POLITECNICO DI TORINO

MASTER's Degree in CYBERSECURITY ENGINEERING

MASTER's Degree Thesis

AI Forensics: An Experimental Analysis of Transparency, Security, and Accountability in Automated Decision-Making Systems

Supervisors

Prof. Andrea ATZENI

Candidate

Giovanni SCAMARDO

Academic Year 2025/26

Summary

Artificial Intelligence systems are increasingly deployed in high-risk domains where their outputs may influence investigative processes and legally relevant decisions. In such contexts, predictive performance alone is insufficient: AI systems must be auditable, reproducible, and forensically analyzable.

This thesis investigates whether modern AI models—often perceived as opaque black boxes—can be systematically examined and treated as structured digital artefacts. The work combines theoretical analysis and experimental validation, focusing on face recognition as a representative high-risk application. The study integrates explainability techniques, adversarial robustness evaluation, demographic representation analysis, and backdoor vulnerability assessment within a unified forensic logging and integrity framework.

The experimental results show that representation-level bias can be quantitatively measured in embedding space and that structural disparities correlate with heterogeneous adversarial susceptibility. Gradient-based attacks induce measurable embedding shifts, while training-time backdoor poisoning produces persistent and targeted manipulation detectable through embedding analysis, explainability artefacts, and cryptographic model fingerprinting.

A central contribution of this thesis is the proposed AI forensic framework, which enforces traceability, integrity verification, and chain-of-custody principles across the entire experimental pipeline. The findings demonstrate that AI systems can be rendered auditable and reconstructable when treated as forensic objects, supporting accountability without replacing human judgment. This work provides a methodological foundation for aligning high-risk AI systems with evidentiary standards and regulatory requirements.

Table of Contents

Chapter 1

Introduction

Over the last few decades, artificial intelligence (AI) has become one of the most influential technologies of our time, with the potential to transform healthcare, finance, justice, industry, and security.[1, 2] This rapid growth has been driven by the combination of increased computing power, the availability of large-scale datasets, and significant advances in machine learning algorithms.[1] With the advent of deep learning and, more recently, large language models (LLMs), AI systems have achieved state-of-the-art performance on a variety of narrowly defined benchmark tasks, including image recognition, natural language processing, machine translation, medical image analysis, and the generation of textual and visual content, in some cases matching or surpassing human expert performance under benchmark-specific evaluation protocols.[3, 4, 5]

However, as AI systems are integrated into high-stakes decision-making processes, their increasing complexity raises equally significant concerns about opacity and lack of control.[6, 7] Modern deep neural networks and LLMs typically operate as black boxes: they may produce highly accurate predictions, yet the internal reasoning that leads from input to output remains largely hidden from users, domain experts, and regulators.[6, 8] In many real-world settings, this opacity makes it difficult to assess whether a model is making decisions for the right reasons, to detect changes in its behaviour over time, or to assign responsibility when something goes wrong.[9, 10]

These concerns are not only theoretical. Several empirical studies have shown that AI systems can systematically reproduce or amplify social biases when deployed in the wild.[11] Commercial facial recognition systems, for instance, have been reported to exhibit markedly higher error rates for women and people with darker skin tones than for light-skinned men, revealing demographic disparities that are unacceptable in forensic or security applications.[11, 12] In the criminal justice domain, risk assessment tools such as COMPAS have been criticised for assigning disproportionately high risk scores to defendants from certain groups with profiles comparable to those of other defendants, raising questions about fairness, discrimination, and due process.[13] In healthcare, predictive models trained on biased proxies of health risk have been shown to underestimate the needs of specific patient populations, with potentially serious consequences for access to care.[14]

At the same time, large language models have introduced a new class of AI systems that interact with users in highly natural ways and are increasingly used to draft legal documents,

summarise court decisions, support medical triage, write software, and mediate access to information for millions of people.[5] Their outputs are often persuasive and fluent, but they may contain hallucinated facts, subtle stereotypes, or sensitive information that has been memorised from training data.[15, 16, 17] Without appropriate explanation and auditing mechanisms, it is difficult to assess when an answer is trustworthy, to identify the sources on which it is based, or to verify whether the model behaves consistently across users, languages, and time.[15]

In this context, explainable artificial intelligence (XAI) and, more broadly, AI forensics seek to turn opaque model behaviour into artefacts that can be understood, scrutinised, and contested.[18, 9] The goal is not only to provide intuitively appealing visualisations or user-friendly narratives, but to build technical and procedural tools that make AI decisions traceable, reproducible, and legally defensible.[10, 7] Explanations should help stakeholders answer questions such as: Why was this specific decision taken? Which features or patterns were most influential? Would the outcome have been different under slightly different conditions? Are similar cases treated in a consistent manner?[9, 7, 10]

From a legal and regulatory perspective, these issues are becoming increasingly urgent.[19] Emerging frameworks such as the EU General Data Protection Regulation and the EU AI Act emphasise transparency, accountability, and risk management for AI systems, especially when they are used in high-risk domains like biometric identification, credit scoring, or access to essential services.[20, 19] In such contexts, an unexplainable model is not merely a technical inconvenience: it can undermine fundamental rights, hinder the possibility of effective redress, and weaken trust in institutions that rely on algorithmic decision-making.[19]

This thesis is situated at the intersection between XAI and AI forensics, building on recent work on interpretable and accountable machine learning. While the case study developed in the following chapters focuses on face verification systems based on deep neural networks, the underlying questions are shared with many other applications, including LLM-based services: How can we systematically analyse and document model behaviour? Which explanation techniques are suitable in forensic settings? How can we evaluate their reliability, fairness, and robustness under realistic and even adversarial conditions?[7, 9, 10, 21] By addressing these questions, the thesis aims to contribute to a more rigorous and operational notion of explainability, one that is adequate not only for debugging and model development, but also for evidential and accountability purposes.[21]

Chapter 2

Explainable Artificial Intelligence (XAI)

2.1 Origins and Purposes of Explainable AI

This chapter introduces the main conceptual foundations and methodological families of Explainable Artificial Intelligence (XAI), with a focus on how explanations are constructed and evaluated.[22, 7] The acronym XAI is often traced back to DARPA's Explainable Artificial Intelligence (XAI) program, launched in 2016, which helped consolidate the term and formalised two main directions: (1) developing learning algorithms that produce more interpretable models without significantly reducing performance and (2) designing explanation interfaces that make complex model behaviour understandable to human users.[18, 23]

Core XAI concepts include:

- **Human-understandability**: Explanations should be accessible to non-expert users and adapted to the needs and expertise of their intended audience.[9]
- **Fidelity (faithfulness)**: Explanations should reflect the model's decision logic as accurately as possible, that is, they should be faithful to the underlying model rather than expressing general reliability or trustworthiness of the system.[7, 24]
- **Performance trade-off**: Interpretability must be balanced against predictive accuracy.

These recurring themes are discussed in several surveys and monographs on XAI.[22, 7, 9] XAI typically distinguishes **global explanations**, which describe the overall behaviour of a model (for example, which features are most influential on average), from **local explanations**, which focus on a single prediction or instance (for this input, these features contributed in this way).[22, 7] Another critical challenge is assessing explanation fidelity (faithfulness), i.e., measuring how accurately an explanation represents the model's internal decision logic rather than providing a misleading simplification.[7] This issue is particularly relevant for model-agnostic techniques, where the internal information of the model is not directly used.

Recent research also emphasises human-centred evaluation of explanations, analysing how useful, understandable, and trustworthy they are for different stakeholders.[9, 18]

2.2 Inherently Interpretable Models and Post-hoc Explanation Methods

Explainable Artificial Intelligence (XAI) techniques are mainly classified into two groups of methods: *inherently interpretable models*, which are understood intuitively, and *post-hoc explanation methods*, where the functioning of complex models that have already been trained is explained afterwards.[22, 7]

2.2.1 Inherently Interpretable Models

Inherently interpretable models are those algorithms that provide a direct and clear insight into the underlying decision-making process.[7] They are designed so that the input–output relationship is directly inspectable, often without requiring additional post-hoc explanation techniques.

The most common inherently interpretable models include:

- **Linear and Logistic Regression:** decisions are based on the weights assigned to individual variables so as to make it easier to identify which characteristics influence the positive or negative prediction.[7]
- **Decision Trees:** a decision tree consists of nodes, each representing a condition on some variable, and leaves that predict the outcome. The path from the root through the nodes to a leaf can be interpreted as a sequence of explicit logical rules.[7]
- **Rule-based Models:** these models operate through a set of *if–then* rules that clearly define the criteria for decision-making.[22]
- **Generalized Additive Models (GAMs)** and more recent variants such as **Explainable Boosting Machines (EBMs):** in these models, the prediction is represented as the sum of contributions from each variable, maintaining interpretability even when moderate non-linear relationships exist.[7, 25]

Such models are widely discussed as intrinsically interpretable approaches in the literature on interpretable machine learning.[7, 22]

The main strengths of these models are their transparency, easy validation, and clarity in justifying decisions. However, they are limited in that they cannot capture high non-linearity or complexity within data, which may lead to lower predictive accuracy compared to more sophisticated models such as deep neural networks or ensemble methods.[7]

In highly critical domains such as medicine and law, several scholars—including Cynthia Rudin—argue that inherently interpretable models should be preferred whenever feasible, since post-hoc explanations may not provide the same level of transparency and verifiability.[6]

2.2.2 Post-hoc Explanation Methods

Post-hoc explanation methods are applied after a model has been trained and are used to clarify how it makes decisions, without modifying its internal structure.[22, 7] They analyze

the input data, the predictions, and, when possible, the internal activations, in order to make complex models that behave like "black boxes" more understandable.[7]

These techniques are essential when highly accurate yet non-interpretable models are employed, such as deep neural networks or ensemble models. Among the main post-hoc methods widely adopted in research and practical applications are LIME, SHAP, Grad-CAM, and Integrated Gradients, each with specific characteristics and application domains.[7]

2.2.2.1 LIME

Local Interpretable Model-agnostic Explanations (LIME) is a technique that emphasizes providing local explanations, meaning that it will interpret a particular prediction made by the model instead of giving an overall interpretation of the model. To achieve this, LIME creates a simplified and interpretable surrogate model, frequently linear, that mimics the decision boundary of the original complex model in a small vicinity surrounding the input under analysis.[26]

LIME's major advantage is its adaptability, since it does not rely on the model's internal structure and can practically be used with any machine learning algorithm.[26] However, this very characteristic has its downsides as well. The local approximations that LIME generates can be unstable, meaning that small changes in the input or sampling procedure may lead to substantially different explanations.[7]

Further, because LIME relies on a locally fitted surrogate model, it may not faithfully capture the original model's decision logic.[7]

2.2.2.2 SHAP

SHAP (*SHapley Additive exPlanations*) is based on cooperative game theory and assigns each variable a Shapley value, which represents the average contribution of that feature to the model's prediction relative to a reference situation.[27, 7] This approach provides strong theoretical guarantees, including *consistency*, meaning that a feature with greater influence cannot receive a lower importance score, and *additivity*, ensuring that, within the additive explanation model, the sum of feature contributions matches the model output.[27]

Thanks to these properties, SHAP provides coherent and interpretable explanations both at the local level, that is, for individual predictions, and at the global level, offering insight into the model's overall behaviour.[27, 7]

The major limitation of SHAP arises from the computation of Shapley values, which is computationally very demanding, particularly in the case of large datasets or models with many features. In addition, when features are strongly correlated, the attribution of importance may be noisy, since the shared influence among variables makes it more difficult to correctly interpret each feature's contribution.[7]

2.2.2.3 Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) is one of the most widely used explanation methods in computer vision.[28] It was originally developed for convolutional neural networks (CNNs) and produces class-discriminative heatmaps that highlight the regions of

an image most relevant to the model's decision.[28] Despite its practicality, Grad-CAM is primarily tailored to convolutional architectures and does not directly extend to non-visual modalities.[28]

2.2.2.4 Integrated Gradients

Integrated Gradients is a gradient-based method that can be applied to differentiable models, such as neural networks.[29] It is used to understand how much each feature contributes to the model's decision. To this end, it calculates and integrates the gradients of the model output along a continuous path from a reference input, representing a neutral point, to the real input being analyzed.[29, 7] Compared to simple gradient-based methods, Integrated Gradients is often more stable and interpretable because it mitigates saturation effects that can mask feature influence.[29] Nonetheless, the quality of the attribution depends on the choice of the reference input, and feature correlations can lead to ambiguous or misleading importance assignments.[7]

2.3 Comparison between Interpretable and Post-hoc Approaches

The comparison between inherently interpretable models and post-hoc explanation methods represents a central debate in Explainable Artificial Intelligence (XAI).[6, 7] While both approaches aim to enhance the transparency of AI decision-making processes, they rely on fundamentally different strategies, leading to distinct trade-offs in terms of transparency, predictive power, and reliability.[22, 7]

Inherently interpretable models embed transparency directly into their structure. Their decision logic is explicitly accessible and can be inspected without auxiliary tools, which facilitates validation, auditing, and justification of outcomes.[7, 6] This structural transparency is particularly valuable in high-stakes settings such as medicine and law, where accountability requirements are stringent.[6] However, such models may struggle to capture highly non-linear relationships or complex data patterns, where more expressive architectures often achieve superior predictive performance.[7]

Post-hoc explanation methods, in contrast, are applied to already trained models—typically high-performing black-box systems—with the goal of providing interpretative insights without altering the original architecture.[22, 7] Techniques such as LIME and SHAP enable the analysis of complex models in a flexible and model-agnostic manner.[26, 27] This adaptability constitutes a major advantage, as explanations can be generated for virtually any predictive system, including deep neural networks and ensemble models.[22, 7]

Despite this flexibility, post-hoc approaches raise important concerns regarding fidelity.[7, 6] Since explanations are derived externally from the model's internal mechanisms, they may not perfectly reflect the true decision logic. Empirical studies have shown that small variations in input data or sampling procedures can lead to substantially different explanations for the same prediction, highlighting issues of instability and local variability.[7, 26, 30, 31] Consequently, the central challenge in XAI research is not merely to generate explanations,

but to rigorously evaluate their robustness, consistency, and faithfulness.[9, 32]

From a practical perspective, inherently interpretable models offer the advantage of providing direct explanations without additional computational overhead. Post-hoc methods, on the other hand, can be computationally demanding—particularly in the case of SHAP—and may not always be suitable for real-time applications.[27, 7] Nevertheless, their ability to operate independently of model architecture makes them uniquely versatile.[22, 7]

In recent years, a third research direction has emerged, aiming to overcome the dichotomy between intrinsic interpretability and post-hoc explanation.[25, 33, 34, 35] Hybrid models seek to combine the predictive strength of complex architectures with structural mechanisms that preserve a degree of transparency.

A prominent example is provided by Explainable Boosting Machines (EBMs), an advanced extension of Generalized Additive Models (GAMs) that integrates interpretable additive functions with feature-level boosting techniques.[25] This design enables the learning of non-linear relationships and low-order feature interactions while maintaining an inspectable structure. Empirical findings suggest that EBMs can achieve accuracy comparable to, and occasionally exceeding, that of traditional black-box models while preserving high interpretability.[25]

Similarly, neural architectures designed to be interpretable by construction—such as Self-Explaining Neural Networks (SENN) and Prototype-Based Neural Networks (ProtoPNet)—embed explanatory components directly into their structure.[33, 34] By incorporating interpretable weights, prototypes, or conceptual bases, these models allow predictions to be systematically verified against their corresponding explanations.[33, 34]

Another hybrid approach is knowledge distillation, in which a complex model transfers its learned representations to a simpler and more interpretable model.[36, 37] This technique aims to retain much of the predictive power of deep networks while enabling greater transparency at inference time.

2.4 Explainability Across Application Domains

Explainable Artificial Intelligence (XAI) is an interdisciplinary field in which challenges and solutions change depending on the application context.[22, 7] Although much of the research focuses on developing general interpretability methods, experience gained in specific domains such as computer vision, Natural Language Processing (NLP), and recommender systems has led to the creation of particularly effective strategies and tools, valuable even beyond their original areas of use.[22]

Therefore, this section will focus on these key application areas to pinpoint approaches, challenges, and best practices that can be transferred to other contexts. In each of these fields, the notion of explainability takes on a well-defined meaning: in computer vision, it regards the identification of those visual regions where influence is exerted upon a decision; in NLP, it deals with building insight into the relationship between the model's inner representations and linguistic concepts; whereas in recommender systems, this means being able to provide understandable, transparent, and personalized justifications for the suggestions proposed by the system.

Across domains, several common principles for explainable model design can be iden-

tified, with particular emphasis on reusable techniques such as concept-based approaches, counterfactual explanations, and verifiable attention mechanisms.[7] The result is not only a deepening of the theoretical understanding of explainability; from a practical point of view, it also gives guidelines for the choice and adaptation of XAI tools in differing scenarios and contributes to a more coherent and generalizable methodological framework.

2.4.1 Computer Vision

Computer vision is a sub-field of artificial intelligence that allows computers to extract, process, and understand meaningful information through images and videos.[22] The purpose of this sub-field is to give models some or all of the capabilities of human vision, including object categorization, scene-context understanding, image segmentation, and tracking, in order to provide useful information for decision-making or predictive tasks that is not readily apparent from raw visual data. Computer vision relies on detecting patterns, structures, and spatial relationships in visual data.

In the area of computer vision, as in most deep neural network (DNN) applications, the need for model interpretability is increasing.[7] DNNs contain an extremely high degree of intrinsic complexity, which makes it challenging to determine which visual patterns most strongly influence model predictions.

To address the challenge of interpreting model decisions with respect to visual tasks, several visual interpretability techniques have been developed. As introduced in Section 2.2.2.3, Grad-CAM and LRP provide complementary attribution views for CNN-based vision models: the former highlights class-discriminative regions via gradient-weighted activations, whereas the latter redistributes relevance scores back to pixels.[28, 38]

These methods were an important milestone in interpretability, providing practical ways to visualise which regions contribute to neural network decisions.[28, 38] However, later studies questioned their reliability, reporting substantial variability in explanations under small model perturbations. This raised serious concerns about their stability and faithfulness.[7]

More recent research has extended interpretability efforts to Vision Transformers (ViTs), a new class of models that use self-attention mechanisms instead of convolutional operations.[39] ViTs partition an image into fixed-size patches and treat each patch as a token sequence.

This architecture allows modeling of long-range dependencies across disparate regions of an image, whereas CNNs operate through local receptive fields and learn visual structures hierarchically.[39] However, this advantage introduces interpretability challenges. In particular, it is difficult to determine how different attention heads distribute relevance across image patches and how attention evolves across layers.[40] Additionally, although attention maps can be extracted, they do not always reliably reflect which image regions are most responsible for a model's final prediction.[30]

To address these challenges, hybrid approaches combining gradient-based information with attention analysis have been proposed. For instance, *Attention Rollout* aggregates attention weights across multiple layers to provide a more comprehensive view of information flow within the model.[41] These advances showcase that, despite structural differences from CNNs, Vision Transformers can be analyzed and interpreted by carefully investigating attention dynamics and information propagation. Despite this, ViT interpretability remains

an active and rapidly evolving field of research, since the highly context-dependent and distributed nature of attention makes it particularly hard to establish clear intuitive connections between image regions and model reasoning.[40]

2.4.2 Natural Language Processing (NLP)

Natural language processing (NLP) is the subfield of artificial intelligence concerned with enabling machines to understand, generate, and interact in human language. In contemporary applications, many NLP systems take the form of large language models (LLMs) based on Transformer architectures, which amplifies the importance of explainability for generated text and interactive outputs.[4, 15] From a forensic perspective, this is particularly relevant because LLM outputs can be introduced as evidence-like artefacts (e.g., summaries, reports, chat transcripts). Consequently, prompts, model versioning, and decoding settings become part of the chain of custody and must be documented for reproducibility and accountability purposes.[42]

Core NLP tasks include text classification, sentiment analysis, machine translation, and conversational agents. These systems must capture linguistic structure, semantics, and contextual dependencies; therefore, language understanding is intrinsically complex and fundamentally grounded in syntactic and semantic relationships between words and sentences.[22]

Explainability is especially challenging in NLP due to the sequential and semantic nature of language. Small variations in word choice, phrasing, or syntactic structure can lead to substantial differences in interpretation and model predictions. Several approaches rely on attention mechanisms within Transformer models, or on attribution techniques such as Integrated Gradients[29] and LIME/SHAP adapted to textual data.[7] However, empirical studies have shown that attention weights are not necessarily faithful explanations of model behaviour, as they may capture statistical correlations rather than the true causal factors underlying a prediction.[30, 31]

To address these limitations, researchers have proposed rationale-based models that explicitly identify which parts of the input text determine a specific decision.[32, 24] In such approaches, the model is required not only to output a prediction, but also to select or generate a subset of tokens, phrases, or spans that serve as justifications for that prediction. A distinction is often made between *extractive rationales*, which highlight segments of the original input text, and *abstractive rationales*, which generate free-text explanations summarizing the reasoning process.

A crucial aspect of rationale-based modeling is the evaluation of faithfulness. It is not sufficient that the highlighted text appears plausible to a human reader; rather, it must be empirically verified that the model's prediction genuinely depends on the identified rationale. To this end, researchers employ techniques such as input erasure and sufficiency/necessity tests, measuring how predictions change when the selected rationale is removed or isolated. If the removal of the rationale significantly alters the output, this provides evidence that the explanation captures information that is causally relevant to the model's decision.[32, 24]

This line of research reinforces a broader methodological principle: explainability in NLP should not be evaluated solely on the basis of human intuitiveness or persuasive clarity, but also—and more importantly—on its alignment with the model's internal decision-making

mechanisms. Explanations that are coherent yet unfaithful risk creating an illusion of understanding, which may undermine technical auditing, forensic analysis, and regulatory assessment.

2.4.3 Recommender Systems

Recommender systems constitute a core component of many modern digital platforms, aiming to suggest content, products, or services that are likely to be relevant to individual users. Their operation typically relies on modeling user preferences, past interactions, and similarities across users and items in order to promote the discovery of content aligned with personal interests.[43] Such systems are widely deployed in domains including streaming services, e-commerce, online advertising, and social media, where they significantly influence user experience and decision-making processes.

In this context, interpretability plays a crucial role and is closely connected to user trust, perceived fairness, and system accountability.[44, 45] Empirical studies have shown that providing explanations for recommendations can increase user satisfaction, perceived transparency, and acceptance of suggested items.[44] Moreover, interpretability supports auditing and bias detection, enabling stakeholders to assess whether certain products, categories, or user groups are systematically favored or disadvantaged by the system.

