluation context. The methods are not used with equal weight. DES and ISO/IEC 27005 form the central methodological structure because they organise the evaluation of governance adequacy: DES makes boundaries, assumptions, dependencies, and visibility gaps explicit, while ISO/IEC 27005 supplies the asset–threat–vulnerability–risk–treatment sequence. CML, MAL, CTI, SIM, and QUC operate as supporting lenses that refine specific analytical judgements rather than define the overall methodology.

DES provides the principal modelling scaffold by identifying system boundaries, actors, dependencies, trust assumptions, and control points. This allows the AI-enabled environment to be examined as a socio-technical system rather than as an isolated technical artefact.

ISO/IEC 27005 provides the procedural structure for asset identification, threat identification, vulnerability analysis, risk assessment, and control evaluation. This ensures that the case study remains traceable to recognised organisational risk management practice.

CML informs the treatment of probabilistic system behaviour by showing why AI integrity cannot be reduced to unauthorised modification alone. It supports the classification of drift, data-quality failure, threshold instability, and retraining effects as governance-relevant integrity risks.

MAL contributes adversarial reasoning by extending the analysis from control failure to intentional manipulation. It informs the treatment of poisoning, evasion, model manipulation, and other attempts to compromise system integrity.

CTI strengthens likelihood assessment by incorporating attacker capability, intent, opportunity, and evolving tactics. This prevents the risk assessment from relying only on static control presence or absence.

SIM informs the treatment of monitoring, escalation, containment, recovery, and post-incident learning. This is particularly relevant where AI integrity degradation emerges gradually through behavioural change rather than through a single discrete breach event.

QUC is incorporated at governance level rather than as a technical case domain. It frames cryptographic agility, dependency awareness, and long-range planning as part of organisational resilience under technological change.

Taken together, these programme-derived methods support a structured evaluation of AI-related integrity risk. Table 3 makes this application explicit by linking each method to its location in the dissertation and to an observable output in the case-study analysis.

MSc source	Concept or method applied	Where applied	Observable evidence of application
DES	Boundary, domain, trust-assumption and dependency modelling	Chapter 3 methodology; Chapter 4 Figure 4 and stages 1-5	AI integrity risk is traced across operational, lifecycle, monitoring, governance and cryptographic layers.
CML	Probabilistic model behaviour, training, validation, inference, drift and retraining	Sections 2.3, 3.3.1, 4.3-4.7	Models, datasets, feature pipelines, thresholds and drift controls are classified as integrity-critical assets and risks.
MAL	Adversarial adaptation and exploitation of systemic weakness	Sections 2.2.3, 3.3.2, 4.4	Evasion, poisoning, model extraction and feature manipulation are modelled as threat pathways.
CTI	Attacker capability, intent, resources and likelihood reasoning	Sections 2.5, 3.3.2, 4.4 and 4.6	Threat table includes attacker, motive, target, resources and risk-aversion; likelihood rationales are CTI-informed.
SIM	Detection, escalation, containment, recovery and post-incident learning	Sections 2.5, 3.3.3, 4.5-4.8	Controls include behavioural monitoring, drift thresholds, escalation triggers, rollback and incident classification.
QUC	Cryptographic agility and long-range security-assumption planning	Sections 2.4, 3.3.4, 4.3-4.8	PQC readiness, model signing, PKI, TLS, archived data and cryptographic agility planning are treated as strategic governance controls.

Table 3. MSc Programme Learning and Dissertation Application Matrix

3.4 Structured Risk Modelling Framework

The methodological core of this dissertation consists of a five-stage structured modelling framework aligned with ISO/IEC 27005 and operationalised through DES principles [5]. The DES framework is operationalised in conjunction with ISO/IEC 27005 risk assessment processes. This integration ensures that the modelling approach remains aligned with established organisational practices while extending them to address AI-specific risks.

Stage	Purpose	Analytical Focus	Output	Primary Analytical Basis
Stage 1	Asset Identification	Identify AI and governance assets	Asset register	DES + ISO/IEC 22989 + ISO/IEC 23053
Stage 2	Threat Identification	Model adversarial and systemic threats	Threat matrix	MAL + CTI + ISO/IEC 23894
Stage 3	Vulnerability Analysis	Identify governance and lifecycle gaps	Vulnerability table	DES + SIM + ISO/IEC 42001
Stage 4	Risk Assessment	Evaluate likelihood and impact	Risk matrix	ISO/IEC 27005 + ISO/IEC 23894
Stage 5	Control Evaluation	Map controls and mitigation strategies	Control framework	ISO/IEC 42001 + ISO/IEC 27001 + QUC

Table 4. Methodological Framework Overview

The five-stage framework in Table 4 operationalises the MSc-derived lenses introduced above by linking them to asset classification, threat modelling, vulnerability analysis, risk assessment, and control evaluation.

3.4.1 Stage 1: Asset Identification

AI-related components — including trained models, training datasets, feature engineering pipelines, monitoring systems, and deployment infrastructure — are explicitly identified and classified according to CIA.

This stage tests whether AI systems are formally recognised as integrity-critical assets within organisational governance structures.

3.4.2 Stage 2: Threat Identification

Threats are identified using a hybrid adversarial framework informed by adversarial ML research, MAL adversarial adaptation principles, and CTI capability modelling [20]. Threat categories include adversarial input manipulation, training data poisoning and feature pipeline abuse.

3.4.3 Stage 3: Vulnerability Identification

Vulnerabilities are conceptualised not solely as software defects but as governance omissions. Examples include absence of adversarial robustness testing, lack of dataset validation controls, and undefined behavioural drift thresholds. This broader description reflects DES system boundary modelling and SIM detection logic [21].

3.4.4 Stage 4: Risk Assessment

Risk is evaluated qualitatively using likelihood–impact matrices consistent with ISO/IEC 27005 [5]. Likelihood assessment incorporates CTI-informed adversary capability considerations, while impact assessment evaluates financial, operational, and reputational consequences.

This stage determines whether AI-related risks are represented explicitly and proportionately within the organisational risk register.

3.4.5 Stage 5: Mitigation Design and Control Mapping

Identified risks are mapped to governance controls using DES traceability principles. Mitigation strategies include:

- Explicit AI asset classification;
- Adversarial robustness testing integration;
- Dataset governance procedures;
- Behavioural monitoring thresholds; and
- Cryptographic agility planning.

This stage tests whether augmentation rather than structural replacement suffices to address identified governance gaps.

3.5 Case Study Strategy

A hypothetical organisational case study is employed to operationalise the modelling framework. The purpose of the case study is analytical rather than empirical. It represents how governance blind spots may emerge when deterministic assumptions remain implicit within AI-enabled operational environments.

The case study is designed to be realistic and structurally consistent with ISO-aligned organisational contexts. However, it does not claim empirical generalisability. Instead, it functions as a structured instrument for testing governance under defined modelling conditions. The case study is intentionally designed to support integrated evaluation of risk entry, propagation, and governance control visibility.

This approach aligns with the evaluative aim of the dissertation and ensures traceability between research question, modelling stages, and findings.

3.6 Validity, Reliability and Limitations

3.6.1 Validity

Construct validity is supported through alignment with established standards and scholarship, including ISO/IEC 27005, adversarial ML research, and DES modelling principles. Construct validity is further strengthened by aligning the conceptualisation of AI assets and lifecycle stages with AI-specific standards, particularly ISO/IEC 23053:2022 for ML-based AI system structure and ISO/IEC 23894 for AI risk management guidance, rather than relying only on conventional information-security risk standards. Internal validity is strengthened by ensuring that each modelling stage directly supports the evaluation of governance adequacy in relation to the research question.

3.6.2 Reliability

Reliability is supported through the structured and repeatable five-stage modelling framework. The analytical sequence is designed to be reproducible across other AI-enabled organisational settings because the same asset–threat–vulnerability–risk–control logic can be applied while preserving a consistent evaluation structure.

3.7 Chapter Summary

This chapter has established a structured qualitative methodology grounded in DES principles and aligned with ISO risk management standards. It has also made explicit how MSc-derived learning from CML, MAL, CTI, SIM, QUC, and DES is applied through the five-stage analytical framework rather than merely described.

The structured five-stage modelling framework provides a systematic mechanism for evaluating whether AI-related integrity risks are explicitly and adequately represented within organisational cyber security governance. The next chapter applies this methodology to a structured case study, generating the evidential basis for the analysis and discussion that follow.

Chapter 4: Case Study Implementation and Risk Analysis

4.1 Introduction

This chapter operationalises the structured modelling framework outlined in Chapter 3 through application to a hypothetical organisational case study. The purpose of this chapter is to evaluate whether existing organisational cyber security risk management frameworks adequately model and mitigate AI-related integrity risks within an AI-enabled operational environment.

This chapter also supports how the programme-derived methods are applied in the case study. DES structures the modelling process, while CML, MAL, CTI, SIM, and QUC inform the treatment of AI integrity, adversarial threats, monitoring, and cryptographic agility.

The case study functions as an analytical instrument rather than an empirical study. It represents how governance blind spots may emerge when deterministic assumptions remain implicit in AI-enabled environments. The objective is structured evaluation of framework adequacy rather than measurement of organisational prevalence.

Figure 4 is the central analytical artefact of this chapter. It connects the technical architecture of the AI-enabled fraud detection system with governance responsibilities, risk entry points, failure propagation paths, and control visibility.

4.2 Organisational Context

4.2.1 Overview of the Organisation

The organisation used in this case study is fictitious. It provides AI-driven fraud detection services to financial institutions. Its core service relies on supervised ML models deployed within client transaction pipelines.

It operates under an ISO/IEC 27001-aligned Information Security Management System. Risk assessments are conducted annually in accordance with ISO/IEC 27005. Existing risk registers document traditional cyber security risks, including network intrusion, malware infection, insider threat, and system misconfiguration.

For the purposes of the case study, the organisation is assumed not to have formalised AI-specific governance practices consistent with ISO/IEC 42001- and ISO/IEC 23894-style expectations. AI lifecycle responsibilities, model assurance activities, AI-specific risk ownership, and monitoring obligations have not been explicitly integrated into governance documentation or routine risk review. This is significant because ISO/IEC 42001 concerns the establishment, implementation, maintenance, and continual improvement of an Artificial Intelligence Management System, while ISO/IEC 23894

provides guidance for integrating AI risk management into organisational AI-related activities and functions.

However, prior to structured modelling, AI models were not explicitly classified as integrity-critical governance assets. They were implicitly treated as software components within the application environment.

It is important to clarify that this case study does not assert that all ISO-certified organisations omit AI classification. Rather, it provides a structured context for evaluating whether governance processes systematically and explicitly model AI-related integrity risks.