Early research on explainable recommender systems focused primarily on matrix factorization techniques. In these models, user–item interactions are represented through latent factors, and some approaches attempted to associate these factors with human-interpretable attributes or categories.[44] Although latent representations are not inherently transparent, efforts were made to extract meaningful signals that could be communicated to users in simplified explanatory formats.

More recent work has shifted toward knowledge-aware and graph-based approaches. By integrating knowledge graphs, items and users can be connected through explicit semantic relations (e.g., actor–movie, brand–product, genre–content), allowing explanations to be constructed as relational paths or attribute-based justifications.[43, 45] For example, a recommendation can be explained through a path such as: “This movie is suggested because it shares the same director as another movie you rated highly.” These graph-based explanations tend to be more interpretable and verifiable, as they rely on explicit semantic structures rather than abstract latent dimensions.

Another important research direction concerns counterfactual explanations in recommender systems.[43] Instead of merely stating why an item was recommended, counterfactual methods clarify how the recommendation would change under alternative conditions (e.g., if certain interactions or attributes were different). This approach enhances transparency by revealing the sensitivity of the system to specific features or user behaviors, thereby supporting user agency and controllability.

The evolution of recommender architectures—particularly with the adoption of deep learning and multimodal models—has further expanded explanation strategies.[45] Modern systems may combine textual justifications, visual cues, and quantitative indicators to provide richer explanations. For instance, a recommendation might be accompanied by statements such as: “This product is recommended because it belongs to the same category and brand

as items you previously purchased," possibly supplemented with visual similarity indicators or popularity metrics. Multimodal explanations can improve usability and user engagement, especially in environments where decisions are influenced by heterogeneous signals.

Overall, existing studies indicate that no single explanation strategy is universally optimal. Content-based explanations are often intuitive and easy to understand, knowledge-graph explanations provide structured and semantically grounded justifications, counterfactual explanations enhance transparency and controllability, and multimodal explanations increase expressiveness and user engagement.[43, 45] Consequently, recent research trends emphasize adaptive and personalized explanation mechanisms that balance user-centered clarity with faithfulness to the underlying recommendation model.

2.5 Transferable Principles of Explainability Across AI Domains

Despite the methodological differences between computer vision, natural language processing, and recommender systems, several general principles emerge in the design and assessment of explainability techniques. Making these principles explicit is particularly relevant for AI forensics, where explanations must support systematic analysis and verification of model behaviour across heterogeneous use cases.[7]

A first principle is the primacy of faithfulness over form: explanations should be evaluated mainly on their ability to accurately represent the model's internal mechanisms, rather than on how intuitive or visually appealing they appear.[7, 24] Transparent but unfaithful explanations risk providing a misleading picture of how the system operates, which can mislead subsequent analysis.

A second principle concerns consistency with the structure of the data. Explanation methods must respect the properties of the underlying domain, such as spatial structure in images, sequential dependencies in text, or relational patterns in user–item interactions.[22] Techniques that ignore these properties may appear plausible while providing misleading information about how the model processes inputs.

A third principle relates to the intended audience. In forensic and audit settings, explanations are primarily addressed to investigators, reviewers, and judges rather than model developers. Consequently, explanation artefacts must be organised, documented, and contextualised in a way that supports rigorous decision-making, accountability assessments, and legal analysis.[32]

Finally, recent work highlights a trend towards hybrid architectures that combine inherently interpretable components with post-hoc explanation mechanisms.[25, 33, 34, 35] Examples include concept bottleneck models and prototype-based networks, which force the model to reason through explicit concepts or representative examples.[35, 34] These approaches aim to reconcile high predictive performance with a degree of native transparency that makes systems more amenable to scrutiny in real-world and legal contexts. From an AI forensics perspective, these properties are essential to ensure that explanations can be systematically audited, reproduced, and contested in evidential and accountability-driven settings.

Chapter 3

AI Forensics: Definition, Principles, and Relationship with Explainability

3.1 Definition and Scope of AI Forensics

AI Forensics is an emerging research area situated at the intersection of Explainable Artificial Intelligence (XAI), cybersecurity, and digital forensics.[46, 47] Its central objective is to treat artificial intelligence systems and their associated artefacts—such as trained models, logs, datasets, learned weights, configuration files, and generated outputs—as forensic objects that can be systematically examined, validated, and reconstructed. Rather than focusing solely on predictive performance, AI Forensics emphasizes the need to analyse how decisions are produced, under which conditions they are generated, and whether they can be reliably reproduced and scrutinised in investigative or judicial settings.

This scope explicitly includes foundation and generative models, such as large language models (LLMs). In these systems, prompts, retrieved contextual information, model versions, fine-tuning parameters, and decoding configurations (e.g., temperature, sampling strategy) constitute critical forensic artefacts. Preserving and documenting these elements is essential to ensure reproducibility, traceability, and contestability of generated outputs, particularly when such outputs may function as evidence-like artefacts (e.g., reports, summaries, chat transcripts).[42]

Accordingly, AI Forensics extends beyond the explanation of individual model predictions. It requires end-to-end provenance tracking, integrity verification, and systematic documentation of the entire operational pipeline—from data ingestion and preprocessing to model deployment and output generation. This comprehensive approach enables AI-generated artefacts to be assessed as reliable, testable, and potentially admissible forms of digital evidence in legal and investigative contexts.[46]

Recent contributions informed by the framework of Explainable AI for Cyber Forensics (XAI-CF) further refine this perspective.[46, 47] Within this paradigm, explainability is not treated merely as a usability feature, but as a foundational requirement for forensic reliability. XAI-CF proposes a holistic integration of transparency, verifiability, and accountability,

aiming to strengthen trust and credibility in AI systems deployed in legal, law-enforcement, and security-sensitive environments. In this view, explanations must be technically grounded, empirically testable, and aligned with established principles of digital evidence handling.

In this thesis, the XAI-CF perspective is adopted and operationalized as a structured instance of AI Forensics in which explainability is explicitly coupled with evidentiary standards. Explanations are therefore conceived not only as interpretative aids, but as forensic artefacts that must be documented, reproducible, and subject to independent audit. This alignment with evidentiary requirements ensures that AI systems can be systematically scrutinised, reconstructed, and, when necessary, contested within accountability-driven and legally regulated settings.

3.2 Core Principles

In this sense, for an artificial intelligence system to be considered a reliable and forensically valid object, the system must comply with a set of basic principles. These principles are rooted in the traditions of digital forensics but adapted to the algorithmic nature of AI; they establish a methodological framework to ensure integrity, traceability, and evidentiary value for algorithmic evidence.[46, 48]

3.2.1 Traceability and Digital Chain of Custody Applied to AI Systems

All stages in the life cycle of an artificial intelligence model, from the initial training phase to deployment and inference, must be thoroughly documented and accompanied by verifiable metadata. Such metadata includes, among other elements, the inputs provided to the system, activity logs, configuration parameters, model versions, and the outputs generated during operation.[48]

This principle establishes what can be defined as a *digital chain of custody applied to AI systems*, a mechanism that enables the transparent and verifiable reconstruction of every step and every decision made by the algorithm.[48, 49] Similar to traditional digital forensics practices, it ensures that the digital evidence associated with the system's functioning remains unaltered, thereby preserving integrity and authenticity, and supporting non-repudiation when cryptographic signatures, secure timestamping, and authenticated identities are in place.

The traceability of processes and systematic recording of evidence not only safeguard the evidential value of the system's behaviour, but also constitute an essential prerequisite for independent review, auditability, and the possibility of technical or legal contestation of automated decisions.[49, 50] Consequently, without an adequate digital chain of custody applied to AI systems, an intelligent system may be difficult to regard as reliable or suitable as a source of evidence in investigative or judicial contexts.[49]

Formal requirements. For an AI system to satisfy the principle of traceability, it must meet the following formal requirements:

1. **Identification:** every artefact (dataset, model, configuration, log, explanation) must possess a unique and persistent identifier.

2. **Provenance Registration:** all inputs, transformations, and operations must be recorded with source, timestamp, and responsible actor.

3. **Life-cycle Logging:** all stages of model operation (training, validation, deployment, inference) must produce structured logs with tamper-evident protection.

4. **Reconstruction Ability:** an independent reviewer must be able to reconstruct the chain of processing and reproduce the relevant output using stored artefacts.

3.2.2 Auditability (White-box and Black-box)

The principle of auditability refers to the ability to subject a model to an independent, systematic, and reproducible examination that verifies how it processes data and what computational steps lead to a specific output. A system can be said to be auditable when a third party can rebuild its decision-making behaviour, assess its consistency, and investigate potential biases or manipulations. This property is a foundational requirement in AI Forensics, providing transparency, accountability, and reliability in automated decision-making applications in regulatory, investigative, and judicial environments.[49, 50]

Given this background, let us consider two main types of auditing. The first type is white-box auditing, which designates a situation in which the internal structure of the model is fully available for inspection. This allows auditors to examine the architecture, parameters, training and validation data (when accessible), internal logs, and training configuration, enabling a detailed reconstruction of the model's decision process and an assessment of whether observed behaviour is consistent with the underlying algorithmic logic.[49] For this reason, white-box auditing represents the highest degree of transparency and control for inspection, since full structural access enables the discovery of structural vulnerabilities, distortions resulting from training data, or evidence of tampering.[49]

When such access is not available, as in the case of proprietary systems, commercial models, or scenarios subject to technical, legal, or security constraints, auditing must instead be carried out in black-box mode. In this context, the analysis is conducted solely through the observation of the system's inputs and outputs, reconstructing the model's behaviour based on its external responses. The absence of internal visibility requires the use of Explainable Artificial Intelligence (XAI) techniques, such as LIME, SHAP, Integrated Gradients, or Grad-CAM, which enable post-hoc explanations and allow the underlying decision logic to be inferred indirectly.[26, 27, 29, 28] Although this method does not provide a level of transparency comparable to white-box auditing, it remains essential for evaluating non-accessible systems, offering a sufficient degree of control for investigative purposes and for verifying operational correctness.

Formal requirements. The auditability principle requires that an AI system fulfils the following conditions:

1. **Access Conditions:** the scope and type of access permitted (white-box or black-box) must be explicitly defined and documented.

2. **Inspectability:** the system must allow independent verification of behaviour through internal inspection (white-box) or structured input–output probing (black-box).

3. **Repeatability:** audit procedures must be repeatable, yielding consistent results when conducted by different reviewers under identical conditions.

4. **Documentation:** all audits must produce detailed records enabling ex post control and supporting forensic analysis.

Failure to meet these requirements prevents independent verification of the system's behaviour and undermines the possibility of contesting AI-generated outputs, thereby weakening their forensic reliability and evidential value.[49, 50]

3.2.3 Stability and Fidelity Requirements for Explanations

In the context of AI Forensics, the principle of stability of explanations plays a central role because an explanatory system can be considered reliable only when it produces consistent results that are not susceptible to manipulation or random variation. Stability implies that explanations generated by an AI model remain consistent when facing small variations in input data, environmental fluctuations, or adversarial attempts to influence the system's behaviour. This requirement reflects robustness criteria discussed in studies on Explainable AI and Adversarial Machine Learning, which establish that explanations must remain stable and coherent even under minor perturbations, random noise, or targeted adversarial attacks.[7, 51, 52]

In forensic settings, this principle becomes even more significant because unstable explanations not only undermine the reliability of the model but may also compromise the evidentiary validity of the digital evidence produced by the system. An explanatory mechanism that varies substantially under equivalent operational conditions risks introducing uncertainty into technical assessments and judicial decisions, creating a situation in which the conclusions of the AI may be considered non-reproducible and therefore inadmissible. Stability is therefore a necessary condition to support accountability, independent verification, and the admissible forensic use of evidence generated by intelligent systems.

Alongside stability, fidelity represents a second essential pillar. Fidelity (faithfulness) refers to how accurately the explanation reflects the model's true internal decision logic rather than providing a misleading simplification.[7] Fidelity means that the explanation is not only plausible to a human observer but also genuinely aligned with the real algorithmic mechanisms that determine the system's behaviour. In forensic contexts, fidelity plays a critical role because inaccurate, overly simplified, or partially misleading explanations may confuse investigators, judges, and experts, hindering the correct evaluation of the system's functioning and any associated responsibilities. Only high-fidelity explanations enable the identification of systematic errors, decision-making biases, or manipulations and make it possible to accurately reconstruct the model's reasoning process. In this sense, fidelity is not only a technical requirement but also an epistemic and legal prerequisite that allows explanations to serve as credible evidence and withstand scrutiny in judicial settings.[32, 24]

Formal requirements. To satisfy this principle, an AI system must meet the following requirements:

1. **Stability Under Perturbation:** explanations must remain consistent when inputs undergo small, controlled variations, within predefined thresholds.
2. **Fidelity to Internal Logic:** explanations must accurately reflect the model's true decision process rather than surrogate or spurious patterns.
3. **Robustness Verification:** stability and fidelity must be tested periodically using established metrics and recorded for forensic examination.
4. **Sensitivity Disclosure:** known limitations or instability regions of the explanation method must be documented to inform evaluators.

Illustrative simplified example of explanation stability and fidelity. To illustrate in a concrete way how stability and fidelity can be evaluated in practice, consider a simple binary classifier operating on two numerical features (x_1, x_2). Let the model be defined as

$$f(x_1, x_2) = \sigma(3x_1 - x_2 + 0.2),$$

where σ denotes the sigmoid activation function. The decision boundary is therefore linear and directly inspectable, which makes this model ideal for analysing the behaviour of explanation methods.

Suppose that for the input point

$$x = (0.90,\ 0.10),$$

the classifier outputs the positive class.

A local surrogate-based attribution method (e.g., a sampling-based SHAP approximation with insufficient samples or an unstable surrogate fit) may assign the following contributions to the features.[27, 52]

$$\phi_{x_1}(x) = +0.82, \qquad \phi_{x_2}(x) = -0.05.$$

Now consider a minimally perturbed version of the same input:

$$x' = (0.92,\ 0.08).$$

The classifier output changes only marginally, and the predicted class remains positive. From a functional perspective, the model behaves in a stable manner around this region of the input space.

However, applying the explanation method to x' may yield substantially different attributions, such as:

$$\phi_{x_1}(x') = +0.15, \qquad \phi_{x_2}(x') = +0.10.$$

This illustrates that even when the underlying model is stable, the explanation procedure itself may introduce instability if not properly configured or validated.

Although x and x' are nearly identical and the model's decision is unchanged, the explanation varies dramatically. This reveals a lack of *explanation stability*, since a tiny perturbation of the input induces a disproportionately large change in the attribution vector. Furthermore, the explanation also lacks *fidelity*: given the true linear structure of the model, a faithful attribution should be approximately proportional to the coefficients of the logit function and should therefore consistently assign dominant importance to x_1.

This simple synthetic example highlights why stability and fidelity are essential requirements for any explanation method intended to support auditing, contestability, and forensics in AI systems.

3.2.4 Legal Compatibility and Procedural Transparency

Another critical principle relates to the legal admissibility of evidence produced by AI systems while ensuring compliance with the transparency, verifiability, and reproducibility conditions mandated by current law. For an algorithmic artefact to qualify as admissible evidence, it must be open to examination by third parties, enabling the reconstruction of the logical process involved in automated decision-making.

Where opaque or non-auditable models fail to meet this standard, the evidence may not be admissible or adequate to support a technical or legal conclusion in court. This expectation of transparency and independent verifiability in decision-making systems is mirrored in the European regulatory framework, which notably includes the General Data Protection Regulation and the European Union Artificial Intelligence Act.[20, 19] Both instruments impose obligations on the transparency, accountability, and documentation of automated decision-making processes, emphasizing the need for systems that can be independently inspected and verified.[20, 19, 53]

Formal requirements. This principle requires that an AI system satisfy the following conditions:

1. **Legal Grounding:** the purposes of use and legal basis must be explicitly documented and compliant with relevant frameworks (e.g. GDPR, AI Act).

2. **Examinability:** the system must allow third parties—investigators, judges, experts—to examine and understand the decision process.

3. **Contestability:** individuals and authorised reviewers must be able to contest outputs and obtain human review.

4. **Forensic Documentation:** the system must provide documentation enabling assessment of admissibility, including limitations, risks, error rates, and conditions for appropriate use.

Without these conditions, AI-generated artefacts may fail to meet procedural standards of admissibility and contestability, regardless of their technical performance.

Taken together, these principles define the minimum conditions under which AI systems and their outputs can be treated as forensically valid objects, suitable for audit, contestation, and evidential use.

3.3 The Connection between Explainability and AI Forensics

Explainable Artificial Intelligence (XAI) is the empirical underpinning of AI Forensics and acts as a bridge to the theoretical and applied methods for understanding the internal workings of artificial intelligence systems.[22, 7] Explanation methods such as *Local Interpretable Model-Agnostic Explanations (LIME)*, *Shapley Additive Explanations (SHAP)*, *Gradient-weighted Class Activation Mapping (Grad-CAM)*, and *Integrated Gradients* can produce understandable and interpretable representations of a model's decision-making process.[26, 27, 28, 29] These techniques may be used to document a model's behaviour and the associated outcomes of decisions, identify biases or systematic deviations in the system, determine the extent of agreement and correctness of decisions during a technical audit or review by an expert, and reconstruct the reasoning of the model in an investigative or legal context.[7]

Nonetheless, AI Forensics is an extension of explainability which provides an additional factor, the forensic verifiability of the explanations. An explanation must be more than understandable; it needs to be reproducible, verifiable, and defensible in the context of dispute, investigation, or litigation. In other words, any explanation offered by an AI system must be capable of being replicated and verified in a manner that ensures it has not been changed or manipulated.[48, 49]

Recent approaches in the domain of AI Forensics meet these requirements by incorporating advanced mechanisms, including explanation logging, model fingerprinting, and traceability graphs, as discussed in recent work on AI forensics and algorithmic accountability.[48, 54, 49] These instruments record, verify, and certify model explanations to ensure that they are kept valid and intact through all states of a system's life cycle.[48, 49]

In this view, XAI and AI Forensics are two complementary dimensions of the same framework. Explainability provides both the technical and interpretive transparency needed in order to understand how the model works internally. On the other hand, AI Forensics aims to support explanations that are verifiable, forensically grounded, and more likely to withstand legal scrutiny.

Schneider and Breitinger, as well as Mökander, emphasize that a full coupling between explainability and algorithmic auditing is a critical prerequisite for AI-based decisions to be regarded as scientifically reliable and forensically admissible digital evidence.[49, 50] In this way, the complementarity of XAI and AI Forensics provides the theoretical foundation for establishing dependable practices of AI accountability in domains characterized by high ethical, social, and legal impact.[49, 50]

3.4 Tools and Operational Practices

A practical implementation of AI Forensics requires a dedicated toolset and operational processes that translate the theoretical principles of transparency, accountability, and verifiability into concrete, repeatable practices.[48, 49, 42]

Functions of these tools include the recording, validation, as well as the preservation of every step in the artificial intelligence system's life cycle; thus, a sequence of decisions can be explained, replicated, and validated.

In operational terms, AI Forensics is founded on a structured set of mechanisms, which, as a whole, constitutes an algorithmic chain of custody (as introduced in Section 3.2.1). This chain also provides a means for every digital object, be it a model, a dataset, a log file, or an explanation, to be traceably associated with a secure record of the object's pedigree.[48]

By following this approach, it not only strengthens trust in AI systems, but it also fulfils the evidentiary requirements necessary to conform to legal and ethical norms, especially within high-stakes domains such as medicine, finance, and law.[53, 49] The tools described in this section indicate how the principles of forensic science can be incorporated within the development, execution, and management of AI technology. Model fingerprint creation and explanation logging can be considered as the first line of defence, facilitating model identification, authentication, and continued logging of the model and decision-making process. These facilitate the independent validation of every version of a model as well as tracing every explanation back to an unalterable source. In a complementary fashion, traceability graphs and audit trails offer a broader view based on tracing the logical, procedural, as well as causal relations underlying the AI system's behaviour. These tools, combined, make abstract ideas such as explainability and accountability verifiable, quantifiable, and legally defensible.[48, 49, 55]

3.4.1 Model Fingerprinting and Explanation Logging

AI Forensics can be conceived as a systematic approach aimed at converting internal processes of artificial intelligence systems into traceable, verifiable, and legally defensible forms of digital evidence.[49, 50] Within this framework, AI systems, datasets, event logs, and even results can be considered forensic artefacts, elements which, if well-documented, can be validated as true representations of an algorithmic decision-making process.

Because this approach promotes transparency while enabling reproducibility and auditability, it is particularly valuable in forensic, regulatory, and scientific contexts. A core element in this approach is model fingerprinting. It involves generating unique cryptographic or structural signatures for AI models and their variants that permit distinct identification, authentication, and integrity verification throughout the model's life cycle. These fingerprints are considered a kind of digital identity of the AI model.[54, 56]

Fingerprinting, from a technical standpoint, may involve hashing model parameters, encoding architectural characteristics, or embedding cryptographic watermarks within network weights. The goal is to enable the researchers, auditors, and model owners to verify that a given model has not been altered, substituted, or tampered with.[54]

For example, the *DeepMarks* framework proposed by Chen et al. introduced a method to embed digital fingerprints directly into the parameters of deep neural networks, which enabled the verification of ownership and detection of misuse.[54] More recent work on causal fingerprints for generative models similarly explored the possibility of embedding invisible yet verifiable marks into AI systems to prevent tampering and enable forensic tracing.[56] These approaches collectively allow for forensic attribution, linking a model to its outputs and enabling the identification of which entity is responsible for generating a given decision or artefact. Model fingerprinting plays a number of critical roles in forensic or auditing settings:

- **Authenticity:** This ensures that the model used to generate a decision is indeed the one declared or certified.
- **Integrity:** It ensures that the parameters, architecture and configuration of the model have not been changed since deployment.
- **Traceability:** It connects the model to its training data, version history, and inference conditions, creating a transparent record of its evolution over time.

Such guarantees are fundamental when AI-generated results serve as technical or legal evidence: any uncertainty about the model's integrity would undermine the evidential validity of the output.[49, 50]

In addition to fingerprinting, explanation logging represents the procedural counterpart that documents and verifies the reasoning behind an AI model's decisions.[48, 49, 55]

This involves systematically recording the model's explanations — such as feature attributions, saliency maps, attention weights, or textual rationales — together with contextual metadata including timestamps, dataset versions, model identifiers, inference details, and operational environments. Thus, every AI-generated decision is accompanied by a comprehensive record of the conditions under which it was produced, along with its corresponding interpretability artefacts.