4.2.2 AI System Architecture

Deployment occurs within a secured cloud infrastructure environment protected by conventional network controls. This architecture reflects a plausible supervised ML operational design and provides a realistic context for structured governance modelling. The fraud detection pipeline consists of the following components:

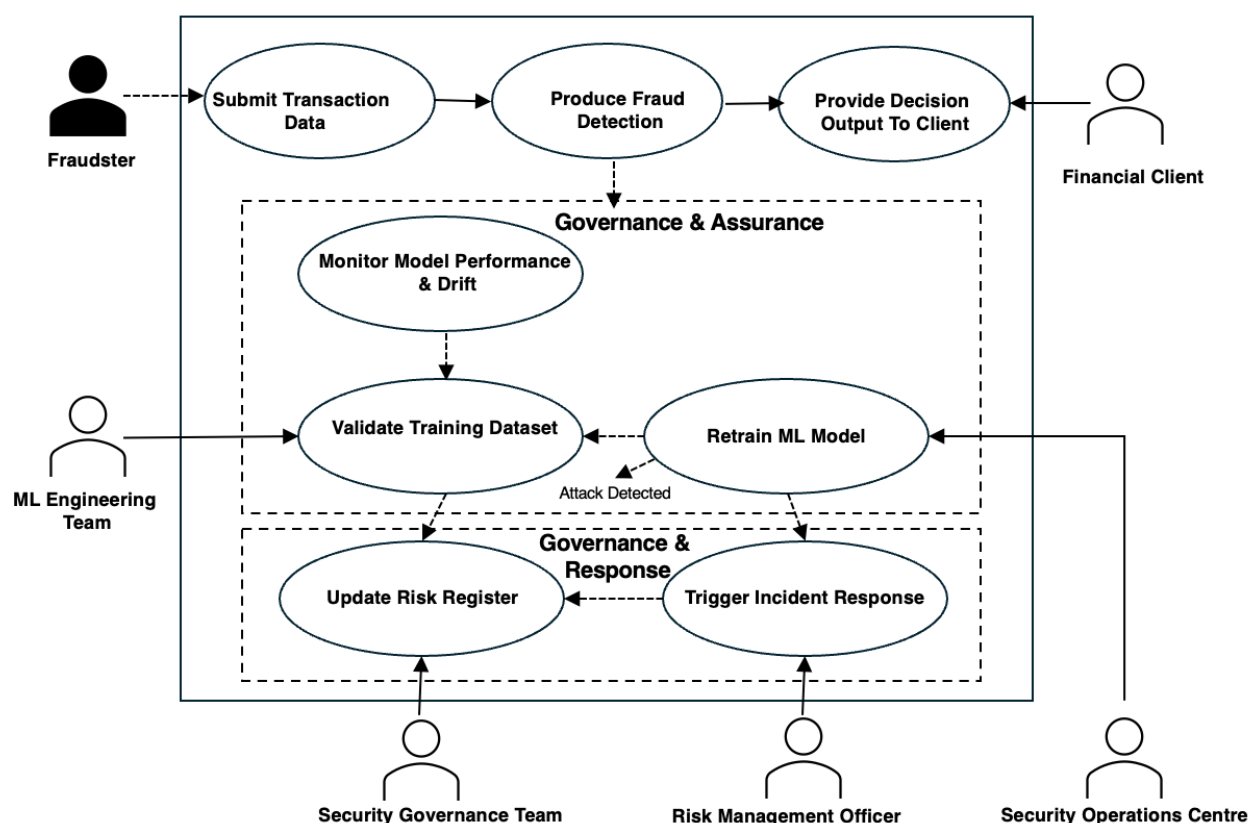
- Transaction ingestion interface;
- Feature engineering module;
- Trained ML classification model;
- Decision threshold logic; and
- Monitoring dashboard [23].

To ensure analytical traceability between governance processes and system behaviour, the operational environment is first modelled through a UML use case diagram. This diagram explicitly represents the interaction between external actors, AI system components, and governance functions. Modelling these relationships makes visible how integrity risk may propagate through legitimate interfaces rather than through unauthorised system compromise. In alignment with the DES approach outlined in Chapter 3, the use case representation clarifies system boundaries, trust assumptions, and responsibility allocation across operational and governance domains.

Figure 1 models how adversarial manipulation may enter the system through authorised transaction interfaces rather than through direct system intrusion. This modelling decision is analytically significant. It reinforces the central argument that AI-related integrity degradation may occur within legitimate operational flows, thereby challenging deterministic governance assumptions.

Furthermore, the explicit inclusion of governance actors — including the Security Governance Team, Risk Management Officer, and Security Operations Centre — illustrates that AI lifecycle activities such as dataset validation, robustness testing, and risk register updates must be treated as formal governance functions rather than informal technical tasks. This representation operationalises Stages 1 and 2 of the structured modelling frameworks by linking actors, assets, and threat vectors.

Figure 1. UML Use Case Diagram of Governance Actors and System Interactions

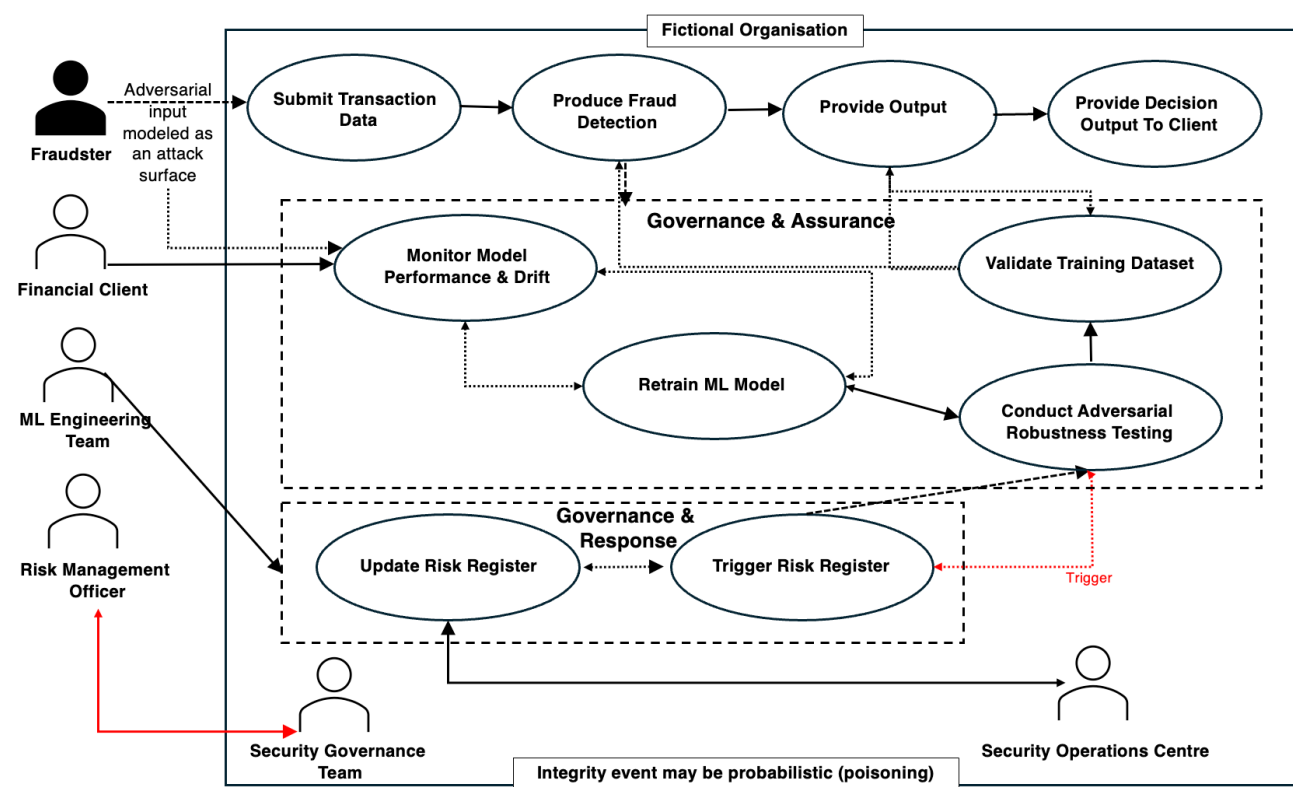


4.2.3 Governance Actors and Use Case Modelling

While the use case diagram models actor interactions, a structural representation of the technical environment is necessary to support asset identification and vulnerability analysis. The following system architecture diagram maps the end-to-end fraud detection pipeline, including ingestion, feature engineering, inference, monitoring, retraining, and governance integration components. This architectural view enables precise classification of integrity-critical assets and provides a concrete foundation for subsequent risk modelling stages.

Figure 2 makes explicit the separation between operational inference services and governance controls such as dataset validation, adversarial testing, monitoring, and risk register updates. Of particular importance is the inclusion of monitoring and drift detection mechanisms as integral system components rather than peripheral add-ons. This modelling reinforces the methodological position advanced in Chapter 3: AI integrity risk is not confined to code-level compromise but may manifest through behavioural deviation within legitimate system processes. The architectural representation therefore directly supports Stage 1 asset classification and provides structural clarity for the threat and vulnerability analysis undertaken in subsequent sections. Figures 1 and 2 provide the operational and structural basis for the integrated governance analysis developed later in the chapter.

Figure 2. System Architecture of the AI-Enabled Fraud Detection Environment



4.3 Stage 1: Asset Identification

Consistent with DES modelling principles, the first stage requires explicit identification and classification of integrity-critical assets. This stage operationalises the research question by testing whether AI assets are explicitly represented within governance structures.

4.3.1 Identified Assets

The modelling exercise identifies the trained ML model and training datasets as primary integrity-critical assets within the case study. Model outputs directly influence fraud detection decisions, while training data integrity affects decision-boundary behaviour [24], [25]. Prior governance documentation emphasised infrastructure protection, but did not explicitly classify the ML model, datasets, feature pipeline, monitoring thresholds, retraining workflow, or model registry as integrity-critical governance assets. This omission is significant because risks that are not formally classified as assets are less likely to be assigned owners, controls, monitoring requirements, or assurance evidence.

The asset model distinguishes AI conceptual assets, ML lifecycle assets, cyber infrastructure assets, and governance artefacts, reflecting AI-specific lifecycle framing as well as conventional ISMS logic.

ID	Asset	Domain	Confidentiality	Integrity	Availability	Integrity Priority
A01	ML Model [24]	AI/ML	Medium	Critical	High	Very High
A02	Model Parameters & Weights	AI/ML	Medium	Critical	Medium	Very High
A03	Training Dataset [26]	AI/ML	High	Critical	Medium	Very High
A04	Validation Dataset	AI/ML	Medium	High	Medium	High
A05	Feature Engineering Pipeline [27]	AI/ML	Medium	High	High	High
A06	Inference Service	AI/ML	Medium	High	High	High
A07	Drift Detection Mechanism	AI/ML	Low	High	High	High
A08	Retraining Workflow	AI/ML	Medium	High	Medium	High
A09	Decision Threshold Configuration	AI/ML	Low	High	High	High
A10	PKI Infrastructure	Cyber / Quantum	High	Critical	High	Very High
A11	TLS Certificates [22]	Cyber / Quantum	High	High	Medium	High
A12	Model Signing Mechanism	Cyber / Quantum	Medium	Critical	Medium	Very High
A13	Secure Model Distribution Channel	Cyber / Quantum	Medium	Critical	High	Very High
A14	API Authentication Mechanism	Cyber	High	High	High	High
A15	Security Logging System	Cyber	Medium	High	High	Medium
A16	Security Incident Management System	Governance	Medium	High	High	Medium
A17	Risk Register / ISMS Documentation	Governance	Low	High	Medium	High
A18	Cryptographic Agility Policy	Governance / Quantum	Low	High	Low	High
A19	Archived Transaction Data	Cyber / Quantum	Critical	High	Low	High

Table 5. Integrated Asset Register

As shown in Table 5, the assets requiring protection are not limited to infrastructure and applications. They include model behaviour, lifecycle data, thresholds, provenance mechanisms, cryptographic dependencies, and governance documentation. The key evaluative point is that the asset register changes the object of governance: AI-related integrity risk becomes visible only when these elements are treated as formal assets rather than background technical processes.