Explanation logging not only improves interpretability but also creates an evidence-based audit trail, whereby independent reviewers can trace and verify how a model arrived at any given output. This is paramount in highly regulated industries such as healthcare, finance, and legal, helping organisations maintain accountability, compliance, and legal defensibility in regulated domains.[53, 42] Other approaches combine explainability techniques with cryptographically secure and tamper-evident logging mechanisms to protect the evidential integrity of AI decisions.[48, 55] By doing so, these mechanisms make records tamper-evident and substantially harder to manipulate retroactively, safeguarding not only the content of explanations but their temporal authenticity as well.

Taken together, model fingerprinting and explanation logging set up a dual foundation for AI accountability and forensic verifiability. Fingerprinting secures the model as an identifiable, immutable artefact, while explanation logging preserves the context and reasoning behind each decision. Put together, they become a comprehensive algorithmic chain of custody that ensures traceability, verifiability, and defensibility of every AI-generated output with scientific and legal rigor.[48, 49]

3.4.2 Traceability Graphs and Audit Trails

This section operationalises the traceability principle introduced in Section 3.2.1 by describing concrete artefacts—traceability graphs and audit trails—used to represent and preserve provenance and decision histories.

Traceability graphs and audit trails are basic tools to guarantee AI systems' transparency, verifiability, and reproducibility of decision processes. Their purpose is to render every stage in the life cycle of a model, from data collection to decision generation, analysable, verifiable, and, when needed, reproducible by independent parties.[48, 49]

The relationships between the various parts of the AI system, data, model elements, training stages, inferences, and generated explanations are represented visually and structurally by traceability graphs. The causal chain that links a particular input to its corresponding output should be reconstructed from these representations, emphasizing the operational and logical dependencies in algorithmic decision-making processes.[48]

Audit trails perform a complementary documentary function in that they record every significant event in the model's life cycle, which includes data preprocessing, training, inference, and explanation generation. Every event is timestamped and coupled with metadata like dataset versions, model configurations, and inference parameters that enable precise verification of when and how a certain decision was produced. In this way, a true chain of custody is created where auditors or investigators can trace back, from the final output to the initial conditions of the input, ensuring traceability and accountability throughout the system's entire operation.[49, 50]

Recent research has pointed out that explainability, logging, and auditing are crucial for making AI systems reliable in regulated or high-impact contexts, such as medical, financial, or legal domains.[53, 42, 55] Similarly, studies in the field of digital forensics emphasize that even the most complex models, such as deep neural networks, must integrate logging and traceability mechanisms to make results verifiable and defensible as technical evidence in legal contexts.[49, 50]

Some crucial elements that form effective traceability graphs and audit trails include the following:

- **Data provenance:** This is the process of documenting the origin, successive transformations, and usage of the data.
- **Model fingerprinting and versioning:** Either structural hashes or cryptographic signatures should be able to distinguish between different iterations of a model.
- **Inference and explanation recording:** Every decision should be followed by the respective explanation, along with metadata that describes its context.
- **Relational graph structure:** Connections explain causal or procedural relationships, while nodes represent entities like datasets, models, or explanations.
- **Immutability and security of logs:** Secure timestamps, cryptographic hashing, access control mechanisms, and controlled evidence preservation procedures can be used to make records tamper-evident and reduce the risk of undetected modifications.[55, 57]

Chapter 4

AI Robustness and Adversarial Machine Learning

4.1 Introduction and Definitions

In AI Forensics, a core challenge is ensuring that AI outputs are not only explainable but also reliable. Robustness and Adversarial Machine Learning are central because they highlight the limitations of learning models and their susceptibility to malicious or subtle input manipulations.[58]

Robustness can be defined as the model's capability to perform well despite adversarial or unexpected changes in the input data. In a forensic context, the significance of this feature is central for judging the validity, verifiability, and scientific defensibility of model results.[59]

A model that can be reliably misled by adversarial inputs or corrupted data may be unsuitable as a source of evidence in investigative or judicial contexts. In other words, a good system would function properly despite the presence of noise, gaps in information, as well as malicious interference, and not just in a noise-free environment. This is an important property because, in a real-world setting, scenarios are rarely as controlled as those in a training environment, as input data could have some level of error, bias, or fluctuations based on the environment. A good model, therefore, should be able to absorb disturbances, cope with distribution shifts, and have consistent decision behaviour, especially under stressed conditions.[59]

In forensic deployments, this distinction is critical because evidential conclusions may depend on model behaviour under non-ideal acquisition and operational conditions.[49] On the other hand, the research area of Adversarial Machine Learning[58] focuses on both the attackability of machine learning systems as well as how to defend against those attacks. The primary aim is to grasp how it is possible for intelligent systems to be fooled using small perturbations of input data, so-called adversarial perturbations.[58, 60]

These changes can be imperceptible or barely noticeable to the human eye for sufficiently small perturbation budgets, yet they can significantly impact the model's output. In deep learning, for instance, small changes to an image's pixels can result in substantially incorrect classifications.[60]

The same has also been observed in natural language processing, where small lexical or

syntactic edits (e.g., synonym substitutions or punctuation changes) can affect downstream predictions.[58, 61] In this field, it has become an ongoing arms race between those trying to attack, often using more powerful and complex adversarial attacks, versus those attempting to defend, building models and systems to be more robust against attacks.[58, 62] These include defensive techniques such as adversarial training on clean and perturbed examples and input sanitization, as well as approaches to certified robustness; however, the literature also warns that techniques like gradient masking may provide a false sense of security if used as a robustness proxy.[60, 62, 58]

In sum, robustness and adversarial machine learning help characterise the conditions under which AI outputs remain reliable under perturbations and threat models relevant to forensic use. They form a basis for evaluating the applicability of AI systems as a credible part of a process of investigation or evidence. Without robustness assessment, explanations risk being attached to decision processes that are not reproducible under the same threat model, undermining contestability and forensic replication.[7, 55]

In the context of AI Forensics as a whole, robustness can be considered a key prerequisite, with direct implications for the authenticity, reproducibility, and verifiability of AI-driven decisions, their rationales, and digital trails.

From an AI Forensics perspective, robustness assessments should be paired with traceable experimental protocols and systematic logging of adversarial parameters (e.g., threat model, attack type, perturbation budget ϵ, and preprocessing steps), so that robustness claims are reproducible, verifiable, and open to forensic contestation, in line with the principles outlined in Chapter 3.[49, 48, 55]

Formal requirements (forensic robustness evaluation). A robustness claim intended for forensic use should satisfy, at minimum, the following conditions:

1. **Threat Model Specification:** the attacker's capabilities and objectives (e.g. evasion vs. poisoning; white-box vs. black-box) must be explicitly defined.
2. **Controlled Perturbation Budget:** the perturbation constraint (e.g. norm and ϵ) and all preprocessing steps must be fixed and reported.
3. **Reproducible Protocol:** the evaluation procedure (attacks, random seeds, datasets, metrics) must be replicable by an independent reviewer.
4. **Tamper-evident Logging:** all relevant artefacts (inputs, adversarial parameters, outputs, explanations) must be recorded with integrity safeguards.[55, 57]
5. **Metric Disclosure:** the robustness indicators and reporting format must be specified (e.g. attack success rate, robust accuracy, false accept/reject impact for verification), including confidence intervals where applicable.

Illustrative simplified adversarial example for face verification. To illustrate concretely how adversarial perturbations can compromise a face verification system, consider a typical pipeline in which a deep neural network (e.g., FaceNet)[63] encodes each preprocessed face into an embedding vector $z \in \mathbb{R}^{128}$.

For simplicity, this illustrative example adopts 128-dimensional embeddings and Euclidean distance; other realistic deployments may instead rely on higher-dimensional embeddings and alternative similarity measures.

Two images are classified as belonging to the same identity if the Euclidean distance between their embeddings is below a threshold τ.

Suppose that two clean images of the same subject produce the distance:

$$d(z_A, z_B) = 0.75 \qquad \text{with} \qquad \tau = 1.00,$$

leading the system to output a correct "same identity" decision.

Now apply a small adversarial perturbation to image A using the Fast Gradient Sign Method (FGSM), defined as:[60]

$$x^{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x L),$$

with $\epsilon = 0.01$.

Here, L denotes a verification-oriented objective (e.g., a contrastive or triplet loss) designed to increase the distance between paired embeddings.

Although such a perturbation is often imperceptible to a human observer (assuming inputs are normalised to the $[0, 1]$ range), it may significantly alter the internal representation of the system.

In a purely illustrative (non-empirical) numeric example, the embedding may shift as follows:

$$\|z_A^{\text{adv}} - z_A\|_2 \approx 0.68.$$

The verification distance then becomes:

$$d(z_A^{\text{adv}}, z_B) = 1.42 > 1.00,$$

leading the system to classify two visually indistinguishable images as "different identity".

This example highlights how a face recognition model can be manipulated by imperceptible input perturbations, which induce large semantic distortions in the embedding space and ultimately reverse the verification outcome. From a forensic perspective, such vulnerabilities undermine the evidential reliability of face verification systems and motivate the need for robustness testing, adversarial auditing, and explanation-based diagnostics. Operationally, robustness evaluation in forensic-oriented settings can be framed as a controlled test protocol including (i) benign perturbations (e.g., noise, compression, illumination shifts), (ii) standard adversarial attacks (e.g., FGSM, Projected Gradient Descent (PGD)), and (iii) stability checks on explanations (e.g., attribution similarity under perturbations), with all parameters logged to ensure reproducibility.[60, 62, 7, 55, 57]

4.2 Categories of Adversarial Attacks

Adversarial Machine Learning (AML) research has identified different categories of attacks that take advantage of the inherent vulnerabilities of learning algorithms.[58] Each category of attack occurs at a different stage of the model's life cycle and has different malicious goals, for instance, to mislead the predictions of the model, or to breach the confidentiality or integrity of the model's training data. It is important to understand these categories for the forensic analysis because it provides a conceptual framework for identifying, documenting, and addressing whatever manipulations could undermine the evidential value of an AI system.

From a forensic standpoint, this taxonomy also informs what artefacts should be preserved and logged (e.g., dataset versions for poisoning, trigger indicators for backdoors, and query traces for extraction), enabling later attribution and contestability.[58, 49]

4.2.1 Evasion Attacks

Evasion attacks are among the most widely studied and well-understood types of adversarial manipulation, and they occur in the inference phase of a trained and deployed model, when an attacker modifies each observation of their input data to make it imperceptible to a human but misleading to the model.[64, 60] For instance, a few pixel-level changes can cause a vision model to interpret a stop sign as a speed-limit sign, or a malware detection model to classify harmful software as benign.[60] Attacks of this nature are typically caused by the model's sensitivity to high-dimensional feature spaces, which shows us that even small perturbations can push an input across a decision boundary.[58]

In forensic settings, evasion attacks are particularly harmful because they produce no visible artefact, the data is authentic, but the model output has been purposefully altered in an undesirable manner. Improved explainability tools, inference comparison testing, and consistency testing that can characterize model behaviour as deviating from expectations are necessary for identifying these kinds of attacks.[7, 22]

4.2.2 Poisoning Attacks

Poisoning attacks manifest earlier in the AI pipeline than other attacks, during the training phase. In this case, the adversary adds compromised or mislabeled samples into the training dataset, which leads to the corruption of the learning process.[65, 58] The intent is to introduce small biases or create scenarios that result in the trained model outputting incorrect predictions in specific conditions, while continuing to output correctly in all other cases. An example of this is a facial recognition system that is "poisoned" such that certain individuals or demographic groups are misidentified, thus embedding bias or backdoor triggers into its learned parameters.

Identifying whether poisoning has occurred is often non-trivial. Because the manipulation exists in the training dataset itself, it can remain present even after the model appears to pass standard validation checks. From a forensic perspective, formulating the approach poses a risk to the trustworthiness of evidence used to allow the model to act, as it compromises the chain of possession regarding the data that is used to develop the model.[58]

4.2.3 Backdoor Attacks

A backdoor attack is a more intricate method of poisoning the model, through the insertion of a concealed trigger that, whenever the trigger is present, such as a specific pattern, symbol, or sequence, forces the model to generate an attacker's desired output.[66, 67] When outside of the trigger condition, the model behaves as expected, making this attack particularly difficult to catch with evaluation metrics. As an example, one might consider a facial recognition network that misclassifies an individual when it detects the individual wearing a specific accessory (like patterned scarf or a pair of glasses) in the image.[66] In forensic and legal settings, and when being used for safety-critical tasks, backdoor attacks may prove the most dangerous, as an attacker is able to selectively manipulate the AI outputs of the system without impacting on the overall performance of the model, and will not need to introduce any identifiable changes to the model that may be apparent during detection.[67]

4.2.4 Model Inversion and Extraction Attacks

Apart from directly dealing with data, attackers can also make use of the inherent vulnerabilities of AI models in two distinct categories of attacks, namely model inversion attacks and model extraction attacks.

In inversion attacks, the attacker uses the model's responses to draw conclusions about the data that was used to train the model. The attack entails the attacker taking advantage of the model in a way that makes it possible to retrieve data, including personal characteristics, from datasets that appear to be anonymous.[68]

In contrast, in the case of extraction attacks, the attacker launches a sequence of carefully crafted queries, motivated by extracting the internal model parameters or duplicating the model's decision-making logic. The attacker can steal valuable, proprietary, as well as secret information.[69]

Both attacks pose a threat to data privacy as well as ownership of the AI model, making it hard to decide which is more at stake, data privacy or security within AI systems. It is particularly important in the forensic environment because it is vital to have traceability, integrity, and verifiability of the digital evidence being generated or processed within AI systems.[58]

In a forensic context, it is absolutely essential to have a clear appreciation of the significance of the aforementioned attack vectors from a forensic perspective.

Every attack occurs on a distinct system layer:

- Backdoor attacks use covert control mechanisms;
- Inversion attacks compromise the confidentiality of the protected data;
- Poisoning attacks disrupt the training process;
- Evasion attacks impact model predictions.

Thus, a thorough forensic assessment will require a consideration of the whole model life cycle, from data acquisition to deployment, looking for points of potential weakness along

the way, in relation to both training as well as inference.[58] In operational practice, it is necessary for the forensic analyst to have robustness testing, adversarial simulation, and mechanisms for model monitoring as part of their investigative toolset. In this way, they would be able to make a distinction between maliciously introduced system flaws versus benign ones introduced inadvertently as a result of an attack.

Robustness, as well as interpretability, have been some of the most important, as well as challenging, topics in AI research, particularly over the past years.[7, 9] At a high level, these two dimensions may appear conceptually distinct. The first concept revolves completely in the realm of reliability, security, and robustness of model behaviors within perturbed environments, while the latter focuses on human comprehensibility of the decision-making processes of a model. In many forensic settings, it is difficult to treat the two dimensions independently. If a model is not robust, explanations may have limited forensic value, because small perturbations can change both predictions and explanatory artefacts.[7]

Conceptually, robustness can support the forensic usefulness of explanations by ensuring that both predictions and explanatory artefacts remain stable under small perturbations. If a small variation in the data leads to radically differing accounts, then the account-explanatory framework fails, since a considered choice becomes arbitrary in slightly differing circumstances.

This uncertainty has been noticed in many post-hoc attribution techniques. For example, the LIME approach as well as SHAP can result in very different values being associated with the input features, even in slightly differing inputs.[26, 27, 7, 70] On the other hand, explanation can also be a means to achieve robustness, as it can indicate sensitivities due to model representation flaws. By analyzing the explanation, it becomes possible for the examiner to identify whether the model is based on a spurious correlation, as well as irrelevant predictors, both of which imply a lack of strong generalization.[7] Thus, it is important to consider explainability and robustness as complementary forces working in tandem. Each can strengthen, inform, and impact the other.[9]

4.2.5 Adversarial Explainability

The point where robustness and explainability meet has generated a novel research area, which has been referred to as adversarial explainability. In this line of research, the notion being examined is how AI models can be attacked, as well as their explanations. In relation to this, the explanation, which has always been utilized as a means for transparency, can actually be a target for manipulation, thereby creating a deceptive explanation for the underlying reasoning process of the model.[51, 71, 70]

In adversarial explainability, the attacker's behaviour consists of slightly modifying the input data in a way such that the model's output gets altered, whereas the explanation does not seem to change. For instance, an indistinguishable small modification to the image may result in the wrong classification, while a Grad-CAM heat map would still emphasize meaningful areas.[51, 71] Thus, these attacks show that explanation can be fallible, misleading, or corrupted just like AI systems.

In fact, this creates a serious challenge for the use of AI as a means of analyzing evidence, as the explanation could be incorporated as a form of evidence. Moreover, being able to

falsify an explanation means the chain of reasoning within the evidence can be corrupted. In a forensic context, the challenge of adversarial explainability questions the trustworthiness of interpretability methods.[7]

A compelling explanation from a visual perspective but not from a semantic perspective can mislead an auditor's perception of model-driven behaviour, yielding an inaccurate assessment about a model's decision. This becomes more problematic in areas like law enforcement, medicine, or finance, whose decision-making processes involve artefacts like saliency maps, attention weights, or feature attributions, which could be considered as a form of proof.[32, 24]

In order to counter such challenges, several strong frameworks have been proposed within the research community, whose objective is to make sure that the generated explanation is not only interpretable but stable, reproducible, as well as robust to perturbation.[72, 33] These often include four distinct strategies as follows:

- **Explanation Regularization:** The addition of stability-focused loss functions during training, which penalize large variations in the explanation across nearby regions of the input.[72] This ensures that small controlled perturbations in the input do not cause disproportionate variations in the corresponding interpretive maps.
- **Ensemble-based Explanation Strategies:** Rather than focusing on a single technique for interpretability, a set of techniques such as LIME, SHAP, Integrated Gradients, and Grad-CAM can be utilized to generate a combined explanation.[26, 27, 29, 28] It is possible to detect irregularities of interpretation, potentially indicative of manipulation, based on the level of agreement between different validation techniques.
- **Stability and Similarity Metrics:** Other metrics include explanation consistency indices, cosine similarity between attribution vectors, or structural similarity indexes for saliency maps, which are measured by controlled noise or adversarial attacks to show robustness.[51, 71] These provide empirical evidence for reliability in interpretation.
- **Adversarially Aware Training:** In some architectures, perturbations generated from adversarial processes are incorporated to optimize both prediction accuracy and explanation constancy.[62, 33] It is a dual defence mechanism, assuring that it resists both adversarial misclassifications as well as preserves stable explanatory behaviors under input variations.

In this way, research on adversarial explainability tackles the challenge of creating forensically provable explanations, which can preserve evidentiary credibility despite being subject to attempts at manipulation. What this means, in the context of AI Forensics, is that explanations have to be subject to the same level of scrutiny as model predictions, being traceable, reproducible, and falsifiable in line with scientific and legal standards. For an explanation to be credible, it should be resistant to manipulation and accompanied by robustness and stability assessments. A key requirement is stability under perturbation, often complemented by logging, reproducibility controls, and cross-method agreement checks.[7]

4.2.6 Forensic Significance and Evidential Implications

In the larger context of AI Forensics, the relationship between robustness and explainability has direct bearing on evidential reliability and auditability of intelligent systems. For example, although a model may be interpretable, if the explanations are inconsistent and/or subject to manipulation, any evaluation is moot and could even be perceived as misleading transparency.[9, 7] On the other hand, while a robust model that is not interpretable poses limited forensic value because a justification of its decisions cannot be meaningfully evaluated.[6]

Thus, forensic auditors and investigators are required to evaluate the two aspects in combination. A system's robustness metrics (e.g. adversarial resistance and stability to noise) can be evaluated but should also consider consistency in its explanation (evaluated by its fidelity, locality, and (human) interpretability tests).[32, 24] Thus, only by consideration of both frames would we be able to determine whether the AI system meets key forensic requirements that support scientific credibility and may facilitate admissibility in investigative or judicial settings.

4.3 Metrics and Integrated Forensic Frameworks

The intersection of explainability and robustness has led to the development of integrated forensic frameworks aimed at ensuring transparency, trustworthiness, and protection throughout the AI system life cycle.[49, 48] Within such frameworks, explainability methods are not limited to illustrating how a model operates; they also function as diagnostic tools for assessing robustness against adversarial manipulation. Conversely, robustness analysis contributes to evaluating whether explanations themselves remain stable, faithful, and reproducible under controlled perturbations. This bidirectional relationship is central to forensic reliability.

Existing governance and risk-management frameworks provide structured foundations for operationalising this integration. For example, the *NIST AI Risk Management Framework (AI RMF)* explicitly identifies validity, reliability, safety, security, resilience, and explainability as core components of trustworthy AI systems.[42] Within this framework, continuous measurement, documentation, and monitoring are required to ensure that system outputs remain reliable under evolving operational and threat conditions.

Similarly, the European Union *Artificial Intelligence Act* establishes obligations for high-risk AI systems, including risk management, technical documentation, automatic logging of operations, transparency requirements, and human oversight.[19] These regulatory requirements directly align with forensic principles such as traceability, auditability, and evidentiary reproducibility. The *General Data Protection Regulation (GDPR)* further reinforces these obligations by requiring accountability, documentation, and meaningful information about automated decision-making processes.[20]

From a forensic perspective, integrated frameworks combine:

- **Robustness Metrics:** including robust accuracy, attack success rate, certified robustness bounds, and stability of outputs under controlled perturbations.[62, 58]

- **Explanation Metrics:** such as fidelity, stability under perturbation, attribution similarity indices, and cross-method agreement measures.[7, 51]
- **Procedural Controls:** structured logging, model fingerprinting, version control, and tamper-evident audit trails to ensure reproducibility and traceability.[48, 55]

In operational deployments, forensic-oriented AI systems should therefore incorporate continuous auditing of explanations and periodic robustness validation as part of their life cycle management. Such systems establish a *chain of trust from data to decision*, ensuring not only that predictions are accurate, but that both predictions and their explanations remain stable, reproducible, and defensible under defined threat models.[49, 48]

Ultimately, integrated forensic frameworks transform explainability and robustness from isolated research objectives into measurable, documentable, and legally defensible properties of AI systems, thereby supporting scientific credibility and potential evidentiary admissibility in investigative and judicial contexts.

4.4 Challenges and Future Directions

Among the most important questions for current research in the fast-evolving domain of AI Forensics is how robustness and explainability can be fruitfully combined within one coherent and consistent methodological framework.[49, 50] These two dimensions, which are often considered as goals, should be treated not as alternatives, but rather as complementary parts of a single foundational principle: the pursuit of verifiable trust in artificial intelligence systems.

Explainability allows a model's decisions to be understood and justified, therefore providing insights into its inner reasoning, while robustness aims to ensure that such decisions are stable, reliable, and resistant to perturbations or manipulations.[9, 7]

It is only by integrating both that AI systems can become not merely transparent at the level of communication but scientifically defensible on technical grounds and more likely to withstand forensic and legal scrutiny in investigative and judicial contexts. The synergy between both aspects will be particularly crucial in digital investigations and judicial applications.[49]

An AI system is more likely to be regarded as trustworthy when its explanations are coherent, verifiable, and resistant to adversarial manipulation, its results are reproducible, and its outcomes traceable over time.[48, 49, 55]

In this respect, future research should shift the current descriptive paradigm of explainability, which is ultimately limited to making model behaviour "visible", toward a certificatory paradigm whereby each explanation is empirically validated, independently audited, and forensically verified throughout the model's life cycle.[32, 24]

The main directions of progress that can be envisioned for the coming years can be put under the heading of three key areas:

- **Self-monitoring and traceable AI Forensics systems.** One important research line is the development of architectures which are able to self-check and verify themselves. These systems need to incorporate mechanisms that will enable continuous

monitoring of explanatory coherence and predictive robustness through an algorithmic chain of custody that automates the documentation at each step of decision-making, right from data acquisition to the generation of output.[49, 48]

- **Establishment of unified standards and evaluation metrics.** The lack of common criteria for assessing robustness and explainability now limits the capability to certify the reliability of AI models. Therefore, it is expected that future work will focus on the development of uniform metrics, recognized internationally, which integrate technical indicators of stability, consistency, and fidelity with legal and ethical dimensions about verifiability and accountability.[9, 53]

- **Algorithmic governance and accountability.** Forensic requirements for AI systems should be explicitly integrated into the design and operational levels of governance frameworks. This would include the development of operational and regulatory guidelines that define procedures for traceability, documentation, and validation of algorithmic processes, making sure that transparency is not just an ethical aspiration, but also a structural and legal one.[53, 19]

Chapter 5

Forensic Significance and Operational Context

5.1 Overview

The intersection of robustness and explainability has significant implications for how AI can be utilized and trusted in forensic investigations.[49, 50] In digital forensics, traditional practices have relied upon a transparent and auditable chain of custody to establish the integrity of the evidence.[48, 55, 57] Analogous to these practices, in AI Forensics, robustness helps ensure that AI model outputs are not easily manipulated, whereas explainability provides investigators with the rationale for how those answers were determined.[7, 22] Together, these properties support what can be *operationally framed* as an *algorithmic chain of evidence*, i.e., a documented link between (i) the inputs and system configuration used at inference time, (ii) the resulting outputs, and (iii) the associated explanatory artefacts, so that the full decision process can be independently reconstructed and contested.[49, 48, 50]

5.2 Ensuring Evidential Integrity

The following criteria summarize recurring requirements discussed in forensic AI auditing and digital evidence governance.[49, 50, 48]

- **Traceability** entails that every step in the model's life cycle (data capture, training, validation, and inference) is documented and capable of being reproduced.
- **Consistency** concerns the ability of the model to produce reproducible results under comparable conditions, which is closely related to robustness.
- **Verifiability** is the capacity to reproduce the same explanations and the same decisions made by the model when led by an independent expert.

Together, these criteria support the probative value of algorithmic outputs by ensuring that they can be reconstructed, independently validated, and meaningfully contested. If any of these are not met, the legitimacy of an AI system as a forensic tool is called into

question.[49, 48] In consequence, explainability and robustness are to be treated as essential evidential properties or safeguards, not optional features, of the model in an AI-driven investigation.[7, 9]

5.2.1 Model Auditability – Operational Criteria

Auditability is one of the fundamental requirements in the forensic evaluation of artificial intelligence systems.[49, 50] An AI model is considered auditable when its operations, results, and the logical process that generated them can be reproduced, examined, and justified according to an objective and transparent criterion. In practical terms, this means that all steps involved in the model decision-making process, from data collection and preparation to training, validation, inference, and output generation, must be traceable, verifiable, and interpretable by independent reviewers or experts.[48, 49, 55, 57]

Therefore, auditability provides the link between technical validation and legal reliability in a forensic context.

Auditing not only verifies the correctness of a result, but also enables the reconstruction of the logical and computational path that led to a specific decision, which is a necessary condition for ensuring evidentiary accountability in AI-based investigations.[48]

To operationalize this concept, it is useful to consider three interconnected types of auditability measures that give meaning to the notion of evidential reliability in AI systems:[9, 32]

- **Robustness metrics:** quantify how well the model behaves consistently, producing consistent outputs even when perturbations, noise, or other adverse conditions are present. These measures ensure that model decisions are not based on brittle or easily manipulated reasoning.[58, 59]
- **Explanation stability metrics:** measure the consistency of interpretive outputs, such as feature attributions, saliency maps, or decision motivations, under repeated runs or variations in input data. Stable explanations indicate that the model's reasoning is not arbitrary and can withstand independent verification over time.[51, 71, 70]
- **System-level auditability indicators:** provide an overall view of traceability, transparency, and compliance with procedural norms. These refer to the conditions that enable the system to keep records of all operations, version control, and documentation at a level sufficient to allow independent forensic review.[48, 42, 53, 55]

5.2.1.1 Robustness Metrics

Robustness metrics quantify how well a model maintains performance under stress or perturbation, whether by accident (e.g., noise, acquisition errors) or intention (adversarial input). From a forensic perspective, they help in the verification of whether outputs remain reliable within realistic input variations.[58, 59]

- **Accuracy under perturbation (ϵ-bounded):** quantifies the degradation of a model's predictive performance when the inputs are systematically modified within a

controlled perturbation budget ϵ.[62, 59] In practice, one considers the accuracy on clean data, $\epsilon = 0$, compared to the accuracy that one gets for $\epsilon > 0$. Ideally, this is visualized by a curve showing how accuracy changes as a function of ϵ, with a detailed description of the perturbation type. This allows one to pinpoint the values of ϵ from which the performance of the model seriously deteriorates.

- **Attack success rate:** gives the percentage of adversarial attempts that successfully change the model's output under a specified threat model, attack type, constraints, and perturbation budget.[58, 60] Higher values indicate lower resistance to manipulation; reporting the details of the scenario explains operational risk.
- **Certified robustness bound (radius):** refers to a formally guaranteed perturbation magnitude within which the model's prediction remains unchanged.[59] Unlike purely empirical checks, certified guarantees can strengthen technical defensibility *when their assumptions (e.g., threat model and perturbation class) match the forensic setting under evaluation.*

5.2.1.2 Explanation Metrics

In the context of forensic auditability, explanation metrics assess the stability and reliability of a model's interpretations as inputs or internal states vary, with the aim of verifying that what is presented as the algorithm's reasoning is coherent, reproducible, and reflective of its actual behaviour.[9, 32, 24]

- **Stability under perturbation:** measures how similar feature attributions (e.g., LIME/SHAP values or saliency maps) remain when the same input is subjected to minimal, controlled modifications. Low variability indicates that the explanation does not depend on incidental details and is therefore more credible in expert review.[26, 27, 7]
- **Fidelity:** quantifies the degree to which the explanation aligns with the model's true decision behaviour (e.g., local decision boundaries), so that high fidelity supports technical validity and evidentiary usefulness.[24, 7]
- **Infidelity:** captures the divergence between the model's predicted output and the output implied by the explanation; high values suggest that the interpretation mischaracterizes the decision logic and risks misleading conclusions.[51, 71, 70]

Taken together, stability, fidelity, and infidelity provide a triangulated view of explanation quality: stability attests to consistency under small perturbations, fidelity to correspondence with decision behaviour, and infidelity highlights systematic discrepancies, enabling more verifiable and reproducible judgments about explanation reliability.[32, 24]

5.2.1.3 Operational Audit Indicators

There is no single performance measure through which to reduce an effective forensic audit; it needs to span the entire model pipeline and address each step with traceability, reproducibility, and long-term control.[49, 48] The aim is to turn auditability from an abstract property

into a concrete practice, so that every AI-derived decision, together with its rationale, can be scientifically verified and defended in legal settings.

Terminology varies across standards; in this thesis, we use *replicability* to denote rerunning the same pipeline under the same code, data, and configuration, while *reproducibility* refers to independent reruns under equivalent conditions, potentially with different implementations.[42]

Replicability means getting the same results with the same seed, configuration, and code and data versions. In practice, this means keeping track of random seeds and documenting pseudo-random generators in use, versioning code and experimental notebooks, versioning data and artefacts (raw and processed datasets, feature stores, and trained models) with explicit links between data versions, training configurations, and results, and capturing the execution environment (library versions, containers, drivers, operating system, hardware).[42] A replicability report should contain, beyond metrics and plots, an executable recipe with configuration files, commit and data-release references, seeds, and hardware that would allow an independent expert to rerun the experiment and get the same result within known margins. This reduces ambiguity and makes outputs more probative.[49, 50]

Provenance concerns the integrity and origin of data and models. Each critical object (datasets, feature sets, source code, model weights, reports) should be associated with immutable identifiers, such as cryptographic hashes, and, when appropriate, digital signatures attesting authenticity and absence of tampering.[48, 42, 50] Version control should enable reconstruction of data lineage: sources, transformations, parameters, and the code version that produced them. For models, documentation should include training and validation datasets, hyperparameters, optimization criteria, checkpoints, and selection metrics. Structured artefacts, such as model cards and datasheets for datasets, covering their purposes, limits, covered populations, known risks, complete the chain of custody, facilitate accountability, and support integrity checks over time.[73, 74]

Drift monitoring concerns the fact that even a valid system can degrade over time due to distribution shift, concept drift, or changes in class prevalence. Continuous monitoring needs to include statistical tests on inputs and representations to detect distributional change; performance surveillance on labeled data or reliable proxies; alert thresholds with response plans, including controlled retraining and rollback.[42, 75] Monitoring should result in versioned logs and reports recording when drift was detected, the supporting evidence, the decisions taken, and their impact on metrics, ensuring probative consistency over time.[48, 42, 55]

From a robustness standpoint, drift monitoring operationalizes robustness to distribution shift by detecting when the deployment environment diverges from the assumptions under which robustness claims were established.[59, 42]

Replicability, provenance, and drift monitoring define an operational framework for the forensic audit of AI systems. Replicability ensures that results are not contingent on unrepeatable conditions; provenance ensures that each artefact is authentic, traceable, and attributable; drift monitoring safeguards the stability of evidence over time by detecting and correcting deviations that could undermine validity.[49, 50] Together, these practices move from generic promises of transparency to repeatable and verifiable procedures in which every prediction is accompanied by context, rationale, and proof of integrity, enabling reconstruc-

tion of the decision chain and supporting the reliability of digital evidence according to clear and reproducible technical criteria.[53, 19]

5.3 Integrating Explainability and Robustness in Forensic Workflows

The integration of explainability and robustness is an essential requirement for the forensic auditability of artificial intelligence systems.[49, 48] It is not sufficient to report a prediction and an explanation; forensic use requires evidence that both remain valid under documented operating conditions and predefined perturbation bounds. It must be verified that the explanation is consistent with the model's actual behaviour and that it retains its validity under controlled variations of the input or internal states.[32, 24] From this perspective, the forensic workflow should combine explanation tools with robustness tests, so that what the model decides, why it decides it, and whether that rationale remains stable when conditions change within predefined limits are assessed simultaneously.[7, 58]

The process starts by defining the use context and the threat model, and specifying which perturbations are realistic or admissible for the application at hand.[42] It is based on this that a baseline of explainability is established on unperturbed data, selecting methods appropriate to the model and the domain. For vision models, techniques such as Grad-CAM can be used to highlight the image regions that support the decision; for tabular or textual models, attribution methods such as SHAP can quantify the contribution of individual features.[28, 27] All the parameters of the various tools employed must be fixed and documented, so as to enable repeatable comparisons.[32]

That is followed by a stress phase where controlled perturbations are applied to the same inputs. The modifications considered in non-adversarial scenarios include noise, changes in color or illumination, compression, and mild occlusions; those considered in adversarial settings are perturbations optimized under a fixed magnitude constraint.[62, 58] For each level of perturbation, the performance and explanation stability will be assessed together. Apart from task-relevant accuracy and error metrics, explanatory stability is measured by the similarity of saliency maps or importance vectors, persistence of the ranking of the most influential features, overlap of salient regions, and functional coherence, e.g., deletion/insertion tests where masking salient regions yields the expected drop in confidence.[51, 71] Reporting should include descriptive statistics and worst cases since anomalies carry probative weight in forensic contexts.[49, 50]

The case of facial recognition clearly illustrates the approach: an explanation obtained with Grad-CAM that identifies plausible anatomical regions is valuable only if, in the face of small changes to the image, those regions remain substantially the same and the model's decision does not shift to irrelevant portions. If minimal perturbations profoundly alter the structure of the explanation, the system cannot be considered reliable for evidentiary purposes, because the explanatory narrative does not stably reflect the decision logic.[7] If, instead, the explanations remain consistent within a defined range of perturbations and performance remains compatible with the baseline, the model becomes more credible and usable as evidence.[49]

All of these activities can be formalized as an explanation auditing procedure within forensic methodologies.[49, 48] Such a procedure sets admissibility criteria and acceptance thresholds for both performance and explanation stability, defines response plans when thresholds are exceeded, and prescribes the documentation required for independent repeatability. In this way, explainability does not remain a merely descriptive attribute, but becomes a verifiable requirement which, together with robustness, supports the scientific and legal defensibility of the model's outputs.[32, 24]

5.3.1 Improving Model Auditability

Besides the foundational building blocks of explainability and robustness, making auditability an operational requirement requires provenance and model identity controls, periodic adversarial red teaming, robust defences, and continuous stability monitoring.[49, 50, 48] Concretely: maintain, in an immutable manner with signed records, verifiable hashes, and versions, data, code, weights, and configurations; perform model fingerprinting and watermarking to ensure evidential traceability; run regular campaigns of simulated attacks to reveal weaknesses and update defences; use hardening strategies like adversarial training, resilient preprocessing, and mechanisms with formal guarantees; and document explanations, feature attributions, and performance over time to find drift and inconsistencies in decisions.[54, 42, 55] These practices altogether can be put together to form an end-to-end accountability loop in which every decision made by a model comes with verifiable metadata, reinforcing scientific reproducibility and legal defensibility.[53, 19]

5.3.1.1 Forensic Documentation and Chain of Custody

Embedding explainability and robustness into forensic documentation practices enhances the transparency of AI systems and preserves evidential integrity.[48, 49] It accompanies the model throughout its entire life cycle, from input acquisition to output production and all related interpretive artefacts, in a traceable, immutable, and independently verifiable manner. Metadata are components of the chain of custody that enable the reconstruction of every relevant step of the decision process.[50, 53]

Along the pipeline, technical metadata, data metadata, operational metadata, and interpretive artefacts should be generated and consistently retained.[42] Technical metadata include model and code versions, configurations, hyperparameters, the execution environment with libraries, drivers, and operating system, immutable identifiers such as cryptographic hashes and digital signatures, and synchronized timestamps.[54, 56] Data metadata include the provenance of training, validation, and test datasets, the transformations and sanitization applied, quality checks, and the criteria used for sampling and splitting.[74] Operational metadata include the use context, the threat model, the policy versions and decision thresholds in force, roles and access permissions, application and security logs, test and validation results, and deployment decisions.[42, 19, 20] Interpretive artefacts include saliency maps, heatmaps, feature attributions, relevance scores, and counterfactuals, together with the parameters used to generate them so that complete replicability is possible.[7, 73] All materials should be archived in repositories with version control, with cross-references among data,

model, and results, and with tamper-evident registers that make any subsequent change visible.[49]

A digital chain of custody requires recording and attesting every significant event. For each phase of the process, including input, preprocessing, inference, output, explanations, any postprocessing, export, and archiving, the identity of the operator or process that performed the action, the timestamp, immutable identifiers of the artefacts involved, the active configurations, and evidence of integrity checks such as verification of hashes and signatures should be available.[48, 50] When the model is updated, for example through new weights, recalibration, or threshold changes, the chain should document the before and after states with references to commits and versions and an approved rationale, thereby making the evolution of the system and its impact on performance and explanations transparent.[55, 57, 42, 53] This approach allows a complete reconstruction in court of the model's chain of reasoning and its production context. It clarifies not only what the system decided but also which data were used, which model version and parameters were active, which robustness guarantees applied, and which explanations were generated, and it enables verification that those explanations remained stable and consistent under admissible perturbations.[49, 32] Verified, transparent, and signed documentation that references hashes and versions, together with audit trails, is often sufficient to support the evidential weight of the conclusions, to facilitate independent expert review, and to promote the acceptance of AI-based evidence in judicial proceedings.[50, 19]

Nevertheless, there remain several challenges to the integration of robustness and explainability into forensic practice: consensus on either an undisputed measure of the robustness of explanations or a definition of interpretability sufficient for forensic purposes does not exist.[24, 9] Similarly, whereas explanation robustness, infidelity, and sensitivity have various measures proposed in the literature, no operational consensus on their application or common threshold values yet exists for evidentiary use.[51, 71, 70] Added computational cost and methodological complexity of robustness testing and adversarial defences increase barriers to uptake in time-pressured investigations.[58, 62] On the regulatory side, balancing transparency with both privacy and accountability is particularly sensitive when the disclosure of explanations or technical artefacts conflicts with trade secret protection or data-protection obligations in frameworks like the EU AI Act and the GDPR.[19, 20]

The field is likely to evolve in three main directions:[49, 42]

- **Standardization:** Elaboration of benchmarks and protocols for the forensic auditing of AI systems should be consistent with existing frameworks for risk management and robustness assessment. Initiatives such as the NIST AI RMF 1.0 and the ISO/IEC 24029 series on neural–network robustness might form natural anchors for common metrics, minimum reporting requirements, and repeatable, traceable test setups.[42, 75]

- **Automation:** Native integration of explainability and robustness controls into the model pipelines, such as auto-generation and versioned logging of saliency and attribution, scheduled stress tests with standardized perturbations, continuous monitoring for explanation stability and drift, and alerting with predefined response playbooks. This is in line with the technical documentation and progressive transparency emphasized

in the EU AI Act.[19]

- **Certification:** Creation of schemes for ascertaining conformance and labeling models as forensically trustworthy before use in investigations or in court. Effective schemes combine regulatory compliance with documented robustness testing and risk-management practices and explicitly require provenance, chain of custody, explanation stability, and reproducibility.[49, 50, 42]

Together, this set of directions can bring the treatment of AI-generated evidence closer to the levels of robustness, reproducibility, and credibility expected of traditional digital or physical evidence, resting on mature evaluation metrics, automated verification processes, and certification mechanisms anchored to recognized standards.[49, 48]

Chapter 6

AI Forensics Case Studies

6.1 Case Study 1 : Facial Recognition Systems

6.1.1 Introduction

Facial recognition is among the most widely studied AI applications for studying explainability and robustness, and it provides a particularly suitable testbed for evaluating forensic trustworthiness under real-world deployment constraints. In the last ten years, face recognition has become embedded in applications for security and surveillance, access control, and even in criminal investigations.[76, 11]

Nonetheless, this technology has not emerged without a number of continuing vulnerabilities, both technical and ethical, regarding trustworthiness and admissibility as a source of evidence.

Facial recognition systems are mostly based on convolutional neural networks (CNNs), which are excellent at both pattern recognition and feature extraction from large sets of images.[2, 3] However, while CNNs are capable of extraordinary performance, they act like black boxes: the process of reasoning that leads the model to associate a specific image with a specific identity is often unknown and not understandable. This black box nature complicates the ability to audit how a decision was made and why certain aspects of a face were deemed more salient than others. In a forensic or judicial context, this opacity compromises the transparency and scientific credibility of the result.

Research has also revealed that many facial recognition systems have demographic biases, yielding elevated error rates for persons of certain ethnicities or gender identities.[11, 76, 77] These disparities can result from unbalanced training datasets and a lack of algorithmic auditing. In a forensic context, such biases can lead to erroneous identifications or false positives, calling into question the evidentiary reliability of the technology as well as its fairness. When these systems are relied upon to help substantiate investigative judgments, even minor errors or biases may carry substantive legal and ethical ramifications.

The threat posed by adversarial attacks complicates matters even further: a model can be easily fooled with little more than a small perturbation to the image, perturbations imperceptible to the human eye that nonetheless produce misclassifications.[60, 78, 79] Such manipulations can be as simple as patterned glasses, printed stickers, or digital noise, demonstrating that high accuracy systems for recognizing faces might not be trusted once deployed

into uncontrolled and adversarial environments. From a forensic perspective, these vulnerabilities directly threaten the probative value of outputs from facial recognition systems: a system that can be fooled, or whose reasoning cannot be ascertained, does not satisfy the traceability, robustness, and auditability requirements typically expected for digital evidence, as discussed in Chapter 3.

In this context, AI Forensics represents a promising methodological framework for exploring, validating, and documenting the actions of facial recognition systems. By merging Explainable Artificial Intelligence (XAI) methods, namely Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), and Integrated Gradients,[28, 38, 29] with aspects of algorithmic auditing and digital chain of custody, it becomes possible to follow and explain the model's internal decision-making process. Given this background, therefore, the current case study has a twofold purpose:

1. to evaluate the robustness and forensic reliability of an open-source facial recognition model in normal and adversarial conditions; and

2. to explore the stability and consistency of the explanations provided by XAI methods and to evaluate whether the interpretative information can be used as verifiable digital evidence in investigative or legal contexts.

6.1.1.1 Methodology

To assess the forensic reliability of facial recognition systems, this case study adopts a structured experimental framework that integrates explainability, robustness testing, and traceability analysis. The approach combines open-source facial recognition models, publicly available datasets, and Explainable AI (XAI) techniques to evaluate both the interpretability and the stability of algorithmic decisions under controlled and adversarial conditions.[22, 7] The framework proceeds in three phases:

Each phase is designed to be reproducible and contestable, with fixed configurations and systematic logging to support an algorithmic chain of custody.

1. model selection and training

2. explanation and visualization

3. robustness and forensic validation

6.1.1.2 Model Selection and Training

The first step consists of choosing a cutting-edge open-source architecture for face recognition, such as FaceNet or VGGFace2, with pre-trained weights and full documentation available.[63, 80] The choice should be motivated with regard to the task at hand, one-to-one verification, one-to-many identification, open set, computational constraints, and forensic traceability, so that design decisions can be justified and replicated.

For bias reduction in demographics, it is advisable to train or fine-tune the model on a balanced dataset, such as FairFace,[81] including subjects distributed among gender, age, and

ethnicity classes. The corpus construction should document the inclusion/exclusion criteria, percentages of each demographic class, eventual re-sampling/ re-weighting for rebalancing, and a stable split into train, validation and test using fixed seeds and stratification for sensitive categories. Every step (face detection, alignment, cropping, photometric normalization) in the preprocessing pipeline should be defined and preserved with its parameters so as not to introduce undocumented drift across environments.