Within this case-study classification, integrity emerges as the dominant security property across AI/ML lifecycle and cryptographic infrastructure assets. This contrasts with conventional information-security contexts where confidentiality is often prioritised. In the AI-enabled fraud detection environment, behavioural manipulation may occur without data exfiltration, service disruption, or code alteration. Integrity must therefore be assessed across data quality, model behaviour, feature transformation, monitoring thresholds, retraining processes, model provenance, and cryptographic trust.

This classification anticipates Figure 4 by showing that AI integrity risk cannot be understood only at the inference layer. It must be traced across operational, lifecycle,

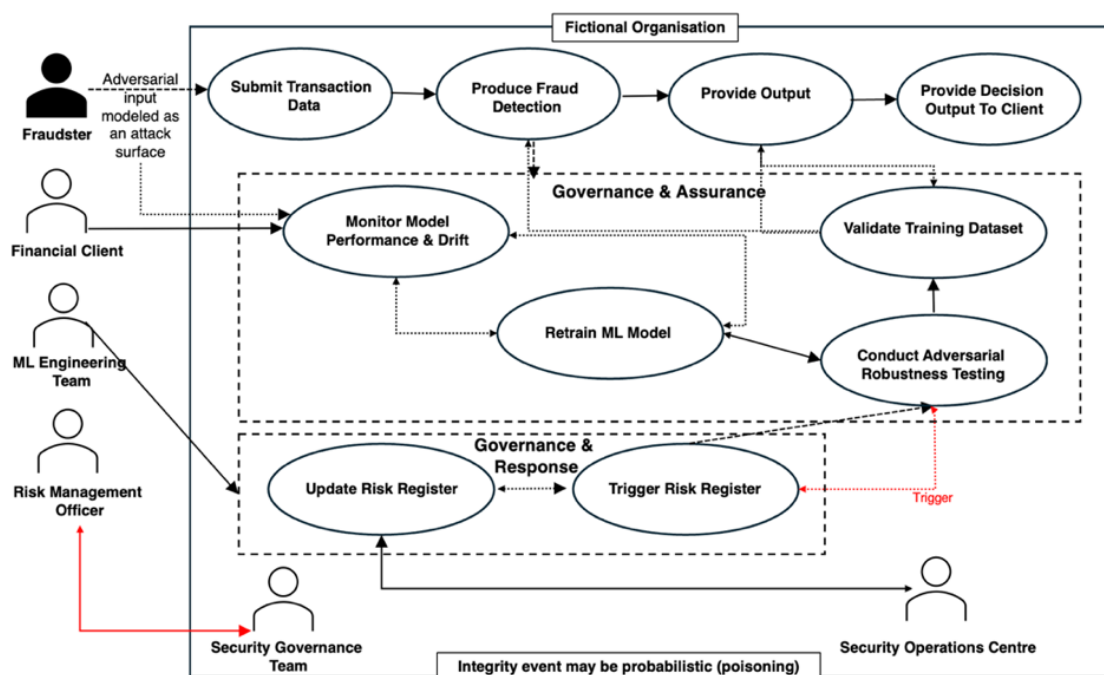
monitoring, cryptographic, and governance dependencies. Several cryptographic assets are also classified as strategically exposed to long-term post-quantum migration risk, reinforcing the need for cryptographic agility planning without turning quantum computing into a second operational case domain.

Asset	Primary Stakeholder	Secondary Stakeholder	Governance Oversight
ML Model	ML Engineering	DevOps	Security Governance
Training Dataset	ML Engineering	Data Team	Risk Officer
Feature Pipeline	ML Engineering	DevOps	Security Governance
Drift Detection System	SOC	ML Engineering	CISO
Inference Service	DevOps	ML Engineering	Security Governance
PKI Infrastructure	Crypto Team	DevOps	CISO
TLS Certificates	Crypto Team	DevOps	CISO
Model Signing Mechanism	Crypto Team	ML Engineering	Security Governance
Risk Register	Security Governance	Risk Officer	CISO
Cryptographic Agility Policy	CISO	Crypto Team	Board

Table 6. Stakeholder Responsibility Mapping

Asset identification was extended through the stakeholder mapping in Table 6 to establish governance traceability. AI lifecycle assets sit primarily with ML Engineering and DevOps, cryptographic infrastructure with Infrastructure Security, and governance artefacts with the Security Governance team and Risk Management Officer. Strategic accountability rests with the CISO. As Figure 3 shows, AI-related integrity risk is therefore distributed across operational, security, and executive domains. In this case study, failure to integrate AI assets into the ISMS is a governance alignment gap, not merely a technical limitation.

Figure 3. Stakeholder Interaction and Governance Responsibility Model



4.4 Stage 2: Threat Identification

Threat identification integrates adversarial ML, adversarial adaptation principles from MAL, and capability modelling from CTI frameworks [20]. Stage 2 operationalises the research question by translating the abstract concern of AI-related cyber security risk governance into specific, analysable threats, attackers, motives, and targets within the case context. This makes it possible to assess whether existing governance and risk management frameworks can adequately identify and represent those AI-related integrity risks in practice. The threats are mapped not only as abstract categories, but as entry paths and propagation routes within the analytical architecture later consolidated in Figure 4.

4.4.1 Adversarial Input Manipulation

Attackers craft transaction inputs designed to evade fraud detection by exploiting model decision boundary sensitivity [28]. Such perturbations do not modify system code and operate through legitimate input channels. Threat actor characteristics may include moderate technical sophistication, access to API outputs, and financial incentive for fraud evasion.

4.4.2 Training Data Poisoning

Adversaries introduce corrupted data into retraining datasets, thereby altering model behaviour gradually over time [29]. This aligns with adversarial data manipulation research and reflects adaptive threat dynamics described in MAL. Impact may include increased false negatives and reduced fraud detection effectiveness.

4.4.3 Model Extraction

Repeated querying of the model may permit reconstruction of approximate decision logic. Extracted surrogate models could facilitate targeted evasion strategies.

4.4.4 Feature Pipeline Abuse

Feature pipeline abuse occurs where an adversary manipulates or exploits the processes that transform raw transaction data into model-ready features. In the case-study system, the feature engineering pipeline is integrity-critical because changes to feature definitions, scaling, encoding, or derived attributes can alter model outputs even where the trained model itself remains unchanged.

This threat may arise through insider manipulation, weak change control, insufficient validation of feature transformations, or compromised development access. Unlike direct model tampering, feature pipeline abuse can appear as normal system operation because the model continues to receive inputs through authorised processes. The resulting integrity degradation may therefore be difficult to detect unless feature changes are explicitly governed, logged, validated, and linked to model-performance monitoring.

The significance of this threat is that it exposes a gap between software change control and AI lifecycle assurance. Conventional controls may protect the application environment, but they may not test whether changes to feature logic distort model behaviour. In this case study, feature pipeline abuse is therefore treated as both a technical threat and a governance visibility problem. It supports the wider argument that AI-related integrity risk must be assessed across the full ML lifecycle, not only at the model or infrastructure layer.

Threat ID	Threat	Goal	Attacker	Motivation	Target Asset	Required Resources	Risk-Aversion
T1	Adversarial Input Manipulation	Evade fraud detection	External fraudster	Financial gain	ML Model / Inference Service	Moderate ML knowledge, API access	Medium
T2	Training Data Poisoning	Corrupt model behaviour	Insider / Data contributor	Financial gain / sabotage	Training Dataset	Access to data pipeline, persistence	Low
T3	Model Extraction	Replicate model decision boundary	External ML-capable adversary	Competitive advantage / targeted evasion	ML Model Parameters	High-volume query access, ML expertise	Medium
T4	Feature Pipeline Manipulation	Distort feature representation	Insider / Developer	Sabotage / negligence	Feature Engineering Pipeline	Development access	Low
T5	Undetected Model Drift Exploitation	Exploit degraded detection	Fraudster / coordinated group	Fraud scalability	Drift Detection Mechanism	Pattern testing, repeated interaction	Medium
T6	Model Substitution via Signature Forgery	Introduce malicious model	Advanced adversary	System compromise	Model Signing Mechanism	Cryptographic exploitation capability	Low
T7	API Authentication Compromise	Intercept or manipulate transactions	External attacker	Data theft / disruption	API Authentication Mechanism	Cryptographic attack capability	Medium
T8	Quantum Decryption (Harvest-Now, Decrypt-Later) [30]	Decrypt stored encrypted data	Future quantum-capable actor	Strategic intelligence	Archived Transaction Data	Large-scale quantum computer	High
T9	PKI Compromise via Shor's Algorithm	Break digital signature trust	Nation-state actor	Systemic disruption	PKI Infrastructure	Quantum computational resources	High
T10	Logging Suppression / Evasion	Avoid detection	Insider / adversary	Conceal fraud	Security Logging System	Privileged access	Low
T11	Governance Omission	Maintain AI asset invisibility	Organisational oversight failure	Expediency / underestimation	Risk Register / ISMS	None (process weakness)	High

Table 7. Threat Identification and Actor Profiling

Table 7 presents threats that reflect CTI-informed adversarial capability and show that AI integrity risk can be exploited through legitimate system interfaces and lifecycle processes. The table therefore changes threat analysis from a generic list of attack types into a governance test: if these threat pathways are absent from the risk register, the organisation's apparent ISMS maturity does not translate into AI integrity visibility.

The threat model suggests that, in this case study, AI-related integrity risk is not limited to highly sophisticated adversaries. Under the assumptions of the case study, several high-impact threats, such as adversarial input manipulation and data poisoning, are assessed as requiring moderate technical capability and may be executed through legitimate system interfaces [28], [29]. In contrast, quantum-enabled threats (e.g., PKI compromise) are currently high-resource and high-risk-averse yet represent strategic long-term exposure. Governance omission is included as a threat-enabling condition rather than a conventional adversarial threat, reflecting the DES principle that system boundaries include organisational processes as well as technical artefacts.

4.5 Stage 3: Vulnerability Identification

Vulnerabilities are conceptualised not only as technical defects, but also as governance omissions, lifecycle gaps, and control weaknesses. Stage 3 operationalises the research question by identifying the specific weaknesses through which AI-related integrity risks can materialise within organisational systems and processes. This shows whether existing risk governance approaches are sufficiently detailed to capture and evaluate those vulnerabilities in a structured way.

Identified vulnerabilities include:

- Absence of adversarial robustness testing prior to deployment;
- No explicit ML lifecycle governance documentation within the ISMS;
- Undefined behavioural drift detection thresholds; and
- Monitoring focused on system uptime rather than output reliability.

These vulnerabilities reflect deterministic assumptions embedded within governance processes. SIM frameworks emphasise discrete incident detection, whereas AI integrity degradation may be gradual and statistical [21].