The training involves fixing the seeds of all pseudo-random libraries, recording the execution environment—versions of frameworks, libraries, drivers, operating system, and hardware—and using repeatable protocols, for example, early stopping, learning-rate scheduling, regularization (weight decay, dropout), and photometric and geometric data augmentation within domain-realistic limits.[1] Fine-tuning should indicate which blocks of the network are frozen and which have been updated along with tables of hyperparameters and effective duration of epochs so that independent replication is possible.

Reference metrics should be monitored and archived as the model's baseline reliability. Besides accuracy and precision, it is useful to include the false acceptance rate (FAR) and the false rejection rate (FRR) at clearly indicated operating thresholds, the Receiver Operating Characteristic and the Detection Error Tradeoff, the Equal Error Rate point, and, when relevant, probability calibration quality, e.g., Expected Calibration Error.[76] For one-to-many use cases, it is appropriate to report CMC curves and Rank-k accuracy. The choice of the operating threshold should be justified by cost-benefit criteria or regulatory requirements, so as to separate variations due to tuning from those due to manipulation or adverse conditions. The whole experimental chain should be traceable and reproducible. Training data, pre- and feature-extraction scripts, configurations, weights, and checkpoints need to be versioned, supplemented by immutable identifiers, or hashes, together with artefact logging of results by epoch and by split. It is a good practice to keep human-readable configuration files (YAML/JSON) and a run log with cross-references to code commits and dataset versions. This way, in further phases of forensic evaluation, one can refer to the baseline reliability and distinguish with greater certainty performance drops due to attacks or manipulations from random fluctuations or undocumented changes in the pipeline.

6.1.1.3 Explanation and Visualization

The second phase focuses on the interpretability of the model's decisions using techniques from Explainable Artificial Intelligence (XAI). These methods help reveal which visual features—such as eyes, nose, or facial contours—contribute most strongly to the model's predictions, thus making the decision process auditable.[7]

Three complementary XAI techniques are applied:

- **Grad-CAM (Gradient-weighted Class Activation Mapping):** generates heatmaps highlighting spatial regions that most influence the model's output for a given class or identity.[28]
- **Layer-wise Relevance Propagation (LRP):** decomposes the model's final prediction backward through its layers to assign relevance scores to individual pixels or features.[38]

- **Integrated Gradients:** provides pixel-level attribution by comparing the gradients of the model's output with a baseline input, quantifying how much each feature contributes to the final classification.[29]

The results of these methods are visualized as activation maps, which make the system's reasoning accessible for forensic review. Each explanation is logged together with its corresponding input and output to form part of the explanation record, a core component of the forensic audit trail.[55, 57, 49, 48] This record can later be used to verify whether the model's explanations remain consistent across versions or under manipulated conditions.

6.1.1.4 Robustness and Forensic Validation

The third phase tests the stability and reliability of the model and its explanations under both normal and adversarial conditions. Three types of evaluations are performed:

1. **Adversarial testing:** Controlled perturbations are made with techniques such as the Fast Gradient Sign Method (FGSM) or Adversarial Patch attacks[60, 82] to create small modifications that may happen unintentionally (e.g., poor lighting, occlusion) or intentionally (e.g., adversarial artefacts designed to deceive the system).[58]

2. **Backdoor and data poisoning simulation:** A small portion of the training data is altered with hidden triggers or mislabeled samples to check if it is possible to force the model to misclassify when the trigger pattern is present.[58] This reflects manipulation scenarios that could undermine the forensic reliability of AI tools.

3. **Explanation stability assessment:** XAI methods used earlier (Grad-CAM, LRP, Integrated Gradients) are applied again with perturbed or manipulated inputs. Stability in the resulting explanations is measured using indices such as attribution variance, similarity scores between maps, or infidelity-based measures that quantify disagreement between the explanation and the model response.[52]

By comparing model predictions and explanation outputs across normal and adversarial states, it becomes possible to evaluate the forensic stability of both the system and its interpretive layer. From a forensic perspective, a model can be considered more reliable when it not only maintains acceptable accuracy under controlled perturbations, but also preserves a reasonable degree of logical coherence in its explanations.[7]

6.1.1.5 Data Logging and Documentation

Throughout all experimental phases, every step from data preprocessing to inference and explanation is logged to maintain full traceability.[49, 48] This includes storing metadata such as model version, dataset hash, random seed, and explanation parameters. Such documentation forms the algorithmic chain of custody, ensuring that each decision and modification is verifiable by independent auditors or experts in judicial contexts.[53] All experimental artefacts, including inputs, explanations, and metrics, are compiled into a forensic evidence package, digitally signed and timestamped to ensure authenticity and nonrepudiation.[55, 57] This practice transforms the AI model's behaviour and its derived evidence into reproducible and legally defensible objects within a forensic investigation.

6.1.2 Expected Results and Forensic Value

The experimental design in the previous section aims to produce results allowing both technical assessment of model performance and forensic evaluation of reliability and interpretability. The expected results can be categorized into three complementary groups: visual evidence, statistical evidence, and forensic evidence.[7]

6.1.2.1 Visual Evidence

The first set of results relates to visualizing how the model makes decisions. This work attempts to apply XAI techniques, namely, Grad-CAM, LRP, and Integrated Gradients, for the derivation of activation maps and attribution heatmaps representing regions within a facial image influencing the model predictions.[28, 38, 29] In a well-trained system, these maps should highlight relevant facial features (eyes, nose, mouth) rather than irrelevant background elements. Consistency across subjects and conditions denotes interpretive stability. If heatmaps change unpredictably after small perturbations, this symptom indicates explanation fragility and the lack of forensic reliability.[52]

6.1.2.2 Statistical Evidence

This section presents the quantitative metrics used to link model performance to forensic reliability in a repeatable and verifiable manner.[76] We report accuracy and precision on unperturbed data and under controlled perturbations such as noise, compression, mild occlusions and low-budget adversarial attacks, so that comparison to the baseline quantifies resilience under stress. We compute differential indicators with respect to the clean baseline, for example ΔAccuracy and ΔFAR/FRR, where FAR (False Acceptance Rate) is the frequency of erroneous acceptances and FRR (False Rejection Rate) is the frequency of erroneous rejections, and we state the operating thresholds to keep results comparable. We assess the stability of explanations by measuring how Grad-CAM and LRP attributions change under minimal input variation, using attribution variance and similarity between maps such as overlap or Intersection over Union and correlation on normalized maps, since low variability indicates explanations consistent with the model's behaviour while high variability signals fragile rationales that are difficult to defend in a forensic setting.[52] We quantify demographic bias by comparing accuracy, FAR/FRR, EER and related metrics across groups such as gender, ethnicity and age, and we report group sizes and confidence intervals because significant and persistent disparities reduce probative credibility and raise fairness concerns.[11, 76] In forensic interpretation a favorable outcome combines stable performance across clean and adversarial scenarios, small deltas, low-variability explanations and no demographic imbalances, while marked drops under perturbation, large deltas, inconsistent heatmaps and systematic group disparities indicate instability or distortion and make scientific defensibility unlikely.

6.1.2.3 Forensic Evidence and Documentation

The final result of the case study is a collection of forensic artefacts, namely digital records that document with transparency how and why the AI system arrived at certain decisions.[49,

50] The goal is twofold: first, to make every step of the pipeline reproducible and independently verifiable; second, to guarantee that the recovered material is tamper-proof in order to support the legal admissibility of the outputs. Evidence in this perspective is not a simple snapshot of results but rather a coherent dossier that links inputs, configurations, inferences, and explanations through a chain of cross-references.

At the core of the evidence are the explanation records: for each inference, the input image is preserved along with an immutable identifier, while the maps generated by the adopted XAI methods (Grad-CAM, LRP, Integrated Gradients) are stored along with their parameters, as are the prediction scores and the decision threshold in force.[55, 57, 49, 48] These are complemented by system metadata describing the state of the model (version, checkpoint, configurations), the provenance of the data (dataset and preprocessing scripts hashes), the execution context (frameworks, libraries, drivers, operating system, hardware), randomization seeds, and synchronized inference logs with timestamps. Taken together, these elements trace the whole technical path that an independent reviewer ought to be able to retrace.

Of equal importance is the full audit trail, the record of events in chronological order across the workflow: data ingestion, preprocessing, inference, explanation generation, and result export.[49] Each event is annotated with the responsible operator or process, with the active configurations, and with integrity proofs such as hash verifications, so that any subsequent tampering becomes detectable. Integrity practices reinforce coherence of the whole: all materials are packaged into a forensic evidence bundle with manifest listing files and their hashes, and the bundle is digitally signed and time-stamped. The signature establishes authenticity and non-repudiation, and the timestamp establishes a reliable date and ordering.[53]

This organization turns the model's behaviour and the derived evidence into objects that are genuinely reproducible and defensible. An external reviewer can verify that the reported outputs derive exactly from the declared inputs and recorded configurations by recalculating hashes, checking signatures, and reproducing the key steps. In this way, the AI system operates as a forensically auditable device: each decision is reconstructible, each stage is inspectable, and each change is traceable, meeting the requirements of transparency, integrity, and accountability expected in investigative and judicial contexts.[49, 50]

6.1.2.4 Clarification of Anticipated Outcomes

This section integrates visual, statistical, and documentary results in order to provide a systematic assessment of the forensic reliability of a facial recognition system. The objective is not limited to measuring predictive performance; rather, it is to determine whether both the model's outputs and their associated explanations satisfy stability, coherence, and verifiability requirements compatible with investigative and judicial use.[49, 50]

A system may be considered forensically reliable only when two fundamental conditions are jointly satisfied.

First, **predictions must remain stable under realistic and controlled perturbations** of the input data, such as digital noise, compression artefacts, mild illumination changes, or limited occlusions.[58, 62] Limited degradation of performance metrics with respect to the clean baseline indicates that the model is not excessively sensitive to environ-

mental fluctuations or minor disturbances. This robustness is essential in forensic contexts, where acquisition conditions are rarely ideal and where evidential conclusions may depend on model behaviour under non-perfect inputs.[49]

Second, **explanations must remain coherent and logically consistent**. Attribution maps and saliency visualisations generated through XAI techniques (e.g., Grad-CAM, LRP, Integrated Gradients) should continue to highlight semantically meaningful facial regions—such as eyes, nose, and facial contours—rather than shifting toward irrelevant background elements.[28, 38, 29] The persistence of such interpretative patterns under small perturbations constitutes an indicator of explanation stability and supports the credibility of the model's reasoning process.[7, 52]

Conversely, if minimal input modifications produce substantial changes in the predicted identity or significantly alter explanatory heatmaps, the system demonstrates structural fragility. From a forensic standpoint, such instability undermines confidence in the model, suggesting that decisions may depend on incidental or uncontrolled factors, thereby reducing their probative value.[49, 50]

Importantly, the proposed framework does not merely identify weaknesses; it also provides operational guidance for mitigation. Corrective strategies include:

- **Targeted retraining**, focusing on failure cases identified during robustness and stress testing;[62]
- **Dataset rebalancing**, aimed at reducing demographic disparities and mitigating bias;[11, 76]
- **Adversarial defences**, such as adversarial training or input hardening, to enhance resilience against intentional manipulation.[58, 62]

6.1.3 Transition to Case Study 2 NLP and Text Analysis Forensics

The insights obtained from the forensic assessment of facial recognition technologies show how explainable, robust and traceable evidence work together to determine the evidential integrity of AI models in visual contexts.[11, 7] Nevertheless, vision-based models are only one part of the forensic picture. Similarly important questions arise with natural language processing (NLP) models, which interpret, classify, or generate textual data that may acquire evidential or legal value. In contrast to facial recognition models, which bring spatial or visual biases and adversarial threats, language models present semantic and linguistic ambiguity that generate opportunities for subtle changes, such as word replacements, change in tone, or modifying punctuation, that influence a model's classification outcomes.[32, 24] This makes the forensic assessment of language models uniquely difficult: models may generate results that appear coherent, while still embedding systemic inconsistencies or representational biases. Thus, the next case study departs from visual recognition technologies to natural language interpretation and assesses how AI Forensics can be applied to text-based AI systems. This section examines how traceability, explainability, and robustness in NLP model evidence under semantic disturbances can lead to coherent forensic analysis, extending the framework from visual anomaly interpretation to textual domains.

6.2 Case Study 2 : NLP and Text Analysis Forensics

6.2.1 Introduction

Traditionally, computer vision has received most of the attention in forensics. The accelerated development of Natural Language Processing (NLP), however, brings another domain to the forefront for AI Forensics. Applications involving NLP include text classification, sentiment analysis, detection of threats or online hate speech, authorship attribution, and support for legal document review.[32] More recently, large language models (LLMs) have extended NLP systems beyond classification to text generation, summarization, and question answering, significantly expanding the range of forensic and evidential scenarios.[5] From a forensic perspective, LLMs introduce additional challenges related to prompt sensitivity, stochastic generation mechanisms, and the need to preserve prompts, system instructions, and decoding parameters as part of the algorithmic chain of custody.[42] Without these controls, identical prompts may yield different outputs across runs, undermining replicability and weakening the probative value of model-generated text in forensic contexts. In these settings, the results produced by such models can constitute potentially usable digital evidence.

Unlike visual models, NLP models operate on meaning and context. Semantics can vary with culture, situation, and subtle linguistic cues.[22] Small changes, such as replacing a word, adjusting punctuation, or shifting tone, can lead a model to a different semantic interpretation and therefore a different decision. For example, a threat classifier may label a string "non threatening" if a direct term is replaced with a euphemism. A sentiment model may reverse an earlier conclusion in the presence of irony, negation, or context driven ambiguity. These are only a few examples that illustrate how complex textual inference can be and why systematic forensic evaluation is required, especially when model outputs may have legal, civil, or disciplinary implications.

From a forensic perspective, three aspects are crucial for NLP models:

- **Interpretability** — what linguistic signals drive the model's decision.
- **Robustness** — whether conclusions remain stable under small linguistic changes.
- **Linguistic Fairness** — absence of systematic bias across groups, dialects, or registers.[32, 24]

Tools that have been developed for numerical or visual data need adaptation for semantic complexity and decision-making logic. AI Forensics for text is about making sure linguistic inference remains practicable and traceable; every decision made by these models should be reconstructible — that is, reverse-engineered — for validation.

6.2.2 Methodology

Building upon the logging and chain-of-custody principles introduced in the facial recognition case study, this section adapts the same forensic controls to textual data, where perturbations are semantic rather than pixel-level.

To address these issues, this case study proposes a forensic analysis of a text classification model trained for sentiment analysis or toxicity detection. These are two NLP tasks in which interpretability and bias play a central role. A widely used open-source model, such as BERT, DistilBERT, or RoBERTa,[83, 84, 85] can be employed and fine-tuned on a public dataset like *IMDB Reviews* for sentiment or the *Jigsaw Toxic Comment* dataset for hate speech.[86, 87]

The forensic assessment is structured into four integrated elements, with end-to-end traceability:

- **Preprocessing** of texts and labels with a record of all the transformations applied and parameters used.
- **Explainability (XAI) analysis**: explaining the model decision and attributing it to specific linguistic cues with methods adapted to text.
- **Adversarial and sensitivity testing**: checking the stability of predictions under small semantic, lexical, or punctuation changes.[61, 88]
- **Forensic documentation**: complete metadata, versions, hashes, and logs required to reconstruct and verify every step of the pipeline in detail.[49]

6.2.2.1 Data Preparation and Model Training

First, the corpus is curated and aligned to the experimental objective to form a reliable and reproducible foundation. Duplicates are removed, as well as any data leakage across splits. The inclusion and exclusion criteria are stated, and there is a balanced representation of classes, such as positive/negative for sentiment or toxic/non-toxic for toxicity.[86, 87] On the other hand, the split into training, validation, and test sets is done in such a way that label proportions are preserved in each partition, and performance estimates are comparable. If perfect balance isn't possible, a mitigation strategy is documented, like class weighting or resampling.

Text preprocessing is specified beforehand and extensively logged. The pipeline explains whether punctuation is normalized or kept, how emojis and mentions are treated, casing, and the policies regarding URLs and stopwords. The tokenization scheme and its version are recorded along with maximum sequence length, padding and truncation policies, special tokens, out-of-vocabulary handling, and casing settings. For BERT-family models, the tokenizer is typically WordPiece, while in families such as RoBERTa, it is byte-level Byte Pair Encoding; in both cases, the vocabulary and the tokenizer version are captured since even small differences can change input sequences and thereby model behaviour.[83, 85]

This allows reproducibility by setting the seeds of all pseudo-random components and by logging the execution environment in as much detail as possible: framework and library versions, drivers, operating system, and hardware. Where possible, deterministic execution is enabled to avoid non-deterministic operators. All transformations, hyperparameters, configuration files, and produced artefacts—checkpoints, logs, plots—are tracked by means of an experiment manager, linking each run with the exact code, data, and results. For foren-

sic traceability, key artefacts are also marked with cryptographic hashes and paired with synchronized timestamps.[49, 48]

The training protocol is defined before execution and specifies the optimizer, learning-rate schedule, early-stopping criteria, and regularization. If residual class imbalance remains, class weights or informed sampling are applied and justified. At the end of training, baseline metrics are computed and archived on validation and test: accuracy, F1 score (macro and micro where appropriate), and the confusion matrix provide an immediate view of behaviour.[22, 7] In the presence of class imbalance, precision–recall curves and PR-AUC are also reported, as they are more informative than ROC in such settings. Probability calibration is evaluated, for example with Expected Calibration Error or temperature scaling, which is important when decision thresholds must be justified in a forensic context.[7] All metrics are accompanied by confidence intervals and, where appropriate, subgroup analyses, so that the baseline is statistically sound.

The combination of prepared data, recorded configurations, fixed seeds, documented environment, integrity checks, and baseline performance serves as the anchor for subsequent forensic comparisons. Differences observed during semantic stress tests or adversarial evaluations can thus be attributed with greater confidence to the perturbations under study, rather than to undocumented changes in preparation or training.

6.2.2.2 Explainability and Feature Attribution

After training, the model's predictions are analyzed with explainability procedures tailored to NLP, so we can see which words influenced the decision and by how much. We use three approaches:

- **LIME**: builds a small local approximation of the model around a single text and highlights which words or phrases contributed most to the prediction.[26]
- **SHAP**: estimates the value of each token with respect to the output and provides both local explanations for a single text and a global, consistent measure of feature importance.[27]
- **Integrated Gradients**: relative to a baseline (like an empty sequence), integrates gradients along the path from baseline to actual input with the goal of measuring how much each word contributed toward the final score.[29]

Explanations are shown as token-level heatmaps in which each word is colored according to its effect on the predicted class. A positive attribution means the word pushes the model toward that class, whereas a negative attribution pulls it away, and low intensity means little or no influence. In a hate speech task, tokens like "hate" or "kill" should have strong positive attribution towards the toxic label, while neutral or function words, like "the" and "and," should have negligible attribution.[32]

If very high attributions appear on irrelevant or benign words, that is a signal that the model is misaligned or that the training contains biases. In those cases, attribution maps serve as a diagnostic tool since they help indicate where to intervene—on data by cleaning or rebalancing, on preprocessing, or on decision thresholds—to fix the issue.[24]

6.2.2.3 Adversarial and Robustness Testing

In measuring robustness, the text inputs are subjected to controlled adversarial perturbations. These perturbations take the following forms:

- **Synonym replacement attacks:** key words are replaced with semantically similar, but lexically different terms (for example, "angry" → "mad").[61]
- **Character-level noise:** characters are inserted or deleted, in order to simulate a text with typos or obfuscation.[88]
- **Negation and punctuation flips:** sentence structure is altered to test if the model accurately processes linguistic polarity.[32]

After the perturbation, the predictions and explanations from the model are analyzed compared to the original text. To test the stability of explanations, the variability of token-level SHAP or LIME values are measured from clean versus adversarial samples. Significant deviations will suggest a considerable level of sensitivity to linguistic noise. This would damage the forensic robustness of the system.[52]

6.2.2.4 Forensic Logging and Documentation

To ensure complete traceability and independent verifiability, every step of the process, from input acquisition to inference, from explanation generation to perturbation test execution, is recorded in an automated fashion with normalized metadata and cross references.[49, 48] The record includes: model and code version (commits and tags); datasets and splits version and hash; random seeds; configurations and hyperparameters; tokenizer and vocabulary version; execution environment with libraries, drivers, and operating system; identity of the process or operator; synchronized timestamps; and results of integrity checks.

The relevant artefacts, including original and perturbed inputs, outputs, attribution maps, application and security logs, and metric reports, are stored in version-controlled repositories and tamper-evident registers, accompanied by cryptographic hashes and, where appropriate, digital signatures.[53] Each item is linked to the case and to the experimental session through unique identifiers so as to enable the reconstruction of the decision path and of the timeline of events. Retention and access policies are defined, too, including the documentation of roles and permissions and minimization measures for whatever sensitive data may be present.

This infrastructure creates a forensic traceability log comparable to a digital chain of custody. Where reproducible documentation exists, an expert can determine what was decided, how and when it was decided, rerun the experiment with the same seeds and configurations, and attest to the integrity of the materials. In this manner, text classification can be transformed from an opaque black box into a more transparent, repeatable, and potentially legally defensible procedure, provided that appropriate forensic controls are applied.[49]

6.2.3 Anticipated Findings and Forensic Value

The present study is designed to establish three converging lines of evidence: linguistic, statistical, and forensic. Together, these will support an assessment of the evidential reliability of NLP systems.[32]

- **Linguistic evidence:** token-level explanations with predictable, coherent importance assignments to truly informative words across comparable samples. A sound system keeps the salient terms constantly influential and avoids high attributions on irrelevant tokens.
- **Statistical evidence:** quantitative measures of the stability of predictions and explanations under controlled text perturbations. We investigate how outputs and attribution maps change when synonyms, typos, or small edits to negation and punctuation are introduced and expect that the variation is kept within predefined bounds.[61, 88, 52]
- **Forensic evidence:** complete, reproducible logs that record every inference and every explanation, together with enough metadata and cross-references to allow independent verification and reconstruction of the decision process.[49]

If small lexical edits yield stable explanations, consistent with prior records, and the performance metrics remain within the agreed ranges, the outputs would be more plausibly regarded as robust and interpretable within the forensic criteria defined in Chapter 3, subject to independent replication. Should explanations or predictions change significantly under minor perturbations, the system will be fragile under forensic analysis and unsuitable for evidential use until re-audited with corrective measures taken.