Vulnerability ID	Vulnerability	Asset(s) Affected	Impact	Associated Threat(s)
V1	Absence of adversarial robustness testing	ML Model, Inference Service	Increased evasion success; false negatives	T1
V2	Lack of dataset validation controls during retraining	Training Dataset, Retraining Workflow	Model corruption; long-term behavioural distortion	T2
V3	Excessive model query exposure via API	ML Model, API Authentication Mechanism	Model extraction; decision boundary leakage	T3
V4	Insufficient access control over feature engineering pipeline	Feature Engineering Pipeline	Feature distortion; degraded model performance	T4

V5	No formal statistical drift detection thresholds	Drift Detection Mechanism, Monitoring System	Undetected degradation; fraud escalation	T5
V6	Classical cryptographic signatures without PQC readiness	Model Signing Mechanism, PKI Infrastructure	Model substitution risk; forged signatures	T6, T9
V7	Dependence on RSA/ECC-based TLS without cryptographic agility	TLS Certificates, Archived Transaction Data	Future decryption of stored sensitive data	T8
V8	Inadequate API rate limiting and anomaly detection	API Authentication Mechanism	Increased likelihood of model extraction and evasion	T1, T3
V9	Logging and monitoring privilege escalation	Security Logging System	Suppressed detection capability	T10
V10	AI assets not explicitly classified as integrity-critical within ISMS	ML Model, Training Dataset, Risk Register	Underrepresentation of AI risk in governance processes	T11
V11	Absence of formal cryptographic agility migration plan	PKI Infrastructure, Cryptographic Agility Policy	Strategic long-term systemic exposure	T8, T9
V12	Over reliance on accuracy metrics without adversarial validation	Validation Dataset, Performance Metrics Repository	False governance assurance; inflated performance perception	T1, T2

Table 8. Vulnerability Identification

As detailed in Table 8, the most significant vulnerabilities are not merely technical weaknesses, but omissions in assurance, ownership, monitoring, and lifecycle documentation. Many AI-related exposures arise not from software defects, but from governance omissions and lifecycle under-specification. V1, V2, and V5 illustrate how probabilistic integrity degradation may occur without unauthorised modification of system code. V6, V7, and V11 show that cryptographic dependencies introduce long-term quantum exposure risks that extend beyond immediate operational compromise. V10 highlights that failure to classify AI models and datasets as integrity-critical assets constitutes a structural governance vulnerability rather than a technical weakness.

The vulnerabilities identified in Table 8 fall into four analytical areas: operational components, lifecycle processes, monitoring mechanisms, and governance artefacts. Operational vulnerabilities affect the ingestion interface, feature engineering pipeline, inference service, and API exposure. Lifecycle vulnerabilities arise in training data, validation data, retraining workflows, and model version control. Monitoring vulnerabilities concern the absence of adversarial testing, undefined drift thresholds, incomplete logging, and weak escalation triggers. Governance vulnerabilities arise where AI assets are not explicitly classified, risk ownership is unclear, or model assurance is not embedded within the ISMS. This distribution is what Figure 4 is designed to make visible: AI-related integrity risk does not sit in a single technical component, but propagates across lifecycle, monitoring, and governance layers.

4.6 Stage 4: Risk Assessment

Risk assessment follows ISO/IEC 27005 qualitative likelihood–impact logic. Table 9 provides the risk lookup used to classify likelihood, impact, and severity in the case study. The ratings are qualitative and case-specific rather than empirical probabilities. Likelihood is assessed by considering attacker capability, access requirements, exposure, and lifecycle opportunity. Impact is assessed by considering financial loss, operational disruption, client trust, regulatory exposure, and strategic dependency.

Risk	Likelihood	Impact	Severity
Adversarial evasion	Medium	High	High
Training data poisoning	Low–Medium	High	High
Model extraction	Medium	Medium	Medium
Undetected model drift	Medium	Medium	Medium

Table 9. Risk Lookup

Likelihood estimation incorporates CTI-informed adversary capability considerations [20]. Impact assessment considers financial loss, reputational damage, operational disruption, and erosion of client trust.

Under the case-study assumptions, Table 10 indicates that adversarial evasion, data poisoning, undetected drift, and governance omission represent high-severity integrity risks that would be under-represented if the organisation relied only on its pre-existing conventional risk register.

Risk ID	Risk Description	Associated Threats	Associated Vulnerabilities	Likelihood (with rationale)	Impact	Risk Severity
R1	Fraud evasion via adversarial inputs	T1	V1, V8	Medium — Moderate technical skill required; accessible via legitimate API	High — Financial loss, reputational damage	High
R2	Long-term model corruption through data poisoning	T2	V2	Low–Medium — Requires sustained access to training data	High — Systemic behavioural degradation	High
R3	Model replication and targeted evasion	T3	V3, V8	Medium — High query access feasible without rate limiting	Medium — Competitive intelligence, fraud optimisation	Medium–High
R4	Feature manipulation leading to misclassification	T4	V4	Low — Requires insider access	Medium — Operational degradation	Medium

R5	Undetected drift enabling fraud scaling	T5	V5	Medium — Natural data drift probable over time	Medium–High — Reduced fraud detection accuracy	High
R6	Malicious model substitution via signature compromise	T6, T9	V6	Low (current quantum capability limited)	Very High — Full system compromise	High (strategic)
R7	Future decryption of archived sensitive data	T8	V7	Low (current) / High (long-term) — Dependent on quantum maturity	High — Regulatory breach, confidentiality loss	High (long-term)
R8	Logging suppression to conceal fraud	T10	V9	Low–Medium — Requires privileged access	Medium — Reduced detection capability	Medium
R9	Governance blind spot for AI integrity risks	T11	V10	Medium — AI often under-classified in ISMS	High — Systematic underestimation of risk	High
R10	Absence of post-quantum migration plan	T8, T9	V11	Low (immediate) / High (strategic)	Very High — Cryptographic trust collapse	Very High (strategic)
R11	False governance assurance from performance metrics	T1, T2	V12	Medium — Accuracy metrics commonly misinterpreted	Medium — Delayed detection of integrity degradation	Medium–High

Table 10. Risk Assessment

The risk assessment reveals that immediate operational exposure is concentrated in AI integrity degradation risks, particularly R1, R2, and R5, whereas quantum-related risks, including R6, R7, and R10, represent lower immediate likelihood but high strategic impact. This prioritisation is significant because risk severity is driven not only by technical exploitability, but also by delayed detectability and governance invisibility. AI-related probabilistic integrity risks therefore require immediate governance integration, while quantum cryptographic exposure requires forward-looking cryptographic agility planning. Governance omission, represented by R9, is assessed as a high-severity systemic risk because it can prevent AI integrity risks from being represented, owned, monitored, or reviewed.

This pattern supports the dissertation's use of deterministic governance bias as an explanatory concept. Adversarial evasion, data poisoning, model drift, and false assurance from performance metrics may operate through authorised interfaces or routine lifecycle activities, meaning that conventional controls may not generate clear incident signals. Where governance remains oriented towards observable intrusion, malware execution, or unauthorised modification, probabilistic degradation may be under-prioritised because it is less visible, slower to escalate, and harder to classify as a discrete incident. This is consistent with AI-specific risk management expectations in ISO/IEC 23894 and ISO/IEC 42001, which emphasise lifecycle risk

identification, treatment, monitoring, and integration into organisational AI-related activities.

4.7 Stage 5: Control Evaluation

Stage 5 evaluates whether controls are correctly positioned across the analytical architecture of the case-study system. It does not ask only whether controls exist in a general organisational sense. This distinction matters because the organisation's existing controls are concentrated around perimeter protection, access management, infrastructure security, secure coding, malware protection, and periodic vulnerability scanning. These controls remain necessary for conventional cyber security, but they do not by themselves provide sufficient visibility over AI-related integrity risks that emerge across the model lifecycle, monitoring environment, and governance layer.

In the case-study context, effective AI integrity governance requires controls distributed across lifecycle processes, behavioural monitoring, assurance review, risk ownership, and strategic cryptographic planning. This control positioning is consistent with ISO/IEC 42001's emphasis on AI management-system governance and ISO/IEC 23894's treatment of AI risk management as an organisational lifecycle activity.

Existing governance controls include:

- network intrusion detection systems;
- endpoint malware protection;
- role-based access control;
- secure coding standards; and
- periodic vulnerability scanning.

While structurally robust, these controls primarily address exploit-based threats. They do not provide sufficient visibility over AI-related integrity risks that arise through training data, feature pipelines, model behaviour, drift, retraining, provenance, or governance ownership. The modelling therefore indicates that the framework remains AI-implicit rather than AI-explicit.

Table 11 sets out the proposed risk treatment strategy, linking each risk to preventative, detective, reactive, and deterrent controls within the governance framework.

Risk ID	Risk Description	Prevent	Detect	React	Deter	Rationale
R1	Fraud evasion via adversarial inputs	Integrate adversarial robustness testing; implement input validation controls	Behavioural anomaly detection; model confidence monitoring	Temporarily tighten decision thresholds; retrain model	Rate limiting and abuse monitoring	Reduces exploitability of decision boundary while maintaining operational continuity
R2	Long-term model corruption	Formal dataset validation	Statistical anomaly	Roll back to previous	Strict contributor authentication	Protects lifecycle integrity and

	through data poisoning	audits; access control on data ingestion	detection in training data	validated model version	and audit logging	prevents long-term degradation
R3	Model replication and targeted evasion	API query rate limiting; response obfuscation	Monitoring of query patterns and high-volume access	API key revocation; rotate authentication credentials	Strict API authentication and monitoring	Limits adversarial learning of decision boundaries
R4	Feature manipulation leading to misclassification	Segregation of duties; strict development access controls	Change monitoring and configuration integrity checks	Revert to validated pipeline configuration	Least-privilege enforcement	Reduces insider-based manipulation risk
R5	Undetected drift enabling fraud scaling	Formal statistical drift thresholds; periodic model evaluation	Continuous monitoring dashboards; performance degradation alerts	Trigger retraining workflow; incident classification	Governance-level performance reporting	Addresses probabilistic integrity degradation
R6	Malicious model substitution via signature compromise	Post-quantum signature readiness planning; secure model registry	Signature verification checks at deployment	Revoke compromised certificates; restore validated model	Code-signing discipline and audit	Ensures authenticity of deployed models
R7	Future decryption of archived sensitive data	Cryptographic agility roadmap; data minimisation	Periodic cryptographic inventory review	Accelerate migration to post-quantum cryptography	Encryption lifecycle governance	Mitigates long-term confidentiality exposure
R8	Logging suppression to conceal fraud	Segregated logging infrastructure; tamper-evident logs	Log integrity monitoring and anomaly alerts	Forensic investigation procedures	Strict privileged access review	Protects detection capability
R9	Governance blind spot for AI integrity risks	Explicit AI asset classification in ISMS; formal ML lifecycle documentation	Governance audits; internal compliance review	Escalate to CISO; update risk register	Board-level reporting of AI risk	Addresses structural under-modelling
R10	Absence of post-quantum migration plan	Develop post-quantum transition strategy	Cryptographic risk tracking	Controlled migration project execution	Executive oversight	Reduces strategic cryptographic collapse risk
R11	False governance assurance from performance metrics	Integrate adversarial evaluation metrics; broaden validation criteria	Periodic robustness testing	Retraining with adversarial examples	Governance review of ML assurance practices	Prevents misleading accuracy-based confidence

Table 11. Risk Treatment Strategy

These risk treatments show that AI-related integrity risks require lifecycle-integrated controls rather than perimeter-focused defences. Preventative measures focus on robustness testing, dataset validation, access control, and cryptographic readiness. Detective measures shift from binary intrusion logic to behavioural and statistical monitoring. Reactive controls emphasise rollback, retraining, credential revocation, and incident classification. Deterrent controls reinforce governance discipline through audit logging, executive oversight, and formal reporting. Quantum-related risks, while lower in immediate likelihood, require proactive cryptographic agility planning to mitigate long-term systemic exposure.