6.2.4 Discussion and Forensic Reflections

The forensic evaluation of NLP systems shows that linguistic reasoning is far more abstract and context-dependent than visual perception. In vision models, spatial correlations can be related to observables such as contours or facial regions. In the case of textual models, the decision-making process is based on semantics, tone, pragmatics, and syntax—dimensions far more difficult to quantify, stabilize, and reproduce.[32, 24] In other words, evidential integrity in AI applied to text cannot be guaranteed through robustness tests alone. Rather, it needs interpretability tools that are sensitive to meaning, cultural nuances, and contextual use.

From a forensic point of view, two different risks arise. The first has to do with the quality of the explanations, which might sound plausible while hiding structural inconsistencies, a situation similar to what happens when methods of attribution assign salience to words that are actually irrelevant.[24] The second risk is manipulability of the system, where strategic paraphrases, minor lexical changes, or data poisoning may change outcomes and rationales without appreciable changes in informational content.[61, 88] Under these conditions, textual evidence needs to be traceable and its metadata complete to enable independent reconstruction of the decision-making process and verification of its integrity.

In other words, AI Forensics, when applied in the textual domain, will be able to combine explainability, robustness testing, and structured documentation in support of semantic transparency and evidential reproducibility.[49] Classification is no longer a black box but an auditable process where every inference will be substantiated by interpretable rationales, stability measures against controlled perturbations, and a digital chain of custody that protects authenticity.

The contribution of this case study lies in its demonstration that the principles of AI Forensics are transferable beyond the visual context to provide a unified methodological basis for the verification of intelligent systems across many modalities.[22] Combining meaning-oriented explainability tools, adversarial testing, and scrupulous documentation, textual outputs become traceable, reproducible, and defendable in court, thereby enhancing the reliability of digital evidence and ensuring algorithmic accountability.

6.2.5 Transition to Case Study 3 : Recommender and Content Moderation Systems

The forensic evaluation of Natural Language Processing (NLP) models has shown that explainability and robustness are closely connected to the semantics and context of human communication.[32, 24] However, the impact of artificial intelligence extends far beyond the analysis of texts or abstract classifications. In today's digital ecosystem, AI systems increasingly act as true filters and mediators of information, shaping what users see, read, or share online.[43] This role of algorithmic gatekeeper is particularly evident in recommender systems, such as advertising engines or the news feeds of social platforms, and in content moderation systems, often managed by automated agents that determine the visibility, ranking, or removal of content.

These technologies create, from a forensic perspective, new and complex challenges. The process of decision-making is dynamic and collective, not predetermined by the model's own internal behaviour but influenced by personalization components, ranking algorithms, and user-generated feedback. Unlike static models, recommender systems constantly output a stream of decisions which gets continuously updated with every user interaction and profile modifications.[43] In most cases, the criteria and decision weights are not explicitly presented, and hence the way priorities are defined or content selected remains obscure.

As already observed in previous case studies, reconstructing the algorithmic reasoning and its evolution over time is a complex and multidimensional task. Establishing an algorithmic fingerprint capable of explaining, for instance, why a user was recommended a specific movie or why a post was flagged or removed requires examining distributed and interconnected decision chains, in which individual components do not necessarily correspond to a single responsible actor or decision point.[45]

This third case study addresses the issue of the forensic auditability of recommender and content moderation systems, with the aim of exploring how explainability, accountability, and data traceability can be combined to make these inherently complex, adaptive, and high-impact systems verifiable in practice. By analyzing representative use cases from the domains of recommendation and online moderation, this research seeks to demonstrate how the principles of AI Forensics can strengthen algorithmic transparency, the validation of digital evidence, and operational accountability within modern digital ecosystems that hold increasing social and legal significance.[43, 45]

6.3 Case Study 3 : Recommender and Content Moderation Systems

6.3.1 Introduction

A recommendation engine is an algorithm that provides recommendation lists to end users such that the suggested items are films, articles, videos, or other products depending on different user behaviours.[43] The most popular recommendation engines include those found on platforms like Netflix, YouTube, Amazon, and those found on social media feeds.

Content moderation platforms are designed to identify and manage content that is deemed to be inappropriate, harmful, offensive, and illegal. This content might include hate speech, misinformation, and other instances that violate the terms and conditions placed on the platforms by users.[45]

The increasing impact of such technologies has led to critical concerns about transparency, fairness, and accountability. The opacity and unpredictability of the logic that determines the rules and ranking in the case of recommenders are seen as the weaknesses. The opacity in the processes of the moderation technology and the inconsistency or unpredictability in the rules applied are the concerns in moderation technology.[43, 45]

When these tools are applied to sensitive domains such as security, discovery of digital evidence, and legal compliance, the need to explain, trace, and validate the decisions made by these tools becomes critical from the perspective of forensics.[49]

In contrast to computer vision models and language models, the output of recommendation models is not static but variable depending on ranking, personalization, and filtering techniques. These models' outputs are determined by more than just the parameters; rather, they are shaped by user-generated content and the interactive feedback process that joins user behaviour with prediction results.[43] Such models pose significant challenges to forensic analysis since the conclusion has to result from a sequence driven by technical and behavioral factors.

In forensic terms, the key difficulty is that identical users and identical content may yield different outcomes over time due to personalization, feedback loops, and system updates, complicating reconstruction and contestation.

In this case study, the principles of AI Forensics are employed to examine the traceability, explainability, and accountability that are essential in either recommendation algorithms or moderation processes. The aim is to establish a methodical process that can record and reconstruct the process of decision making in such algorithms in a verifiable and repeatable manner.[49, 50]

6.3.1.1 Methodology

This work proposes an integrated forensic framework that combines tools for explainability, data provenance tracking, and adversarial evaluation in order to systematically analyze, reconstruct, and verify the decisions produced by recommender and content moderation systems.[43, 45] The objective is to provide a structured methodology that enables the understanding not only of the outcomes of algorithmic decisions but also of the logical and technical

processes that generate them, thereby ensuring transparency, verifiability, and reproducibility of the decision-making pipeline.

The methodological approach is organized into four main components:

1. **System Analysis and Data Capture:** identification of the system architecture, data sources, and decision flows, with controlled collection of relevant data, logs, and metadata.

2. **Explainability and Reasoning Reconstruction:** analysis of how recommendations or moderation decisions are produced through interpretability techniques.

3. **Adversarial Testing and Bias Testing:** evaluation of system robustness and fairness under controlled perturbations and bias-sensitive scenarios.

4. **Forensic Audit and Documentation:** validation of results and creation of a structured evidential archive in accordance with digital chain-of-custody principles.[49]

6.3.1.2 System Analysis and Data Capture

A forensic audit of a recommender system or a content moderation system starts with a detailed reconstruction and mapping of its architecture, which is critical to ensure a proper understanding of how the system functions internally and how the decision-making logic is applied by the model.[43] Knowledge and understanding of the structure make it possible to identify the critical observation and data acquisition points that will later be targeted in the audit process.

For instance, a case study could involve a recommendation model trained on datasets such as *MovieLens*, a recognized benchmark for evaluating the performance of recommendation algorithms.[89] After defining the context, the system is decomposed into its main components. The *candidate generation model* is responsible for identifying potentially relevant items for the user. The *ranking model* orders the results according to their relevance or likelihood of interaction. The *feedback or personalization component* dynamically adjusts the recommendations based on the user's behaviour and preferences.[43]

Forensic data acquisition should systematically collect all relevant data that is necessary to trace the system's behaviour and reproduce it later under controlled conditions. The main categories of data may include user interaction logs that record clicks, viewing times, and shares; the outputs generated by the system for a set of defined reference users; and technical metadata related to model parameters, operational settings, and timestamps associated with each processing event.[45]

All the gathered data and artefacts are stored in a secure repository that follows strict integrity and verifiability requirements. Each item is linked to a cryptographic hash and a timestamp to ensure authenticity and immutability over time. This process establishes a verifiable digital chain of custody, which is a fundamental requirement for ensuring that the forensic analysis can be repeated and validated by other researchers or independent experts under the same operational conditions.[53, 49] Such a protocol guarantees that the results of the forensic examination are reproducible and forensically defensible.

6.3.1.3 Explainability and Reasoning Reconstruction

In recommender and moderation systems, explainability entails an understanding of why some items are highlighted, others demoted, and yet others removed. What the model decides is not the focus anymore; rather, how it reaches that decision is. In this regard, the reconstruction of its reasoning becomes central from a forensic perspective since it allows for the assessment of coherence, non-discrimination, and legal defensibility of algorithmic decisions.[43, 45]

A first family of methods is represented by feature attribution techniques, such as SHAP or Integrated Gradients.[27, 29] These approaches aim at identifying which user features—namely, preferences, history, profile attributes—or content features—such as text, metadata, engagement signals—have most strongly influenced the outcome, for example, a higher position in a recommendation list, or the decision to moderate a certain item. From a forensic standpoint, feature attribution enables one to move from an opaque output to an explicit representation of the factors that drove the decision, useful both for evaluating its reasonableness and detecting possible systematic patterns of bias.

A second category involves counterfactual explanations: the idea here is to simulate "what if" scenarios, in a controlled way, by changing certain variables—for example, suppressing a user preference, modifying profile attributes, or changing metadata about the format, category, or modality of the content.[43] Analysing how the output changes when such elements are perturbed, it becomes possible to understand which are the real influential factors in recommendation or moderation decisions. Forensically, this helps to assess the robustness of algorithmic decisions and clarify to what extent a given outcome was inevitable, or instead depended on few sensitive parameters.

A third strand of work concerns graph-based explanations. In this approach, relationships among users, items, and recommendation paths are modelled as a network, within which ties, information flows, and emergent structures can be visualized.[45] This view thus makes it feasible to show the formation of echo chambers, to pinpoint clusters in which certain categories of content are systematically rewarded or penalized, and to rebuild the specific "path" that led to the suggestion or removal of a given item. Such analyses are particularly significant for AI Forensics because they locate a single decision within the greater relational context, which is often crucial to making sense of impact and responsibility. In addition to these methodological tools, moderation systems should be able to generate explanation reports that detail the standards used for flagging or removal, primary linguistic or visual features which are associated with escalation, and the human review steps that intervened in the process. This information, in a forensic setting, should be recorded within a structured audit log, which enables independent reviewers to verify whether the system has applied consistent standards over time and across comparable cases; whether the decisions are documented to a degree that makes them legally defensible; and whether systematic errors or discriminatory effects can be traced back to the model's operation.[49, 50]

6.3.1.4 Adversarial Testing and Bias Testing

It is not sufficient to observe how the recommender or moderation system behaves under ordinary conditions. Rather, in order to make rigorous assessments about the resilience and

fairness of such systems, they need to be exposed to test scenarios designed to carefully probe the presence of structural bias and the vulnerability to adversarial behaviour and manipulation.[43, 45] These tests do not serve merely to improve the model technically but to generate probative material that could be used, in audits or legal proceedings, to show whether the system is actually robust or, on the contrary, easily circumvented.

In bias testing, one wants to evaluate whether the system systematically favors certain topics, user groups, or content types over others when user preferences are kept constant.[45] In practice, this usually involves creating test scenarios in which the user's context and intentions remain identical, while only sensitive or correlated attributes are varied: for instance, belonging to a specific category or being associated with a particular content type. If, under entirely identical conditions, some content is always given more visibility or systematically penalized purely because of these associations, this provides strong indications of algorithmic discrimination or exposure bias, with direct implications for the assessment of the fairness, correctness, and legitimacy of the decisions made by the system.

By contrast, manipulation testing focuses on the system's response to artificial engagement. Here, one simulates situations in which interaction metrics such as likes, shares, or comments are artificially inflated, or where traffic appears coordinated in a way that resembles bot networks or organised campaigns.[43] The purpose is to understand whether the model is capable of resisting these manipulation attempts, or whether it automatically promotes content that exhibits abnormal engagement patterns, regardless of quality or compliance with platform rules. From a forensic point of view, such experiments document the extent to which the platform can be exploited in order to amplify misleading, harmful, or unlawful content; they also help to define the responsibilities of those who design and operate the system.

Finally, adversarial evasion testing concerns the possibility of bypassing moderation filters by minimal but targeted changes in content or metadata. One constructs, in this respect, potentially harmful or borderline content—for example, hate speech or disinformation—and subtly modifies it at the lexical, visual, or categorisation level.[45] The idea is to check whether the system still detects the problem with such content or whether those small modifications are sufficient to enable the content to avoid detection. Such a test makes it possible to map the system's vulnerabilities onto deliberate evasion strategies, an aspect that might turn out to be very relevant in court when reconstructing events and assessing whether the moderation measures in place were adequate. Taken together, these forms of testing provide a structured picture of system resilience and of its genuine auditability. They make it possible to determine whether outputs remain stable in the face of deliberate manipulation attempts, or whether the model can easily be distorted through targeted interventions.[49, 50]

6.3.1.5 Forensic Audit and Documentation

This is the final stage of AI Forensics, which does not deal with carrying out further tests or fine-grained analyses of the model but rather with translating all the previous work into a structured documentary output: the forensic audit report.[49] At this point, every element collected along the pipeline—system logs, test results, explanations, robustness evaluations, and adversarial experiments—is organized into a coherent body of technical evidence that can be invoked, understood, and assessed in the context of inspections or legal proceedings.

A forensic audit report must show, above all, detailed information about the provenance of the datasets and models used. That is, the origin of the data, the criteria by which they were selected, their transformations, model versions in use, and execution environments in which analysis took place.[53] This level of detail is not merely descriptive; it allows for a precise reproduction of which conjunctions of data, code, and configuration produced a given output, hence supporting authenticity and traceability of algorithmic evidence.

The report should also include the outputs of explainability methods and counterfactual analyses, together with the experimental context in which they have been generated and the corresponding expert interpretations.[43, 45] From a forensic point of view, it is not enough to demonstrate that the model provided some sort of explanation. It is necessary to document which variables were decisive for the decision under scrutiny, which alternative scenarios were simulated, and how these elements affect the evaluation of the contested outcome.

A further element involves the systematic logging of perturbation tests and changes to metadata—including experiments pertaining to bias, manipulation, and evasion of moderation filters. For each test, the report should indicate the starting conditions, the modifications implemented, the results of those modifications on the outputs, and any anomalies observed.[49] This enables—at a later stage—an informed judgment about the system's robustness against intentional manipulation and provides a factual basis for discussing whether the technical measures adopted by the provider and by the deploying entity can be considered adequate.

Finally, explicit mechanisms for safeguarding the integrity of the collected evidence must be included in the report. This can be done through digital signatures, hashing schemes and trusted time stamping, or by equivalent procedures that make it technically and legally demonstrable that logs, test results and explanation artefacts have not been altered after their production.[53] In practice, this amounts to defining a digital chain of custody, which is a precondition for the procedural usability of the technical material. Taken together, these practices ensure that every stage of the recommendation or moderation process—from input data to final decisions—can be traced, explained and subjected to independent verification. An algorithm that would otherwise remain a 'black box' can thus be rendered significantly more transparent and auditable with respect to its decision processes.[49]

6.3.2 Anticipated Outcomes and Forensic Value

The likely results of this framework can be seen in three dimensions that are closely connected: transparency, accountability, and the integrity of digital evidence.[43, 45]

From the point of view of transparency, it would be expected that the systematic application of such proposed techniques should bring about interpretable explanations for system behaviour. More concretely, this would mean being able to explain, in an individual case, why certain items were promoted, down-ranked, or removed and which factors were decisive in arriving at that outcome, such as user activity, content metadata, or patterns of prior engagement with comparable items. The consistency of such explanations, when identical or comparable inputs are concerned, is an indication of reliable model behaviour. Conversely, erratic or contradictory explanations bring into question the forensic robustness of the system and give grounds for concern about its suitability as a source of evidence in investigative or

judicial contexts.

On the accountability dimension, it is the possibility of documenting model actions and decision paths in a structured way that allows the construction of an actual audit trail relevant both technically and legally.[49, 50] This trail would give regulators, investigative bodies, and affected parties a proper right to understand the reasoning behind a given algorithmic action, becoming relevant in light of the transparency and contestability requirements following from instruments such as the EU AI Act and the GDPR, particularly for decisions based on automated processing.[53, 20, 19] In this sense, the audit trail becomes a point of contact between judicial oversight and technical documentation practices. Finally, regarding integrity, the combination of cryptographically protected logging and fully reproducible analyses enables attesting that the entire forensic process has been conducted in a tamper-resistant and verifiable manner. Showing that the input data, the transformations applied, the tests results, and the generated explanations have not been altered ex post is a necessary condition if the resulting artefacts are to be treated as reliable digital evidence. Only under these conditions will algorithmic descriptions be turned into something usable and defensible as evidentiary material in both investigations and court proceedings, thereby bringing the operation of recommender and moderation systems in line with the standards of authenticity, integrity, and controllability that characterise contemporary forensic science.[49, 53]

6.4 Comparative Synthesis of Case Studies

The three case studies analyzed in this chapter, regarding computer vision, natural language processing, and recommendation and moderation systems, make a collective point that AI Forensics is not bound to a technology, model type, or data modality.[22, 7] All systems, while moving across different modalities (visual, textual, behavioral), purposes (recognition, classification, recommendation), and representational formats, share the same fundamental lines of tension: opaque decision making, vulnerability to strategic manipulation, difficulties in clearly attributing legal responsibility for algorithmic outcomes. From a comparative perspective, these would therefore be indicative of structural features of contemporary AI systems that a forensic approach should be able to address coherently.[49, 50]

The three analyses also converge on a common methodological core. First, explainability emerges as central to transforming raw model behaviour into something open to critical scrutiny.[22, 7] Attention maps in facial recognition, token level attributions in language models, and counterfactual or graph based explanations in recommender systems all serve the same function: they make the internal logic of decisions intelligible and thus open to challenge.[26, 27, 29, 43] Secondly, robustness testing, conducted through controlled perturbations or adversarial scenarios, puts the stability of that logic to the test, showing that forensic validity depends not only on predictive performance but also on the model's ability to provide coherent explanations under varying conditions.[58, 52] Thirdly, systematic traceability and documentation—through logging, metadata, versioning and, where appropriate, cryptographic mechanisms—delineate a genuine algorithmic chain of custody within which individual decisions can be reconstructed and subjected to independent verification.[49, 53]

Along with this common architecture, each domain exposes particular forensic challenges.

For instance, in face identification, the potential for demographic bias and ease of manipulation of visual inputs have direct bearing on evidentiary integrity, whereas in NLP, semantic and cultural ambiguity create special problems in crafting explanations that are both adequately stable and sensitive to contexts.[11, 32] In recommendation and moderation systems, the dynamic and multi-user setting gives rise to issues of traceability and distributed responsibility within algorithmic ecosystems evolving over time.[43, 45] Taken together, these differences indicate that AI Forensics must operate on multiple levels, from the micro-analysis of individual predictions to the macro-assessment of system-level behaviours. Overall, the chapter supports such an understanding of AI Forensics as a unifying methodological framework that couples machine learning practice with the requirements of forensic science and the law of evidence.[49] Core commitments concern the reconstructability of decisions, the stability and robustness of explanations, careful attention to provenance and chain of custody for data and models, and a continuous alignment with legal and ethical constraints in high-impact contexts.[53, 20, 19] This would suggest that in investigative and evidential settings, forensic audit functions should not be conceived of as a reactive ex post safeguard but as a design requirement to be integrated from the earliest stages of development, deployment, and monitoring of AI systems.[49, 50]

Overall, these case studies support the view that AI Forensics should be treated not as an ex post add-on, but as a design-oriented methodology accompanying AI systems throughout their lifecycle, ensuring that outputs remain traceable, testable, and contestable.

Chapter 7

Legal and Regulatory Dimensions of AI Forensics

7.1 The European Regulatory Framework: AI Act and GDPR

The development of Artificial Intelligence applications in law enforcement, investigative practice, and the evaluation of evidence raises foundational questions about the degree of trust that can be placed in AI outputs, and how these outputs can or may be verified or challenged.[49, 50] In this chapter, we shift from the methodological and experimental foundations of AI Forensics with which the preceding chapters were concerned to the legal and regulatory frameworks that already, or will in the near future, shape the conditions under which these methodologies can or – in some cases – cannot operate in court and within institutions.

Central to the investigation into forensic approaches to AI – and to a fundamental underlying tension – is the understanding that AI systems can produce information or evidence that appears objective and precise, but the internal reasoning structure that leads to these outputs is not usually visible or directly verifiable.[8, 6, 22] Accordingly, the opacity of AI systems, and issues of visibility associated with their application, raise core rule-of-law concerns around transparency, accountability, and due process. When an algorithmic decision is pertinent to a legal judgment or investigative action, much as would be expected from traditional forensic evidence, it should be traceable, reproducible, and contestable.[7, 49]

European law has already begun to address this convergence through two complementary frameworks: the Artificial Intelligence Act (AI Act) and the General Data Protection Regulation (GDPR).[19, 20] At the European level, these instruments do more than address AI at a purely technical level; taken together, they shape the governance of Artificial Intelligence by framing its development and deployment within a legally enforceable structure based on risk, fundamental rights, and accountability. In the context of AI Forensics, they outline a legal route toward algorithmic accountability, directly connecting processes of technological explainability with individual rights, organizational responsibilities, and the requirements for evidential admissibility.[49, 50]

7.1.1 The AI Act: From Compliance to Accountability

The EU Artificial Intelligence Act, adopted in 2024, is the first-ever regulatory framework fully premised on regulating AI systems due to the risk they create for individuals and society.[19] The regulation clearly takes a risk-based approach, wherein all AI applications fall into one of four categories: unacceptable risk, high risk, limited risk, and minimal risk. This taxonomy is not merely descriptive but forms the immediate basis for the level of legal requirements and control mechanisms applied to systems falling within each category.

In such a framework, AI systems used for biometric identification, law enforcement, the administration of justice, and digital forensics would generally fall within the category of high-risk AI systems.[19] In this regard, it has been assessed that such systems are capable of significantly affecting fundamental rights, the fairness of judicial proceedings, and the reliability of evidential processes. For this reason, the European legislator has introduced more stringent requirements concerning transparency, governance, data quality, and human oversight. In other words, where the AI is deployed in investigative or evidential settings, the AI Act assumes that the stakes are high enough to justify more stringent safeguards against error, bias, and misuse.[49]

In the context of AI Forensics, several provisions of the AI Act stand out for particular salience:

Article 9 – Risk Management System Article 9 requires providers of high-risk AI systems to establish a structured risk management system that spans the system's entire life cycle.[19] This involves the continuous identification, analysis, monitoring, and mitigation of risks to safety, fundamental rights, and the integrity of data. From a forensic perspective, this means that potential impacts on the fairness of investigations, on the contestability of evidence, and on the possibility of systematic errors or discriminatory outcomes should already be taken into account during the design and development phase.[42] Risk is not something to be addressed ex post, but is built into the model's governance from the outset.