As shown in Table 12, the selected controls are assessed as reducing most risks to lower residual levels within the qualitative scoring model.

Risk ID	Risk Severity	Control (Prevent / Detect / React / Deter with Rationale)	Cost (L/M/H)	Benefit (L/M/H)	Chosen (Y/N)	Residual Risk	Justification
R1	High	Prevent: Adversarial robustness testing prior to deployment — reduces exploitability of decision boundary	M	H	Y	Medium	Directly reduces fraud evasion likelihood; cost justified by financial exposure
R2	High	Prevent: Dataset validation and controlled data ingestion workflow	M	H	Y	Low–Medium	Protects lifecycle integrity; essential governance control
R3	Medium–High	Prevent: API rate limiting and query anomaly detection	L	H	Y	Medium	Low cost, high mitigation impact against extraction
R4	Medium	Prevent: Segregation of duties in feature engineering	L	M	Y	Low	Governance control; inexpensive organisational adjustment
R5	High	Detect: Formal drift detection thresholds with automated alerts	M	H	Y	Medium	Addresses probabilistic integrity degradation
R6	High (Strategic)	Prevent: Post-quantum signature readiness assessment	M	M	Y	Medium–High	Strategic mitigation; reduces long-term substitution risk
R7	High (Long-Term)	Prevent: Cryptographic agility roadmap aligned to PQC standards	M	H	Y	Medium	Reduces future decryption exposure
R8	Medium	Detect: Tamper-evident logging with integrity verification	M	M	Y	Low	Protects detection capability

R9	High	Prevent: Explicit AI asset classification within ISMS	L	H	Y	Low	Governance enhancement with minimal cost
R10	Very High (Strategic)	Prevent: Executive-approved PQC migration programme	H	H	Y	Medium	High cost but essential for long-term resilience
R11	Medium–High	Detect: Adversarial validation integrated into model assurance	M	H	Y	Medium	Prevents false performance confidence

Table 12. Security Controls and Residual Risks

Figure 4 synthesises the preceding control evaluation by locating AI-related integrity risks across operational, lifecycle, monitoring, governance, and strategic cryptographic layers. Its purpose is to show how risk entry points, propagation paths, control visibility, and governance artefacts interact within the case-study system.

The figure also links the control analysis to wider governance concerns, including asset classification, visibility, assurance review, incident escalation, provenance, and cryptographic agility.

Figure 4. Analytical Architecture of AI Integrity Governance: Risk Entry Points, Propagation Paths and Control Visibility

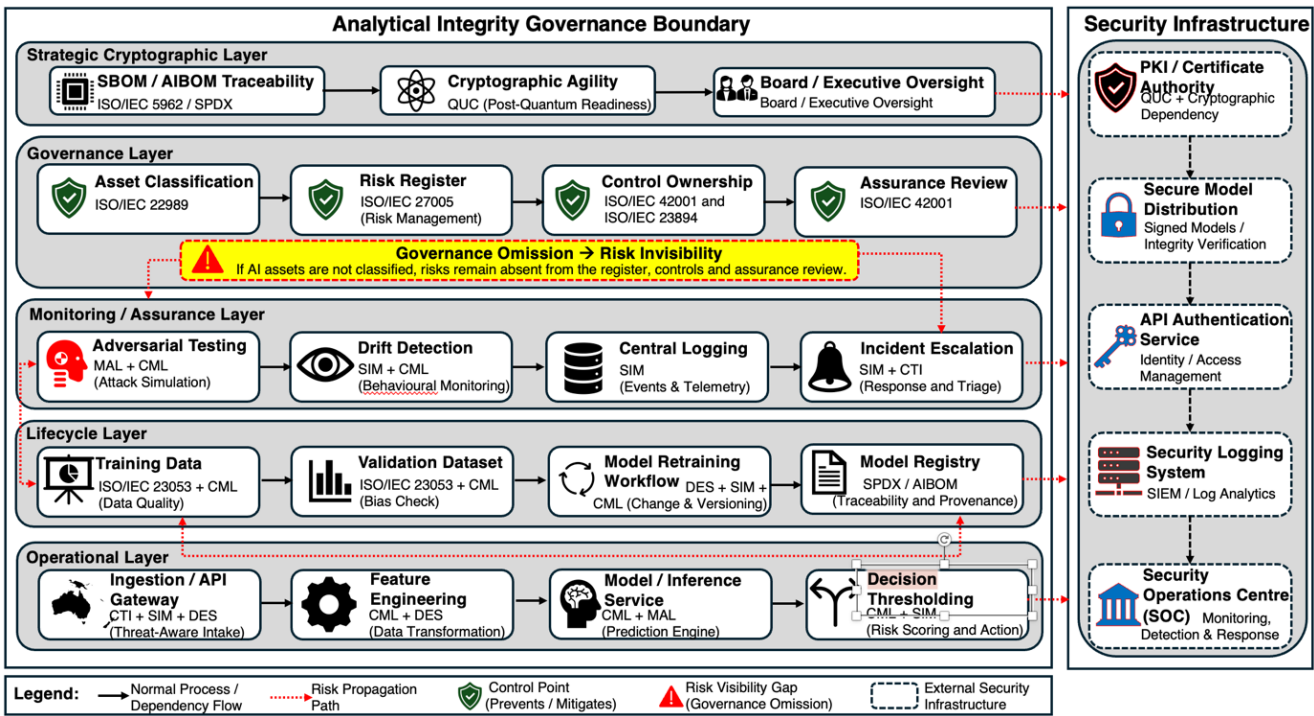


Table 13 summarises how the analytical model is grounded in recognised governance and technical standards rather than in an abstract conceptual framework alone. It shows that governance adequacy depends on integrating terminology, lifecycle structure, risk management, organisational accountability, and traceability across the full AI system environment rather than treating these elements in isolation.

Standards	Relevance	Location
ISO/IEC 22989	Defines AI concepts and terminology	AI assets, lifecycle labels
ISO/IEC 23053	Frames ML-based AI system structure	Training, inference, feedback paths
ISO/IEC 42001	Management system lens for AI governance	Governance layer, accountability, review
ISO/IEC 23894	AI risk management lens	Risk identification, lifecycle controls
ISO/IEC 27001	ISMS context	Asset register, ownership, policy
ISO/IEC 27005	Risk assessment logic	Threat, vulnerability, likelihood, treatment
ISO/IEC 5962 (SPDX)	Provenance / dependency / bill-of-materials traceability	Model registry, signed artefacts, dependency visibility

Table 13. Standards and Analytical Relevance to Figure 4

Taken together, these standards do not function as parallel frameworks within the dissertation. They operate as layered reference points that clarify terminology, lifecycle structure, governance expectations, risk treatment logic, and traceability requirements within the case analysis.

4.7.1 Analytical Reading

Figure 4 is not a descriptive system diagram. It is the central analytical model used to evaluate governance adequacy. It shows where AI-related integrity risk enters the system, how it propagates across technical and organisational boundaries, and where governance mechanisms either make that risk visible or allow it to remain under-modelled.

Its value is that it reveals relationships a conventional risk register would not show. A risk register may list adversarial evasion, data poisoning, model drift, or model extraction as separate risks, but it does not show how they move across authorised input channels, lifecycle processes, monitoring thresholds, governance artefacts, and executive accountability. Without the layered model, adversarial input manipulation may appear to be an application issue, data poisoning a data-quality issue, drift a performance issue, and model signing a cryptographic issue. Figure 4 shows that these are connected governance problems because each depends on asset classification, ownership, monitoring, assurance evidence, and escalation.

First, the figure shows that key integrity exposures arise through authorised operational pathways, not only through perimeter breach. Adversarial input manipulation enters through legitimate transaction ingestion; poisoning enters through retraining processes; model extraction exploits permitted query interfaces; and drift emerges gradually through normal operation. This supports the argument that AI integrity risk is behavioural and lifecycle-dependent, not limited to unauthorised structural modification.

Second, the figure shows that governance adequacy depends on whether technical components have corresponding governance artefacts. The ML model, training dataset, feature pipeline, retraining workflow, monitoring thresholds, and model-

signing mechanism are governance-relevant assets, not merely engineering components. Their omission from asset registers and risk registers produces under-modelling. This is where DES contributes most directly by making boundaries, dependencies, and trust assumptions explicit.

Third, Figure 4 integrates the MSc-derived lenses. CML frames the model and data pipeline as probabilistic assets; MAL identifies adversarial pathways; CTI supports capability-sensitive threat interpretation; SIM informs drift detection, escalation, rollback, and post-incident learning; and QUC positions model signing, PKI dependency, and cryptographic agility as strategic governance concerns. The figure therefore operates as the dissertation's synthetic model of application.

Fourth, the figure links the case evidence to AI-specific standards. ISO/IEC 22989 and ISO/IEC 23053 help define the AI system and ML lifecycle components. ISO/IEC 42001 and ISO/IEC 23894 support explicit lifecycle governance and risk management. ISO/IEC 5962 supports provenance, dependency, and traceability reasoning. Figure 4 is therefore the point where standards, methods, and case-study evidence are brought together.

The central interpretation is that governance weakness in the case study does not arise from the absence of organisational cyber security frameworks, but from incomplete visibility within them. Figure 4 makes that incompleteness visible by converting separate technical risks into a single governance pattern across operational, lifecycle, monitoring, governance, and strategic layers.

4.8 Mitigation Design and Governance Augmentation

The purpose of this section is to translate the risks identified in Sections 4.3–4.7 into concrete governance and control measures. The focus here is practical rather than evaluative: it specifies how identified weaknesses can be addressed through lifecycle controls, behavioural monitoring, governance documentation, and strategic cryptographic planning within an existing organisational framework.

4.8.1 Lifecycle Controls

ML models, training datasets, validation datasets, feature engineering pipelines, retraining workflows, and model registries should be formally treated as integrity-critical lifecycle assets. Mitigations include dataset validation, controlled retraining, model version approval, secure model registry management, and explicit documentation of model provenance. These controls address poisoning, feature manipulation, and model-substitution risks by ensuring that lifecycle changes are governed rather than treated as routine engineering activity.

4.8.2 Behavioural Monitoring Controls

AI integrity cannot be assured solely through access control or infrastructure monitoring. The organisation should therefore implement adversarial robustness

testing, drift detection thresholds, model-confidence monitoring, API query-pattern monitoring, and escalation triggers for behavioural anomalies. These controls address adversarial evasion, model extraction, gradual performance degradation, and delayed detection by making probabilistic integrity failure visible to security and governance teams.

4.8.3 Governance Documentation Controls

The ISMS should be augmented so that AI-specific assets, risks, controls, owners, review cycles, and assurance evidence are explicitly recorded. This includes AI asset classification, AI-specific entries, assigned control ownership, assurance review records, and board or CISO-level reporting where risks exceed defined thresholds. These controls address governance omission by ensuring that AI-related integrity risk is visible, traceable, and accountable within organisational risk management.

4.8.4 Strategic Cryptographic Controls

Cryptographic dependencies supporting model signing, secure model distribution, TLS, PKI, and archived transaction data should be governed through cryptographic agility planning. This includes maintaining a cryptographic inventory, monitoring post-quantum transition guidance, assessing signature and key-management dependencies, and planning migration pathways for high-value assets. These controls address long-term substitution, trust-collapse, and harvest-now-decrypt-later risks without turning quantum computing into a second operational case domain.

These mitigation strategies also show how the Chapter 3 analytical toolkit can be translated into governance artefacts within an ISMS environment. DES informs asset classification and trust boundaries; CML and MAL inform robustness and validation; CTI informs prioritisation; SIM informs monitoring and escalation; and QUC informs cryptographic agility planning.

Quantum-related controls are retained only to the extent that they represent present governance planning duties. In the case study, the absence of a cryptographic inventory and migration roadmap is treated as a strategic governance gap, not as the central operational risk analysed in this dissertation. Post-quantum readiness is therefore retained here as a strategic planning requirement rather than as a second operational evaluation domain.

4.9 Governance Evaluation Outcome

This section evaluates what the case-study results mean for the adequacy of existing organisational cyber security governance frameworks. The framework examined is procedurally robust: it includes an ISO/IEC 27001-aligned ISMS, recurring risk assessment, documented controls, and a conventional asset–threat–vulnerability–risk–treatment structure.

However, the case study also demonstrates operational incompleteness. AI-related integrity risks do not become visible automatically within the existing governance

process. Without explicit modelling and ownership, risks such as adversarial evasion, training-data poisoning, drift, model extraction, and governance omission remain under-represented in the risk register and associated control structure.

Figure 4 explains why this incompleteness occurs. It shows that AI-related integrity risk is distributed across operational, lifecycle, monitoring, governance, and strategic cryptographic layers. The central Chapter 4 result is therefore that governance inadequacy arises less from the absence of framework structure than from insufficient visibility of AI-related integrity risk within that structure.

Within the limits of the case study, this outcome supports the dissertation's wider argument that existing organisational cyber security governance frameworks are conditionally adequate. They remain usable, but their adequacy depends on whether AI-related integrity risks are explicitly represented, assigned, monitored, and reviewed within governance processes.

4.10 Chapter Conclusion and Transition

This chapter has applied the five-stage modelling framework to the AI-enabled fraud detection case study and has shown how AI-related integrity risks become visible only when technical, lifecycle, monitoring, and governance elements are analysed together. Through asset identification, threat modelling, vulnerability analysis, risk assessment, and control evaluation, the chapter supported that risks such as adversarial evasion, training data poisoning, model extraction, drift, and governance omission cannot be adequately understood at the operational inference layer alone.

Figure 4 synthesised this analysis by showing how risk entry points, propagation paths, control points, governance artefacts, and accountability relationships interact across the operational, lifecycle, monitoring/assurance, governance, and strategic cryptographic layers. This supports the chapter's central finding that existing cyber security governance frameworks remain procedurally useful but become operationally incomplete where AI-specific lifecycle risks, behavioural monitoring requirements, and governance responsibilities are not made explicit. This is consistent with ISO/IEC 42001's management-system emphasis on organisational AI governance and ISO/IEC 23894's guidance on integrating AI risk management into organisational AI-related activities and functions.

The next chapter interprets these results to evaluate whether the identified gaps represent framework failure, framework incompleteness, or the effects of deterministic governance assumptions carried into AI-enabled operational contexts.

Chapter 5: Discussion and Conclusion

5.1 Introduction

This chapter discusses and concludes the dissertation. It interprets the Chapter 4 findings in relation to the research question, evaluates the adequacy of existing organisational cyber security risk management frameworks, and considers the wider conceptual and practical implications of the analysis. It then states the dissertation's contribution, situates quantum computing as a strategic governance context, identifies limitations and future research, provides a final response to the research question, reflects on the research process, and closes with a concluding statement.

The discussion is anchored in the case-based evaluation developed in Chapter 4, particularly Figure 4, which connected technical architecture, risk entry points, lifecycle dependencies, governance artefacts, and control visibility. On that basis, this chapter argues that existing organisational cyber security governance frameworks remain usable and coherent, but do not fully address AI-related integrity risk unless those risks are made explicit across asset classification, lifecycle control design, monitoring, assurance, and governance ownership.

5.2 Reconceptualising Integrity in AI-Enabled Systems

Traditional interpretations of integrity within ISO-aligned cyber security governance frameworks focus on protecting data and systems from unauthorised modification, corruption, or tampering. Within this view, compromise is generally understood as a discrete and observable event, such as database alteration, malware injection, unauthorised access, or configuration change. Governance practices built on this logic therefore prioritise access control, configuration management, logging, and intrusion detection.

This interpretation remains appropriate for many conventional information systems, particularly where system behaviour is largely deterministic and failures can be traced to identifiable technical or procedural violations. In such settings, integrity degradation is commonly understood as structural interference with system state.

The findings in Chapter 4 indicate that this interpretation is insufficient for the AI-enabled fraud detection system examined in the case study. In the case study, integrity degradation may arise through adversarial input manipulation, training data poisoning, model extraction, feature pipeline abuse, and model drift, none of which necessarily requires unauthorised code modification or direct system intrusion [9], [18], [19], [17]. In these cases, the system may continue to operate within authorised boundaries while still producing degraded, manipulated, or unreliable outputs.

For the purposes of AI-enabled governance, integrity should therefore be understood more broadly as behavioural reliability under probabilistic conditions. This includes robustness against adversarial manipulation, resilience to corrupted or shifting data,

and stability of model behaviour across lifecycle stages. The question is no longer only whether a system has been altered, but whether it continues to behave reliably under statistical and adversarial pressure.

This broader concept of integrity is consistent with AI-specific governance expectations, particularly those emphasising lifecycle visibility, accountability, monitoring, and continual review [1], [2], [3], [4]. It also explains why AI-related integrity risk cannot be adequately governed if AI models, datasets, feature pipelines, retraining processes, and monitoring thresholds remain implicit within organisational cyber security risk management.

5.3 Evaluation of Framework Adequacy

The analysis does not support an unqualified claim that established organisational cyber security risk management frameworks are either fully adequate or obsolete. The asset–threat–vulnerability–risk–treatment structure associated with ISO-aligned governance remains useful because it provides a disciplined process for classifying assets, identifying threats, assessing risks, assigning controls, and reviewing treatment decisions [5]. However, the case study shows that this structure becomes adequate for AI-related integrity risk only when it is deliberately populated with AI-specific lifecycle assets, behavioural failure modes, monitoring requirements, assurance evidence, and ownership responsibilities.

This is a qualified finding. Organisations should not assume that conventional cyber security governance automatically covers AI-related integrity risk. The emergence of AI-specific standards, AI-focused risk frameworks, adversarial AI taxonomies, and dedicated AI regulation indicates that AI creates governance requirements that general cyber security frameworks do not always make explicit. The case-study evidence therefore supports augmentation and integration, not reliance on conventional frameworks alone.

The Chapter 4 analysis shows why this qualification matters. The asset register identified AI models, datasets, feature pipelines, drift thresholds, retraining workflows, model registries, and model-signing mechanisms as integrity-critical governance assets. The vulnerability analysis identified absent robustness testing, weak dataset validation, incomplete lifecycle documentation, undefined drift thresholds, and gaps in governance ownership. The risk assessment showed that these omissions could produce high-severity risks, including adversarial evasion, training-data poisoning, undetected drift, false governance assurance, and governance omission. The control evaluation then showed that existing controls remained primarily oriented towards exploit-based compromise rather than probabilistic behavioural degradation.

In the case-study analysis, existing frameworks therefore appear adequate in structure but insufficient in operational visibility. The problem is not that the frameworks lack risk management logic. Rather, important AI-related risks may remain under-classified unless they are deliberately modelled across the AI lifecycle, monitoring environment, and governance layer. This is particularly important because AI-related integrity failure may occur through authorised input channels, routine retraining activities, gradual

model drift, or incomplete assurance processes rather than through discrete intrusion or unauthorised modification.

This conclusion is reinforced when the case-study findings are read alongside AI-specific governance expectations. AI governance standards emphasise lifecycle visibility, defined responsibilities, documented assurance, traceability, monitoring, and continual review. Measured against those expectations, conventional governance processes may remain incomplete unless AI lifecycle dependencies, behavioural monitoring obligations, provenance mechanisms, assurance practices, and control ownership are made explicit.

The dissertation therefore concludes that, within the limits of the case study, established organisational cyber security governance frameworks may remain useful as organisational foundations, but should not be treated as automatically sufficient for AI-related integrity risk. Their adequacy depends on integration with AI-specific asset classification, lifecycle governance, behavioural assurance, monitoring thresholds, provenance controls, explicit ownership, and strategic cryptographic agility. This conclusion should not be read as a universal claim about all organisations or all AI systems. It is a case-study-based judgement about the conditions under which established organisational cyber security frameworks can remain useful when confronted with AI-related integrity risks.

5.4 Deterministic Governance Bias

The findings suggest the presence of what may be described as deterministic governance bias. In this dissertation, the term is used as an interpretive concept rather than as an empirically validated organisational phenomenon. It refers to a tendency for governance artefacts to prioritise risks associated with intrusion, tampering, control failure, and unauthorised modification, while giving less explicit attention to probabilistic degradation, behavioural manipulation, lifecycle drift, and incomplete assurance visibility.

This concept helps explain the pattern identified in Chapter 4 and evaluated in section 5.3. The case study did not show an absence of governance processes; it showed that those processes were shaped by assumptions more appropriate to conventional software systems than to AI-enabled systems. As a result, probabilistic degradation, authorised-interface manipulation, lifecycle drift, and incomplete assurance visibility could remain under-modelled even within a formally structured ISMS.

Deterministic governance bias therefore functions as an explanatory lens rather than as a criticism of the standards themselves. It clarifies how established governance practices may lag behind the characteristics of the systems they are intended to govern. Recognising this bias helps explain why explicit AI asset classification, lifecycle governance, behavioural monitoring, model assurance, provenance visibility, and clear control ownership are necessary within otherwise mature frameworks.

5.5 Contribution to Knowledge and Practice

The contribution of this dissertation is integrative and evaluative. It does not propose a new technical defence, AI risk framework, or cryptographic mechanism. Instead, it shows how AI-related integrity risks may remain insufficiently visible within organisational governance when probabilistic system behaviour is assessed through assumptions inherited from deterministic information systems. The contribution is developed in three areas: conceptual clarification, methodological integration, and practical governance adaptation.

5.5.1 Conceptual Contribution

Conceptually, the dissertation reconceptualises integrity in AI-enabled governance as behavioural reliability, not merely protection against unauthorised modification. Integrity in AI-enabled systems also depends on resilience to adversarial manipulation, robustness to corrupted or shifting data, and stability across model lifecycle processes. The associated concept of deterministic governance bias provides a governance-level explanation for why mature frameworks may still fail to make AI-related integrity risks sufficiently visible.