Article 12 – Record-keeping and Logging Article 12 requires that high-risk AI systems have logging capabilities to maintain records of the operation, the key inputs and outputs, and other relevant events.[19] In other words, an audit trail emerges that will enable reconstruction ex post of the system's behaviour. This provision is important for AI Forensics: without accurate and tamper-resistant logs, verification, replication, or challenging the outputs of an AI system used in an evidentiary context becomes extremely difficult.[49, 53] Logging thus becomes a condition for any serious forensic analysis of evidence generated via AI.

Article 13 – Transparency and Information to Users Article 13 makes the following requirement: Users of high-risk AI systems need to be informed about the intended purpose of the system and about its main limitations, including conditions under which outputs may or may not be reliable.[19] In the forensic or investigative context, it means that the users relying on AI, including experts, law enforcement officers, or judicial authorities, are aware of how and under what conditions the model was trained, what

kinds of errors it tends to make, and when its use is inappropriate. This is a form of transparency that forms the preconditions for a critical evaluation of evidential strength, rather than a reception of AI outputs as opaque or overly authoritative verdicts.[22, 7]

Article 14 – Human Oversight Article 14 requires that high-risk AI systems be designed in such a way as to ensure effective human oversight.[19] It is not sufficient for a human to be nominally "in the loop": the oversight must enable the human actor to understand, question, and, where appropriate, correct or override the system's output. In a forensic context, this is essential to preventing AI from assuming the role of an *unchallengeable* decision-making authority. The ultimate decision-maker – whether a judge, prosecutor, investigator, or expert – must retain the final responsibility and be able to discount or reframe algorithmic evidence where it conflicts with other elements in the case.[49, 50]

Taken together, these obligations turn what might previously have been considered "best practices" in AI development – such as explainability, model robustness, and operational traceability – into legally enforceable duties.[42, 75] The AI Act does not just encourage high standards of technical quality; it makes them legally binding and attaches sanctions to non-compliance. From the viewpoint of AI Forensics, this means that the conditions needed to render a system forensically sound – replicable, auditable, and intelligible – can no longer be regarded as optional add-ons but part of the baseline legal requirements for deploying high-risk AI systems.[49]

7.1.2 The GDPR and "Right to Explanation"

The General Data Protection Regulation (GDPR), which entered into force in 2018, remains the principal legal instrument governing the use of personal data within the European Union and, as such, one of the central regulatory frameworks shaping the development and deployment of Artificial Intelligence.[20] Its relevance for AI Forensics rests particularly on two key provisions: Article 22, which guarantees that individuals shall not be subject to decisions based solely on automated processing, and Article 15, which introduces a right of access to "meaningful information about the logic involved" with regard to automated decisions falling within its scope. Taken together, these provisions form the foundation of what is often described in the literature as a European "right to explanation".

In the context of AI Forensics, these provisions assume particular importance:

Article 22 – Automated Decision-Making and Human Intervention Article 22 introduces an explicit constraint on exclusive reliance on automated systems, stipulating that individuals should not be subject to decisions that produce legal effects concerning them, or similarly significantly affect them, where such decisions are based solely on automated processing, except under specific conditions and with appropriate safeguards.[20] From a forensic perspective, this implies that AI models cannot be conceived as autonomously providing decision support in sensitive domains, including investigative and judicial contexts. A meaningful space for human decision-making must remain, within which human actors can critically evaluate, contextualize, and, where necessary, deviate from algorithmic outputs.[49]

Article 15 – Right of Access and Logic of Processing Article 15 further develops this framework by stating that data subjects, when exercising their right of access, are entitled not only to obtain copies of the personal data processed about them, but also, in cases where automated decision-making is involved, to receive "meaningful information about the logic involved".[20] Although the precise scope of this expression has been subject to doctrinal debate, it is generally interpreted as providing a legal basis for a certain degree of explainability: individuals should be able, at least in general terms, to understand which main factors the system has considered and how these factors contributed to the outcome.[22, 7] The GDPR does not, however, establish a fully fledged, case-specific right to a complete explanation of complex algorithmic models. Rather, it guarantees access to meaningful information about the logic involved at an appropriate level of abstraction, sufficient to enable understanding, contestability, and the exercise of procedural rights.[22, 7]

Taken together, these rights create a legal basis for explainability in AI systems. Whenever algorithmic processes are used to influence individuals or legal outcomes – for example through risk assessment, content classification, or decision support in judicial proceedings – data subjects gain a legal entitlement to an explanation of how a given output has been generated.[20, 19] Scholars such as Wachter, Mittelstadt, and Floridi (2017) have argued that this "right to explanation" functions as an important mechanism for preserving human agency and procedural fairness in algorithmic decision-making, providing a counterbalance to the tendency of AI systems to operate as opaque "black boxes."[8]

This dimension of the GDPR becomes even more significant in forensic contexts. When an AI model contributes, even partially, to the assessment of evidence – for instance in facial recognition, automated text classification, or visual content moderation – its underlying rationale cannot remain entirely opaque. It should instead be intelligible to a sufficient extent to enable experts, judges, and parties to understand which features, data, or patterns the system relied upon and to contest those assumptions where appropriate.[49, 50] Scrutiny of the functioning of the model therefore becomes an integral component of the right to a fair trial and of adversarial procedural guarantees.

7.2 Evidential Utilization and the Legal Characteristics of AI-Obtained Evidence

These regulatory principles acquire concrete relevance when outputs generated by AI systems are introduced into investigative and judicial proceedings as potential evidence.

The growing deployment of AI systems within investigations and judicial proceedings raises a crucial question regarding the law of evidence: can results generated through AI be regarded as admissible evidence in court?

It is generally not sufficient that a model produce an output that, from a technical standpoint, is correct, since for such output to take on evidential value, it must meet those same requirements of reliability and controllability already traditionally applied to scientific and forensic evidence.[49, 48]

Admissibility is typically evaluated against four core criteria in European and international forensic practice: authenticity, integrity, reliability, and relevance.[49, 57, 55] AI-generated evidence can pose challenges along all four dimensions, precisely because the process leading from input data to output is often opaque, technically complex, and highly dependent on design choices that are not always properly documented.[22, 7]

- **Authenticity**

  Authenticity deals with the ability to prove that a given piece of digital evidence is genuinely attributable to a certain model, to a particular version of that model, and to one specific dataset or set of inputs. In the context of AI, that means proving the output in court comes from the exact system configuration at a given time, with declared data, rather than from a later version, a modified model, or an altered dataset.[49] This challenge is addressed by AI Forensics, allowing mechanisms like model hashing, systematic versioning of configurations, and a chain of custody that encompasses not only output files but also models, weights, data, and execution environments.[54, 53]

- **Integrity**

  Integrity is the assurance that the output of the system has not been tampered with, either maliciously or unintentionally, after its creation. In evidential terms, it should be possible to demonstrate that what is presented to the court is exactly what the system produced at a particular moment in time. This requires intrinsic logging mechanisms, digital signatures, reliable timestamps, and secure audit trails documenting every transformation the data undergoes along the pipeline (preprocessing, inference, post-processing, transmission, and storage).[55, 57, 49, 53] Without such safeguards, the ability both to defend and contest the integrity of the evidence is considerably diminished.[49]

- **Reliability**

  Reliability refers to the system's ability to generate consistent and reproducible results under the same conditions. In the forensic context, this means the ability to reproduce the result of the model if the same inputs, configuration, and environment are used, and to demonstrate that the performance of the system has undergone serious scrutiny – e.g., through robustness testing, error analysis, and studies on bias and fairness.[75, 59, 58] Stability of explanations here plays a special role: if the explanations provided by the model – or by explainability methods – change arbitrarily for small changes in input, then the evidentiary power of the output is compromised because one cannot anchor the inference to a constant and intelligible pattern of behaviour.[52, 51]

- **Relevance**

  Relevance requires that the AI output make a concrete and meaningful contribution to the reconstruction of the facts at issue, which should be intelligible to the actors in the proceedings. A label classification or numerical score itself is not sufficient. This relationship between the output and the legally relevant questions needs to be explicable – for instance, through explanation tools, visualizations, technical reports,

and expert testimony that will translate algorithmic information into terms that judges, lawyers, and, where applicable, jurors can understand.[26, 27, 7] In the absence of such intelligibility, the evidence risks functioning like an opaque "technical verdict" that does not integrate well into overall evidential reasoning.

In the view of some commentators, including Schwartz and Gawande,[90] questions about the admissibility of AI-generated evidence will depend less and less on whether a model can be described as a "black box", and more on whether it can produce a transparent evidential trail that is open to scrutiny. Complex algorithms are not the core difficulty, in other words; rather, the core problem is the lack of mechanisms for documenting, explaining, and reviewing the inferential path followed by the model.[22, 49] From this perspective, AI Forensics does not just "read" the algorithmic outputs after the fact; rather, it is the technical-legal process whereby the computational inferences used as evidence are made legally defensible. Through structured logging, versioning, robustness analysis, experiment replication, and explanation techniques, AI Forensics tries to bridge the gap from internal functioning to the demands of legal procedure.[49, 48] Its ultimate goal is to guarantee that algorithmic evidence can meet the established standards of authenticity, integrity, reliability, and relevance on which both forensic science and the law of evidence stand.

7.3 Liability, Responsibility, and Forensic Accountability

When AI systems, directly or indirectly, contribute to evidentiary outcomes or legal decisions, one key question immediately arises: who is responsible for the consequence of an algorithmic error? If a model leads to the false accusation of a suspect, or to the unjustified exclusion of a suspect from an investigation, or to a misleading assessment of evidence, where does responsibility lie – with the developer, with the party who deployed the system, with the end user – or, improperly, with the "algorithm" itself?

The AI Act proposes an answer that departs from such a simplistic view and introduces a model of shared, triadic responsibility, in which obligations and duties are distributed across several actors along the entire life cycle of the AI system.[19, 50] There is no single "culprit" in this schema; rather, there is a network of responsibilities that must be traceable and legally reconstructible.[49]

In a nutshell, the model differentiates between:

- **Provider**
  This is the actor responsible for the design, development, and training of the AI model. They bear duties in regard to technical transparency, quality of data, risk management, and documentation.[19, 42] From a forensic perspective, the provider shall ensure the system is so designed that auditing, logging, reproducibility of outputs, and at least some explainability of its functioning are possible.[49] In case this model is structurally opaque, poorly documented, or trained on inadequate data, the primary responsibility for systematic errors shall lie with this actor.

- **Deployer**
  It is the actor who decides to adopt and integrate the system in a concrete context –

say, a law enforcement agency, a judicial authority, or a public body. The deployer does not simply "push a button": they select the purposes for which the system is used, at which stages of the investigative or decision-making process it is integrated and which internal procedures are established in order to control its effects.[50] They thus have a duty to use the system in line with the provider's instructions, to define verification protocols, to avoid function creep or distorted uses, and to implement forms of internal oversight. If a model originally designed to support exploratory analysis is used as an almost exclusive basis for determining guilt, a significant part of the responsibility for this shift in use lies with the deployer.[49]

- **User / Investigator**
  This is the actor who, in practice, interacts with the system and integrates its output into evidentiary reasoning or legal decision-making: for example, an expert, an investigator, a prosecutor, or a judge. Even where the system is technically in compliance, the user has a duty not to treat its output as a definitive verdict, but to verify it, put it in context, and compare it with other evidence.[50] From the point of view of professional diligence, this implies knowing the limitations of the model, recognizing margins of error, and rejecting or downplaying the algorithmic result if it runs afoul of other elements in the case file.[49]

From a forensic perspective, this triadic configuration reinforces the notion of forensic accountability, whereby each step in the chain – design, deployment, use – plays a role in the quality and reliability of algorithmic evidence and should therefore be reproducible ex post.[49] In this vein, AI Forensics is not confined to mere technical tools to analyze models but allows:

- identify **who** did **what**, **when**, and with **which** system configuration;[49, 50, 55]
- map an output – let's say, face recognition for a given match – to a particular model version, to a specific dataset, and to a specific context of use;[54]
- verify whether each actor acted with due diligence, complying with the legal and procedural obligations incumbent upon them.[50, 19]

In an algorithmically mediated environment, the classic legal notion of due diligence is in effect distributed across the entire life-cycle of the AI system. The provider must demonstrate that they designed a system that is auditable and documented; the deployer must show that they introduced it in a proportionate manner, with appropriate safeguards; and the user must be able to justify how they interpreted and integrated the output into their evidential reasoning.[49] The lack of such multi-level traceability tends to shift responsibility onto the algorithm, as if it were an autonomous subject, with the risk of creating a grey area in which no one is truly accountable for errors. The idea of forensic accountability developed in your thesis points in the opposite direction: using AI Forensics, it supports the reconstruction of the contribution of each actor, turning the path from model training to final decision into a chain of responsibilities that can be legally attributed.[49, 50]

7.4 Policy Developments and Outlook

The rule of law, procedural guarantees, and protection of fundamental rights intersect with forensic practice on AI to suggest that the future will not be shaped by good technical practices alone but by the construction of a genuine institutional infrastructure for the control, audit, and accountability of algorithmic systems.[20, 19] In this perspective, several policy developments seem particularly relevant to the evolution of AI Forensics within European and international regulatory frameworks.[49, 50]

A first axis concerns standardisation. Bodies such as ISO, CEN-CENELEC, and NIST are already developing norms and guidelines on risk management, safety, and performance evaluation for AI systems.[42, 75] For AI Forensics, these standards can provide shared methodologies for testing and documenting systems in evidential contexts, create a common vocabulary between technical experts and legal professionals, and offer judges an external benchmark to assess whether a system has been designed and used in line with the technological "state of the art."[59, 58] Standardisation thus becomes a bridge between the needs of legal proceedings – controllability of evidence – and the practices of model development and audit.

A second policy strand is certification. There is growing discussion about the possibility of recognizing specific certifications for "forensic-grade" AI systems, capable of attesting compliance with minimum requirements of explainability, robustness, and provenance traceability.[49, 48] The resulting schemes could be integrated with the conformity assessment procedures provided for high-risk systems under the AI Act, adding an explicitly forensic dimension – logging, replicability, audit trails, digital chain of custody – and entrusting evaluation to independent third parties.[19] In this way, the forensic quality of a system would not be only a matter of contention in individual cases, but a property recognizable ex ante through certifications that courts can refer to and scrutinize.

Another essential factor is indeed the education and training of the actors involved. In the case of judges, prosecutors, and lawyers, it is necessary to develop solid AI literacy, enabling them to understand what a model does, what its limits are, how statistical reliability and explainability shall be assessed, and when a system can be considered forensically sound.[50] For police forces and investigative practitioners, training must deal not only with the operational use of tools but also with the implications for chain of custody, logging, documentation of experiments, and replicability of results.[49] For policy-makers, finally, there is the need to establish critical reading of technical reports, impact assessments, and algorithmic audits, and to translate them into coherent regulatory choices.[42]

Finally, the development of AI Forensics requires authentic inter-professional collaboration: forensic analysis of AI systems cannot be left solely in the hands of computer scientists, just as it cannot disregard the competence of legal scholars or the contributions of ethicists and scholars specialized in the social implications of technologies.[48] There is a need for stable spaces in which collaboration can happen: interdisciplinary working groups, joint guidelines, mixed technical-legal committees, but maybe also dedicated AI forensic units supporting judicial authorities.[49, 50] Taken together, these policy developments outline the prospects of a future in which AI Forensics is institutionalised as an integral part of digital governance. Algorithmic systems would not simply be authorized or prohibited in the abstract but subjected

to standards, certifications, periodic audits, and informed scrutiny by legal practitioners. In that future scenario, AI Forensics becomes one of the central instruments through which AI systems are held accountable, governed, and examined according to standards comparable to those applied to the assessment of the competence, methodology, and credibility of human expert testimony.[49]

7.5 Conclusion

The analysis developed in this chapter has demonstrated that AI Forensics cannot be apprehended in isolation from the evolving legal and regulatory architecture that governs the use of Artificial Intelligence in Europe.[49, 50] The AI Act and the GDPR together build a dual framework: on the one hand, a risk-based regime classifying and constraining high-risk AI systems; on the other, a rights-based regime which grounds explainability and contestability in data protection and fundamental rights.[20, 19] In this framework, AI Forensics does not consist of an "optional" set of technical tools but rather the operational mechanism through which the legal requirements of traceability, transparency, and accountability may be made effective in practice.[42, 75]

This becomes clear above all when AI-generated outputs are treated as evidence. Questions of admissibility are framed anew in light of classical forensic criteria – authenticity, integrity, reliability, and relevance – which AI systems can only meet if they are accompanied by solid mechanisms of logging, versioning, explanation, and reconstruction of the inferential path followed by the model.[49, 7] At the same time, the AI Act's model of shared, triadic responsibility among providers, deployers, and users shows that evidential failures can rarely be attributed to "technology alone": they emerge along a chain of design, implementation, and use in which each actor bears a specific share of forensic responsibility.[19, 50]

Looking ahead, the consolidation of AI Forensics as a distinct field will depend not only on doctrinal developments but also on broader policy trajectories: the emergence of technical standards for audit and testing; the possible introduction of "forensic-grade" certification schemes; the diffusion of AI literacy among legal and investigative practitioners; and the institutionalization of stable forms of inter-professional cooperation.[42, 48] These might be seen to point, collectively, to a future in which AI Forensics becomes an integral component of digital governance, furnishing the interface through which algorithmic systems are rendered contestable, reviewable, and compatible with the evidential standards of the rule of law.[49, 50] In this sense, AI Forensics operates as a bridge between code and law, contributing to aligning intelligent systems with the expectations of truth, fairness, and verifiability that underpin contemporary forensic science and the law of evidence.[53]

Chapter 8

Forensic Experimental Analysis of Face Recognition Models

8.1 Introduction

This chapter develops the experimental part of the thesis through the definition and analysis of a set of concrete forensic use cases. The objective is to clearly define the perimeter of the experimental analysis and to assess whether the behavior of a face recognition model can be examined, documented, and reconstructed in a manner consistent with a forensic approach.

The selected use case is the forensic analysis of an open-source face recognition model, evaluated using publicly available and demographically annotated image datasets. Within this context, the experimental analysis focuses on demographic effects observable at the representation level and on the identification of potential model manipulations, by observing the behavior of the system through outputs and internal representations.

The model is analyzed both under normal operating conditions and under attack scenarios. In particular, the experiments consider errors and anomalies that naturally emerge during standard model usage, as well as effects intentionally introduced through adversarial techniques, including adversarial perturbations and the injection of a backdoor during training. This allows for a direct comparison between benign and malicious conditions within a controlled experimental setting.

To support the inspection and interpretation of model behavior, output analysis is complemented by the use of explainable artificial intelligence techniques. Methods such as Grad-CAM, Integrated Gradients, SHAP, and LIME are employed as support tools to visualize and compare model decisions across different conditions. These techniques are not used to infer causality, but rather as auxiliary forensic artefacts to assist forensic analysis.

A central objective of the experimental work is the production of structured and verifiable documentation of the entire analysis process. Execution logs, metrics, intermediate outputs, and explanation artefacts are systematically collected and organized with the goal of treating the model's decision process and the produced artefacts as forensic objects. This enables post-hoc verification, traceability, and consistency checks analogous to those performed in traditional digital forensics.

The experimental results presented in this chapter are therefore organized into a sequence

of studies reflecting the defined use cases: baseline characterization of model behavior, explainability analysis, demographic bias assessment, adversarial robustness testing, and backdoor vulnerability evaluation. All experiments are conducted within a unified forensic logging framework designed to preserve traceability, integrity, and reproducibility of the generated evidence.

8.2 Experimental Setup and Methodology

This section describes the experimental architecture and methodological choices adopted to implement the forensic use cases defined in this chapter. The objective is to provide an operational account of the experimental conditions, resources, and procedures used throughout the analysis, enabling structured evaluation and systematic comparison of results.

The experimental setup is designed to allow reconstruction of the model's behaviour through systematic observation of its outputs and of the information generated during execution. To this end, the analysis relies on the coordinated use of:

- the trained face recognition model and its operational configurations;
- demographically annotated input datasets;
- quantitative evaluation metrics;
- execution logs and intermediate artefacts;
- explainability tools employed to support interpretation of model decisions.

The experimental procedures are defined to ensure methodological consistency across all studies conducted in this chapter, including baseline evaluation, explainability analysis, demographic bias assessment, adversarial robustness testing, and backdoor vulnerability analysis. Each experiment follows a controlled pipeline that standardizes data processing, model execution, metric computation, and artefact collection.

All outputs and intermediate results are systematically recorded within a unified logging structure, enabling structured comparison between experimental conditions. This design supports replicability of the experiments and facilitates post-hoc inspection of the analytical process.

Overall, the adopted methodology establishes a coherent experimental framework in which the behaviour of the model can be examined in a controlled and reproducible manner, forming the basis for the subsequent forensic analysis.

8.2.1 Dataset

All experiments are conducted using the FairFace dataset, a publicly available face image dataset designed to support fairness and bias analysis in face recognition systems. The dataset provides demographic annotations for race and gender, enabling structured representation-level analysis across demographic groups.

The validation split of the dataset is used throughout all experiments and consists of 10,954 images. Each image belongs to one of seven racial categories and one of two gender

classes. Using the validation split ensures that the evaluation is performed on data independent from any training or fine-tuning procedure, preventing data leakage.

Images are loaded and preprocessed according to the input requirements of the analyzed face recognition model. For forensic traceability, a cryptographic hash is computed for each image after loading, allowing each input sample to be uniquely identified and consistently linked to outputs, metrics, and execution logs across experiments.

8.2.2 Model

The analyzed model is an open-source face recognition system based on the FaceNet architecture. The model maps input face images into a fixed-length embedding space, where identity similarity is expressed through geometric distance rather than explicit class predictions.

The implementation used in the experiments produces 512-dimensional embedding vectors that are L2-normalized. Similarity between faces is measured using Euclidean distance between embeddings. This representation-level formulation is particularly suitable for forensic analysis, as model behavior can be examined through the geometry and statistical properties of the embedding space.

Prior to experimentation, a cryptographic fingerprint of the model parameters is computed using a SHA-256 hash. This fingerprint uniquely identifies the exact model instance under analysis recorded as a forensic artefact, enabling verification of model integrity and detection of any subsequent modification.