5.5.2 Methodological Contribution

Methodologically, the dissertation shows how DES can be combined with ISO/IEC 27005 and programme-derived analytical lenses from CML, MAL, CTI, SIM, and QUC to evaluate governance completeness in a structured and traceable way [8]. The contribution lies in integrating technical and governance perspectives into an evaluative method that links assets, threats, vulnerabilities, controls, governance artefacts, and accountability structures.

The MSc-derived lenses changed the analysis in specific ways. CML reframed integrity as behavioural reliability under probabilistic conditions [10]. MAL sharpened the analysis of adversarial adaptation [16]. CTI informed likelihood assessment through attacker capability, intent, access, and opportunity [20]. SIM shaped the evaluation of detection thresholds, escalation, containment, rollback, and post-incident learning [21]. QUC positioned quantum computing as a strategic governance pressure around cryptographic agility and dependency planning, rather than as a parallel operational case [11].

This methodological contribution is integrative rather than technically novel. It shows how existing methods can be prioritised and combined to test whether AI-related integrity risks are visible, assessable, and governable within established organisational cyber security risk management structures.

5.5.3 Practical Contribution

Practically, the dissertation proposes a governance pathway for adapting existing organisational structures to AI-enabled environments. Its value lies in showing how AI-

related integrity risks can be made more visible and manageable through governance enhancement rather than framework replacement.

5.6 Quantum Computing as Strategic Governance Context

Quantum computing is not the primary operational focus of this dissertation, but it remains relevant as a strategic governance context. Its inclusion reinforces the broader argument that cyber security governance must remain adaptable when technological change destabilises long-standing assumptions. Just as AI-enabled systems challenge deterministic integrity assumptions, quantum computing challenges assumptions about the durability of classical cryptographic trust.

The Chapter 4 analysis showed that quantum-related risks in the case study differ from the immediate operational logic of AI-related integrity risks. Their immediate likelihood is lower, but their long-term strategic impact may be severe, especially in relation to PKI compromise, future decryption exposure, and the absence of a formal cryptographic-agility migration plan [7], [6], [22], [30]. For that reason, quantum computing enters the dissertation not as a second full operational case domain, but as a governance pressure that reinforces the need for forward-looking risk management, dependency awareness, and migration planning.

This positioning supports the dissertation's overall claim. Governance adequacy must be judged not only by how well frameworks manage current threats, but also by whether they can make emerging technological pressures visible early enough for strategic action.

5.7 Limitations and Future Research

This study is governance-focused and analytical in design. It does not conduct empirical adversarial robustness testing, quantitative risk modelling, or sector-wide survey research. The case study is hypothetical rather than empirical and, although designed to reflect realistic organisational conditions, does not claim direct generalisability across sectors or institutions.

These limitations are consistent with the chosen research aim. The dissertation was designed to evaluate governance adequacy rather than measure empirical prevalence or technical attack performance. Nonetheless, they define clear boundaries around the strength of the claims made. The findings should therefore be read as case-study-based analytical conclusions, not as evidence that all organisations using ISO-aligned cyber security frameworks will experience the same governance gaps.

The hypothetical design also means that the study does not capture the organisational politics, resource constraints, legacy system dependencies, procurement pressures, vendor relationships, audit practices, or informal decision-making processes that would affect AI governance in a real organisation. In practice, the visibility of AI-related integrity risk may depend not only on formal frameworks, but also on organisational

maturity, security culture, board attention, regulatory exposure, technical capability, and the relationship between ML engineering, security operations, risk management, and executive governance. A real organisational setting might therefore reveal additional implementation barriers, different likelihood judgements, or stronger existing AI controls than those assumed in the case study.

The study is also limited by its deliberately narrow scope. It focuses on AI-related integrity risk within organisational cyber security governance, while treating quantum computing as a strategic governance context rather than as a second operational case domain. This supports analytical coherence, but it means the dissertation does not evaluate post-quantum migration in technical detail, compare cryptographic transition strategies, or assess the operational implementation of post-quantum controls.

Future research could extend this work in four directions. Empirical studies could examine how ISO-certified organisations currently represent AI-related integrity risks in asset registers, risk registers, assurance processes, and control ownership structures. Quantitative methods could support risk scoring for probabilistic integrity degradation, including adversarial evasion, data poisoning, model drift, and false assurance from performance metrics. Sector-specific studies could compare governance maturity across financial services, defence, health, and critical infrastructure. Further work could also examine how post-quantum migration planning is integrated into broader AI-enabled governance environments.

5.8 Final Response to the Research Question

The case-study analysis supports a qualified answer to the research question. Existing organisational cyber security risk management frameworks are adequate only where AI-specific risks are explicitly represented, owned, monitored, and reviewed. They provide a useful structure for asset identification, threat analysis, vulnerability assessment, risk treatment, control ownership, and continual review. However, they do not automatically make AI-related integrity risks visible when those risks arise through probabilistic model behaviour, authorised input manipulation, training-data corruption, model drift, lifecycle dependencies, or incomplete assurance evidence.

In the case study, adequacy depended on whether AI-specific assets and risks were explicitly represented within the governance process. Once the ML model, training dataset, validation dataset, feature pipeline, drift detection mechanism, retraining workflow, model registry, signing mechanism, and governance artefacts were classified and mapped, the ISO-aligned risk process became capable of supporting meaningful governance. Before that modelling step, the same organisational framework appeared procedurally mature but remained operationally incomplete.

The answer is therefore not that established frameworks should be rejected, nor that they are automatically sufficient. Rather, they are conditionally adequate as governance foundations when integrated with AI-specific lifecycle visibility, behavioural monitoring, adversarial assurance, provenance controls, explicit ownership, and strategic cryptographic-agility planning.

5.9 Reflection of the Research Process

This dissertation began with a broad ambition to examine cyber security risks and mitigations at the intersection of AI/ML and quantum computing. One of the most important lessons from the research process was that intellectual breadth does not necessarily produce analytical strength. Early framing risked dispersing the study across technical, regulatory, ethical, and strategic themes without a sufficiently unified method. Narrowing the focus to AI-related integrity risk within organisational cyber security governance was therefore not merely a practical adjustment, but a substantive improvement in the research design.

A second important lesson concerned the value of methodological discipline. Design for Security (DES) proved particularly useful because it enabled explicit modelling of system boundaries, trust assumptions, governance actors, and interacting domains. This was critical because the central problem was not simply whether vulnerabilities existed, but whether governance structures made them visible, assessable, and controllable. DES therefore provided both a method of analysis and a means of exposing omissions, implicit assumptions, and incomplete modelling.

The research process also showed that a descriptive treatment of architecture was insufficient for dissertation-level evaluation. The analysis required architecture, governance responsibilities, standards, and risk artefacts to be integrated in a traceable way. This helped ensure that findings were not presented as general claims, but were grounded in the case study's asset classification, threat analysis, vulnerability mapping, risk assessment, and control evaluation.

The MSc course lenses further strengthened the analysis. Classical Machine Learning clarified the probabilistic nature of AI-enabled systems. Malware Analysis and Cyber Threat Intelligence strengthened adversarial reasoning and attacker profiling. Security Incident Management helped frame monitoring, escalation, and response in contexts where integrity degradation may emerge gradually rather than through discrete incidents. Quantum Cryptography helped reposition quantum computing from an over-broad parallel topic into a more analytically useful strategic governance context. This reframing allowed quantum security to be retained as a strategic governance concern without displacing the core AI-integrity analysis.

Overall, the research process reinforced the importance of scope discipline, evidential traceability, and methodological alignment. These were lessons not only about completing the dissertation, but also about analysing interdisciplinary security problems more effectively.

5.10 Concluding Statement

The central implication of this case-study-based dissertation is that AI-related integrity risk should not be governed through uncritical reliance on conventional cyber security frameworks. At the same time, the case study does not establish a need for wholesale replacement of organisational risk management structures. The more defensible conclusion is that established structures can remain useful where they are explicitly

integrated with AI-specific governance standards, lifecycle assurance, adversarial risk modelling, behavioural monitoring, provenance mechanisms, defined control ownership, and cryptographic agility planning.

The dissertation's contribution is to show how AI-related integrity risks may become visible, assessable, and governable when DES, ISO/IEC 27005, AI-specific standards, and MSc-derived analytical lenses are applied in a structured and prioritised way. Governance adequacy depends not only on the existence of frameworks, but on whether those frameworks are applied with sufficient visibility into the lifecycle, behavioural, and strategic dependencies of AI-enabled systems.

References

- [1] International Organization for Standardization, “ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology,” *International Standard*, pp. 1–60, Jul. 2022, Accessed: Apr. 18, 2026. [Online]. Available: <https://www.iso.org/standard/74296.html>
- [2] International Organization for Standardization, “ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML),” *International Standard*, pp. 1–36, Dec. 2022, Accessed: Apr. 18, 2026. [Online]. Available: <https://www.iso.org/standard/74438.html>
- [3] International Organization for Standardization, “ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system,” *International Standard*, pp. 1–51, Dec. 2023, Accessed: Apr. 18, 2026. [Online]. Available: <https://www.iso.org/standard/42001>
- [4] International Organization for Standardization, “ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management,” *International Standard*, pp. 1–26, Feb. 2023, Accessed: Apr. 16, 2026. [Online]. Available: <https://www.iso.org/standard/77304.html>
- [5] International Organization for Standardization, “ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection — Information security management systems — Requirements,” *International Standards*, pp. 1–19, Oct. 2022, Accessed: Apr. 18, 2026. [Online]. Available: <https://www.iso.org/standard/27001>
- [6] D. Moody, R. Perlner, A. Regenscheid, A. Robinson, and D. Cooper, “Transition to Post-Quantum Cryptography Standards,” *National Institute of Standards and Technology*, 2024, doi: <https://doi.org/10.6028/NIST.IR.8547.ipd>.
- [7] National Cyber Security Centre, “Timelines for migration to post-quantum cryptography,” Online. Accessed: Apr. 14, 2026. [Online]. Available: <https://www.ncsc.gov.uk/guidance/pqc-migration-timelines>
- [8] M. Jackson, *Problem Frames: Analyzing and structuring software development problems*. Addison-Wesley Longman Publishing Co., Inc., 2000.
- [9] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014, Accessed: Apr. 14, 2026. [Online]. Available: <https://doi.org/10.48550/arXiv.1412.6572>
- [10] R. Collins, “Classical Machine Learning (CML) lecture materials,” Professional Masters Programme, University of Oxford, 2022.
- [11] S. Gogioso, “Quantum Cryptography (QUC) lecture materials,” Professional Masters Programme, University of Oxford, 2023.
- [12] National Institute of Standards and Technology, “Artificial intelligence risk management framework (AI RMF 1.0),” Online. Accessed: Apr. 14, 2026. [Online]. Available: <https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10?ref=decisionarts.io>
- [13] European Union, “Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act),” *Official Journal of the European Union*, pp. 1–13, Jul. 2024, Accessed: Apr. 16, 2026. [Online]. Available: <https://ec.europa.eu/newsroom/dae/redirection/document/112455>
- [14] MITRE ATLAS, “Adversarial Threat Landscape for Artificial-Intelligence Systems,” Online. Accessed: Apr. 25, 2026. [Online]. Available: <https://atlas.mitre.org/>
- [15] International Organization for Standardization, “ISO/IEC 5962:2021 Information technology — SPDX® Specification V2.2.1,” *International Standard*, pp. 1–145, Aug.

- 2021, Accessed: Apr. 18, 2026. [Online]. Available: <https://www.iso.org/standard/81870.html>
- [16] G. Suarez-Tangil, G. Stringhini, and H. Menendez, “Malware Analysis (MAL) lecture materials,” University of Oxford, 2022.
- [17] X. Zhang, X. Zheng, and W. Mao, “Adversarial perturbation defense on deep neural networks,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–36, 2021.
- [18] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *2017 IEEE Symposium on Security and Privacy*, IEEE, 2017, pp. 39–57.
- [19] C. Zhang, X. Costa-Perez, and P. Patras, “Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms,” *IEEE/ACM Transactions on Networking*, vol. 30, no. 3, pp. 1294–1311, 2022.
- [20] J. Lusthaus, “Cyber Threat Intelligence (CTI) lecture materials,” Professional Masters Programme, University of Oxford, 2022.
- [21] D. King, “Security Incident Management (SIM) lecture materials,” Professional Masters Programme, University of Oxford, 2021.
- [22] R. L. Rivest, A. Shamir, and L. Adleman, “A method for obtaining digital signatures and public-key cryptosystems,” *Commun. ACM*, vol. 21, no. 2, pp. 120–126, 1978.
- [23] C. Esposito, G. A. Landrum, N. Schneider, N. Stiefl, and S. Riniker, “GHOST: adjusting the decision threshold to handle imbalanced data in machine learning,” *J. Chem. Inf. Model.*, vol. 61, no. 6, pp. 2623–2640, 2021.
- [24] S. Idowu, D. Strüber, and T. Berger, “Asset management in machine learning: A survey,” in *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, IEEE, 2021, pp. 51–60.
- [25] Z. Tian, Z. Wang, A. M. Abdelmoniem, G. Liu, and C. Wang, “Knowledge representation of training data with adversarial examples supporting decision boundary,” *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4116–4127, 2023.
- [26] L. Mauri and E. Damiani, “Estimating degradation of machine learning data assets,” *ACM Journal of Data and Information Quality (JDIQ)*, vol. 14, no. 2, pp. 1–15, 2021.
- [27] D. Kreuzberger, N. Kühl, and S. Hirschl, “Machine learning operations (MLOps): Overview, definition, and architecture,” *IEEE access*, vol. 11, pp. 31866–31879, 2023.
- [28] T. Paladini, F. Monti, M. Polino, M. Carminati, and S. Zanero, “Fraud detection under siege: Practical poisoning attacks and defense strategies,” *ACM Transactions on Privacy and Security*, vol. 26, no. 4, pp. 1–35, 2023.
- [29] P.-F. Zhang, Z. Huang, X.-S. Xu, and G. Bai, “Effective and robust adversarial training against data and label corruptions,” *IEEE Trans. Multimedia*, vol. 26, pp. 9477–9488, 2024.
- [30] A. Akbar, “Analyzing the Harvest Now, Decrypt Later Threat and Post-Quantum Cryptography Solutions: A Systematic Literature Review,” *Migration*, vol. 2, p. 10, 2025.

Appendices

Annex A Risk Management

Risk management was applied throughout the dissertation as a structured element of project planning and control. The principal purpose of this annex is to identify the main risks affecting the successful completion and quality of the dissertation, assess their significance, and record the mitigation measures adopted.

The principal risks were excessive scope, time pressure arising from professional and personal commitments, and the possibility that weaknesses identified in the previous assessment might remain insufficiently addressed. These risks were assessed in terms of likelihood and impact and were treated through progressive narrowing of scope, adoption of a single overarching research question, use of a structured five-stage analytical framework, and scheduled revision against assessor and supervisory feedback.

Table 14 records the main risks, mitigation actions, and review dates applied during the dissertation process.

Risk ID	Risk Description	Likelihood (Low / Medium / High)	Impact (Low / Medium / High)	Mitigation Strategy	Status (Open / Closed)	Last Reviewed
1	Non-submission of thesis	High	High	Prioritised feedback from the two assessors. Sustained self-motivation.	Closed	01 Sep 2025
2	Scope too broad	High	High	Reduced from five research questions to one question. Narrowed dissertation to governance adequacy. Refocused on AI-related integrity risk. Treated quantum as strategic context.	Closed	01 Oct 2025
3	Methodology insufficiently systematic	High	High	Replaced diffuse approach with a five-stage modelling framework. Mapped stages to DES, ISO/IEC 27005, and MSc methods. Clarified method selection and analytical purpose.	Closed	01 Dec 2025
4	Analysis and conclusion misaligned	High	High	Anchored Chapter 5 to Chapter 4 outputs. Linked discussion directly to Figure 4 and case tables. Removed or reduced unsupported framework-style drift.	Closed	12 Jan 2026
5	Figure 4 lacked analytical force	Medium	High	Redesigned as layered analytical architecture. Added risk entry, propagation, controls, artefacts, accountability.	Closed	15 Apr 2026
6	Scope too broad / unfocused	Medium	High	Narrowed to AI-related integrity governance. Treated quantum as strategic context only. Realigned title, abstract, and research question.	Closed	15 Apr 2026
7	Methods not sufficiently applied	Medium	High	Mapped DES, CML, MAL, CTI, SIM, QUC across stages. Added method selection justification. Linked methods to case analysis.	Closed	15 Apr 2026
8	Dissertation too descriptive	Medium	High	Strengthened Chapter 4 analysis- Added explicit evaluative reading of Figure 4. Shifted from description to governance evaluation.	Closed	16 Apr 2026

9	Contribution and argument unclear	Medium	High	Clarified governance gap and deterministic governance bias. Added context, application, reflection contribution framing.	Closed	16 Apr 2026
10	AI governance framing too weak	Medium	High	Moved AI-specific standards to front- Added ISO/IEC 22989, 23053, 42001, 23894, 5962. Reframed ISO 27001/27005 as governance backbone.	Closed	16 Apr 2026
11	AI lifecycle risks	Medium	High	Expanded asset, threat, vulnerability, and risk tables. Added lifecycle, monitoring, retraining, registry, provenance visibility.	Closed	23 Apr 2026
12	Governance visibility too weak	Medium	High	Added governance omission → risk invisibility logic. Added control ownership, assurance review, monitoring/escalation pathways	Closed	23 Apr 2026
13	Discussion insufficiently evidenced	Medium	High	Anchored Chapter 5 in Chapter 4 tables and Figure 4. Added explicit signposting from findings to interpretation.	Closed	23 Apr 2026
14	Presentation and compliance risks	Medium	High	Corrected headings, numbering, wording, references. Updated TOC/figures/tables. Retained risk/project management and ELSI annexes.	Closed	23 Apr 2026
15	Evaluation insufficiently explicit	Medium	High	Revised 3.2.2 to justify DES against alternatives such as NIST AI RMF, STRIDE, technical ML robustness testing, malware-focused modelling, and CTI-only profiling. Added explicit limitations of DES.	Closed	02 May 2026
16	Conclusions too broad or insufficiently evidenced	Medium	High	Tightened the Abstract, 5.3, 5.8, and 5.10 to frame findings as case-study-based and conditional, avoiding absolute claims about framework sufficiency or replacement.	Closed	02 May 2026
17	Method prioritisation unclear	Medium	High	Revised 3.3.5 to state that DES and ISO/IEC 27005 are central methods, while CML, MAL, CTI, SIM, and QUC operate as supporting analytical lenses.	Closed	03 May 2026
18	Interpretation of Figure 4 and key tables insufficiently deep	Medium	High	Revised 4.7.1 to explain what Figure 4 reveals that a conventional risk register would not: risk entry, propagation, lifecycle dependencies, control visibility, and governance gaps. Added interpretive sentences after key tables.	Closed	03 May 2026
19	Limitations insufficiently critical	Medium	High	Revised 5.7 to explain what the hypothetical design does not capture, including organisational politics, resource constraints, legacy systems, vendor relationships, audit practices, and informal governance behaviours.	Closed	04 May 2026
20	MSc learning not sufficiently explicit in application	Medium	High	Revised 5.5.2 to explain how each MSc lens changed the analysis: CML reframed integrity, MAL sharpened adversarial reasoning, CTI informed likelihood, SIM shaped monitoring/escalation, and QUC positioned quantum as strategic governance context.	Closed	04 May 2026
21	Failure to submit on 08May26 due to personal, technical and/or other circumstances	High	High	Submitted early. Ensured that a downloadable version of the file named Submit_06May was successful.	Closed	06 May 2026

Table 14. Risk Register

Annex B Project Management

This dissertation was managed as a structured analytical project organised across successive phases of development. The principal objective was to evaluate whether existing organisational cyber security risk management frameworks adequately address AI-related integrity risks, while maintaining a scope appropriate to a master's dissertation.

Project delivery was organised in five phases: topic definition, literature review, methodological design, case-study analysis, and synthesis of findings. This phased structure supported progression from problem framing to final evaluation and helped maintain alignment between the research question, selected methods, Chapter 4 evidence, and Chapter 5 conclusions.

Project management also required disciplined coordination of drafting, revision, and integration. Attention was given to maintaining consistency between chapters, strengthening the evidential link between analysis and conclusions, refining the structure of the final discussion chapter, and ensuring that references, figures, tables, and annexes were correctly aligned with the argument of the dissertation.

Quality assurance was addressed through repeated editing cycles, chapter-to-chapter consistency checks, and final review of presentation, citation accuracy, and document structure. In this way, project management supported coherent delivery of the dissertation as a complete academic submission.

Annex C Ethical, Legal, Social and Professional Issues

On 24 March 2025, clarification was sought from the University's Ethics Committee regarding whether formal ethics approval was required for this project. The Ethics Officer confirmed that approval was not necessary, as the study did not involve human participants, personal or sensitive data, classified material, or activities presenting material legal, social, or professional risk, in accordance with the University's Ethics Policy.

The project scope was subsequently refined from an initial intention to use open-source datasets. The final study relies exclusively on publicly available sources and did not involve interviews, surveys, experimental procedures, or personally identifiable information.

Although formal ethics approval was not required, ethical and professional considerations remained important throughout the research. These included accurate representation of cited sources, clear distinction between analytical findings and speculative future developments, and avoidance of exaggerated claims regarding current AI or quantum capabilities. Given the pace and uncertainty of both fields, particular care was taken to frame conclusions proportionately and to avoid overstating either immediate technical risk or long-term strategic impact.

Legal, social, and professional considerations were also relevant because the dissertation examines organisational cyber security governance. The analysis is therefore framed at a general governance level and does not disclose sensitive operational procedures, exploit techniques, or confidential organisational information. More broadly, the study recognises that weaknesses in the governance of AI-enabled systems may affect trust, accountability, service reliability, and organisational resilience. By focusing on governance adequacy rather than technical novelty alone, the dissertation aims to support responsible, transparent, and sustainable organisational adoption of emerging technologies.