8.2.3 Experimental Pipeline

The experimental pipeline is implemented as a structured and reproducible sequence of analysis stages, each corresponding to a specific forensic use case defined in this chapter. The pipeline is designed to reflect the progressive workflow of a forensic investigation, in which reference behavior is first established and subsequently compared against anomalous or adversarial conditions.

The pipeline consists of five main stages: baseline embedding characterization, explainability analysis, demographic representation analysis, adversarial robustness testing, and backdoor vulnerability assessment.

At each stage, the model is evaluated under explicitly defined conditions and produces structured outputs, including numerical metrics (e.g., embedding statistics, distance distributions, attack success rates), visual artefacts (e.g., explainability maps and comparative plots), and machine-readable files (CSV and JSON) summarizing the results. These outputs are persisted to disk and linked to the corresponding experimental configuration and execution context.

Intermediate artefacts generated by earlier stages, such as baseline embedding statistics and demographic group partitions, are reused in subsequent analyses. This design allows later results to be interpreted relative to previously established reference behavior, enabling consistent comparison across experiments.

All stages of the pipeline are executed under the control of the forensic logging framework. Each execution step is recorded with associated metadata, including input identifiers, param-

eter settings, timestamps, and cryptographic hashes of produced artefacts. This ensures that the full experimental workflow can be reconstructed post hoc and that observed deviations across experiments can be traced back to specific conditions or manipulations.

Overall, the pipeline structure supports both analytical rigor and forensic soundness, enabling systematic comparison between normal, biased, adversarial, and backdoored model behavior within a unified evidentiary framework.

8.2.4 Forensic Logging and Evidence Collection

All experiments are conducted under the control of a dedicated forensic logging framework specifically designed to support traceability, integrity verification, and post-hoc reconstruction of the experimental process. The framework treats each relevant action performed during the experiments as a potential forensic event and records it in a structured and machine-verifiable form.

For each experimental stage, operations such as dataset loading and preprocessing, model initialization and fingerprinting, inference execution, metric computation, explainability generation, and attack simulation are explicitly logged. Each event is associated with contextual metadata, including timestamps, configuration parameters, identifiers of the involved artefacts, and cryptographic hashes.

Logged events are linked through cryptographic hash chaining, forming a chain-of-custody structure that preserves the ordering and integrity of the experimental trace. Any modification, omission, or reordering of logged events can therefore be detected during subsequent verification, in analogy with integrity checks performed in traditional digital forensic investigations.

All artefacts produced during the experiments—including execution logs, numerical metrics, configuration files, visual outputs, and summary reports—are persisted to disk and referenced within the logging framework using unique identifiers. This enables systematic correlation between inputs, intermediate results, and final outputs across different experimental stages.

At the conclusion of each experiment, the collected logs and artefacts constitute a coherent evidentiary record that can be independently inspected and verified. This record provides the basis for the experimental results presented in the following sections and supports consistency checks, reproducibility assessment, and forensic validation of the analysis workflow.

As a forensic policy, all numerical values reported in this chapter are automatically extracted from machine-readable result artefacts (e.g., `results/metrics/*.json`, `results/metrics/*.csv`, `results/bias/*.csv`, `results/xai/*.csv`) and consolidated into `CHAPTER\_8\_DATA\_CONSOLIDATED.txt`. These artefacts, together with the corresponding audit reports (`results/forensic\_reports/*.json`), are referenced by cryptographic hashes and covered by the forensic integrity certificate (`FORENSIC\_INTEGRITY\_CERTIFICATE.json`), from which all tables and plots are generated, ensuring full traceability and minimizing transcription errors.

8.3 Experiment 1: Baseline Embedding Characterization

8.3.1 Objective

The objective of this experiment is to establish a reference characterization of the embedding space produced by the analyzed face recognition model under standard operational conditions. This characterization defines a baseline notion of normal model behavior at the representation level, which is required to interpret and contextualize all subsequent experimental findings.

Rather than evaluating recognition accuracy or task-specific performance, the experiment focuses on the geometric and statistical properties of the embeddings from which verification decisions are derived. By defining this reference behavior on a demographically diverse dataset, the experiment provides the foundational point of comparison for the forensic analyses of explainability, demographic bias, adversarial robustness, and backdoor manipulation presented in the following sections.

8.3.2 Methodology

The baseline evaluation was performed on the full validation split of the FairFace dataset, consisting of 10,954 images. Each image was processed by the reference FaceNet model to extract a 512-dimensional, L2-normalized embedding, which represents the fundamental output used in all subsequent analyses.

The experiment was implemented through the script `01\_baseline\_evaluation.py`, which orchestrates the loading of the dataset, the execution of model inference, and the computation of embedding-level metrics. For each input sample, the corresponding embedding was generated and associated with its demographic annotations, enabling structured aggregation and analysis.

To characterize the structure of the embedding space, pairwise Euclidean distances were computed exhaustively between all embeddings, resulting in approximately 60 million unique distance measurements. This exhaustive approach was deliberately adopted to avoid sampling effects and to obtain a complete and statistically stable description of the distance distribution induced by the model on the validation data.

In addition to pairwise distances, summary statistics were computed on the embeddings and distance distributions, including verification of embedding normalization and aggregation of central tendency and dispersion measures. The demographic distribution of the processed samples was also verified to ensure consistency with the dataset annotations and to provide a reference for the bias analyses conducted in later experiments.

Throughout the execution of the experiment, all relevant operations and parameters were recorded by the forensic logging framework. The experiment produced structured numerical results, a complete execution trace, and dedicated audit artefacts, which together constitute the baseline forensic artefacts used as a reference in the subsequent experimental sections.

8.3.3 Results

The baseline analysis produced a set of quantitative indicators describing the representation-level behavior of the analyzed face recognition model on the FairFace validation set.

The extracted embeddings show a highly consistent normalization behavior. The distribution of L2 norms is tightly concentrated around unity, with a standard deviation on the order of 10^{-8}.

The exhaustive computation of pairwise Euclidean distances yields a stable and well-defined distance distribution. The observed mean pairwise distance is 1.3471, with a standard deviation of 0.1312. These values indicate that the embedding space exhibits a coherent geometric structure, without evidence of degenerate behavior such as embedding collapse or excessive dispersion.

Table 8.1 reports the main numerical results obtained from the baseline evaluation, summarizing the key statistics used to characterize the embedding space under non-adversarial conditions.

Table 8.1: Baseline embedding statistics on the FairFace validation set.

Metric	**Value**
Embedding dimensionality	512
Mean L2 norm	1.0000
Std L2 norm	3.6×10^{-8}
Mean pairwise distance	1.3471
Std pairwise distance	0.1312

Figure 8.1 provides a visual summary of the baseline results, including the distribution of embedding norms, the histogram of pairwise distances, and the demographic composition of the validation set. These visual representations complement the numerical metrics and support qualitative inspection of the baseline embedding behavior.

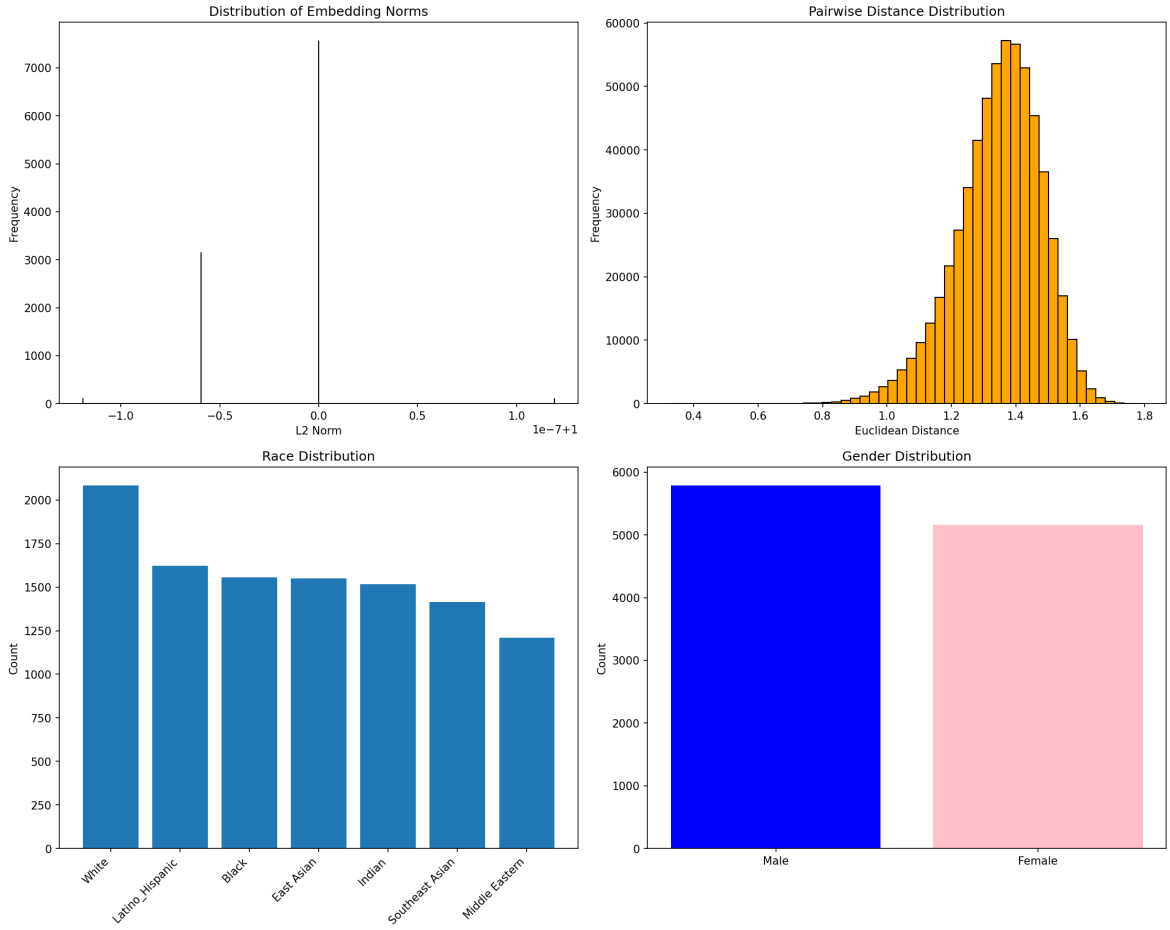

Figure 8.1: Baseline characterization of the FaceNet embedding space on the FairFace validation set, including embedding norm distribution, pairwise distance distribution, and demographic composition of the dataset.

8.3.4 Forensic Interpretation

From a forensic perspective, the baseline evaluation establishes a documented reference of how the analyzed face recognition model behaves under standard, non-adversarial operating conditions. Rather than focusing on task-level accuracy or semantic correctness, the analysis characterizes the representation-level properties of the embedding space produced by the model when no external interference is present.

The observed stability of the embedding norms and the consistency of the pairwise distance distribution provide objective indicators that the model operates in a numerically stable and coherent manner on the reference dataset. These properties are derived from measurable quantities computed exhaustively on the full validation set and are supported by both numerical metrics and visual summaries, as well as by the corresponding execution logs recorded during the experiment.

The role of the baseline analysis is therefore to document how the embedding space behaves when the model operates without external interference. By fixing this reference behavior through quantitative metrics, visual summaries, and execution logs, the baseline enables later experiments to isolate and explain deviations caused by demographic representation effects, adversarial perturbations, or training-time manipulation.

In this sense, the baseline does not constitute an evaluation of correctness or fairness, but rather defines a reproducible and verifiable reference state of the model. This reference state is essential for forensic comparison, as it allows subsequent observations to be contextualized and attributed to specific experimental conditions, supporting structured anomaly detection and evidentiary reasoning in the analyses that follow.

8.3.5 Discussion

The baseline experiment provides a structured and verifiable characterization of the representation-level behavior of the analyzed face recognition model on the FairFace validation set. The numerical metrics collected during this phase, stored in `results/metrics/baseline\_results.json`, document how the model maps input images into the embedding space under non-adversarial conditions.

The results show that the model consistently produces 512-dimensional embeddings that satisfy the expected normalization constraints imposed by the FaceNet architecture. In particular, the mean L2 norm of the embeddings is equal to 1.0, with a very small standard deviation (3.63×10^{-8}), indicating stable and uniform embedding generation across all 10,954 validation samples.

The geometry of the embedding space is further described through the computation of pairwise Euclidean distances. A total of 59,989,581 distances were evaluated, yielding a distribution with a mean of 1.3471 and a standard deviation of 0.1312. These values describe a well-defined similarity structure, in which embeddings are neither concentrated in a narrow region nor excessively spread. This geometric profile provides a concrete reference for interpreting distance variations observed in later experiments.

In addition to embedding statistics, the baseline experiment records the demographic composition of the validation set, including the distribution of samples across race and gender categories. This contextual information is stored together with the baseline metrics and is later reused to support stratified analysis in the bias and adversarial robustness experiments.

From a forensic standpoint, the baseline analysis is strengthened by the systematic documentation of the experimental process. The audit report `results/forensic\_reports/baseline\_audit.json` records the execution of the experiment as an ordered sequence of logged events, including model loading, data processing, and metric computation. These records enable reconstruction of the analysis workflow and verification of the conditions under which the reported results were obtained.

The baseline experiment establishes a quantitative profile of the model's representation-level behavior, grounded in reproducible metrics and traceable forensic artefacts. This profile is used as a reference point in the subsequent analyses to interpret variations observed in the embedding space under conditions of demographic bias, adversarial perturbations, or modifications introduced during the training process.

8.4 Experiment 2: Explainability Analysis

8.4.1 Objective

The objective of this experiment is to apply and evaluate post-hoc explainability techniques in order to generate interpretable artefacts that support the forensic analysis of a face recognition model. The experiment focuses on observing how the model leverages visual information from input images during the embedding generation process.

Specifically, the goal is to assess whether explainability techniques can highlight the regions of the input images that contribute significantly to the extraction of facial embeddings and whether such information can be consistently associated with the representations produced by the model.

The experiment also aims to produce explainability outputs that are comparable across different samples, enabling the identification of recurring patterns and the comparison of model behavior across images belonging to different demographic groups. The generated explainability artefacts are intended to be integrated with quantitative metrics and execution logs as part of the adopted forensic analysis framework.

8.4.2 Methodology

The explainability analysis was conducted on a stratified subset of the FairFace validation set, selected to balance demographic coverage and computational feasibility. A total of 1,000 images were selected using stratification to ensure coverage of all race and gender categories. The selection guarantees inclusion of each category, while not enforcing perfect class balance. This trade-off preserves demographic coverage while keeping the explainability workload computationally feasible.

The resulting sample distribution across racial categories was: Indian (144 images), White (143), Latino/Hispanic (143), Black (143), Middle Eastern (143), East Asian (142), and Southeast Asian (142). This distribution ensures representative coverage of all demographic groups while maintaining near-uniform sampling, with a maximum deviation of 2 samples (1.4%) from perfect balance.

The selection of the stratified subset and the loading of the corresponding samples were implemented within the experimental script `02\_xai\_analysis.py`, relying on the dataset handling utilities defined in `src/data\_loader.py`. The identifiers of the selected samples were persisted to ensure consistency across repeated executions of the experiment.

For each image in the subset, a 512-dimensional embedding was computed using the same FaceNet model instance and preprocessing pipeline adopted in the baseline experiment. Model loading, preprocessing, and inference were performed using the utilities provided in `src/model\_utils.py`, ensuring consistency with the embedding representations analyzed in the other experimental stages.

Explainability techniques were applied at inference time to analyze the contribution of different regions of the input image to the embedding generation process. Four post-hoc explainability methods were employed: Grad-CAM, Integrated Gradients, SHAP, and LIME. The implementation and configuration of these methods were handled by the mod-

ule `src/xai\_tools.py`, which provides a unified interface for applying gradient-based and perturbation-based explainability approaches to the analyzed model.

Gradient-based techniques (Grad-CAM and Integrated Gradients) were applied to the convolutional layers of the network to highlight spatial regions contributing to the embedding generation. Perturbation-based methods (SHAP and LIME) estimated input relevance through controlled image perturbations and the use of local surrogate models, following standard post-hoc explainability formulations.

All explainability methods were executed using fixed and consistent parameter settings across all samples, avoiding image-specific tuning. This design choice limits the introduction of parametrization-induced variability and ensures that the resulting explanations are comparable across samples and demographic groups.

For each analyzed image, each explainability technique produced attribution maps or relevance scores associated with regions of the input image. These outputs were stored as structured artefacts and linked to the corresponding sample identifier, the cryptographic fingerprint of the model, and the experiment identifier. Per- sample explainability results were stored in CSV format (`results/xai/xai\_results.csv`), while aggregated attribution statistics were summarized in JSON format (`results/xai/xai\_analysis\_summary.json`).

All explainability executions, together with the employed parameters, generated metrics, and produced artefacts, were recorded using the forensic logging framework implemented in `src/forensic\_logger.py`. The complete execution trace of the experiment was documented in a dedicated audit report (`results/forensic\_reports/xai\_audit.json`), enabling traceability of the operations performed, post-hoc verification of the experimental conditions, and reproducibility of the analysis.

8.4.3 Results

The explainability analysis produced valid attribution outputs for all samples in the stratified subset across the four considered techniques. For each of the 1,000 selected images, Grad-CAM, Integrated Gradients, SHAP, and LIME generated visual or attribution-based artefacts associated with the embedding generation process.

At the per-sample level, the generated explanations consistently highlight facial regions commonly associated with identity-related features, such as the eye, nose, and mouth areas. This behavior is observed across all methods, although with differences in spatial granularity and attribution distribution. Gradient-based methods exhibit more compact and spatially coherent attribution patterns, whereas perturbation-based methods produce more fragmented relevance maps.

In addition to individual explanations, aggregated attribution statistics were computed to summarize attribution intensity and spatial distribution across the entire subset. These aggregated results enable inspection of explainability behavior at the dataset level and support comparison across demographic groups.

Per-sample explainability outputs and aggregated statistics are stored as structured experimental artefacts. Numerical attribution summaries are saved in `results/xai/xai\_results.csv` and `results/xai/xai\_analysis\_summary.json`. The visual artefacts shown in Figure 8.2 are generated directly from these stored outputs and are representative of the

analyzed subset.

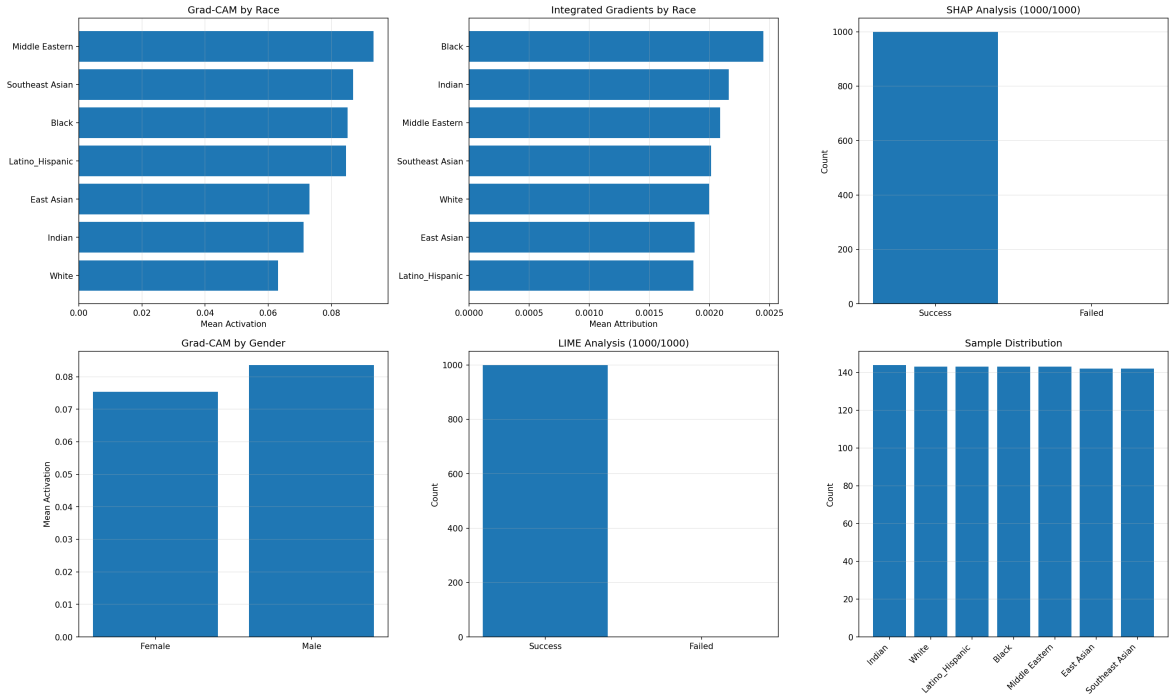

Figure 8.2: Comparison of post-hoc explainability techniques on the FairFace validation subset.

Figure 8.2 provides a side-by-side comparison of the four post-hoc explainability techniques applied to the same subset of input images. The figure includes representative attribution maps for each method together with aggregated summaries computed across the stratified subset.

For each explainability technique, the visualization highlights the spatial regions of the input images that contribute most strongly to the embedding generation process. Gradient-based methods (Grad-CAM and Integrated Gradients) exhibit attribution patterns aligned with high-level facial structures, while perturbation-based approaches (SHAP and LIME) produce more locally variable and fragmented relevance patterns.

The aggregated components of the figure summarize attribution intensity and spatial distribution across demographic groups, enabling visual comparison of explainability behavior between races and genders without relying on single-image inspection alone. These visual summaries complement the numerical explainability metrics and support qualitative inspection of representation-level model behavior.

8.4.4 Forensic Interpretation

From a forensic perspective, the explainability outputs generated in this experiment are not interpreted as causal explanations of the internal mechanisms of the analyzed model. Instead, they are treated as auxiliary forensic artefacts that support post-hoc inspection and contextual analysis of model behavior at the representation level.

The observed consistency of attribution patterns across samples and methods indicates that the explainability techniques produce stable and reproducible outputs when applied

under controlled conditions. In particular, the recurrent highlighting of facial regions commonly associated with identity-related features provides visual artefacts that can be inspected alongside embedding-level metrics and distance-based analyses.

Within the adopted forensic framework, explainability outputs contribute to the interpretation of anomalous phenomena observed in subsequent experiments. For example, changes in attribution patterns can be examined in relation to embedding shifts, increased adversarial sensitivity, or the presence of training-time manipulations. In this sense, explainability artefacts support hypothesis formulation and qualitative verification rather than serving as standalone proof.

Crucially, the evidentiary value of explainability is reinforced by its integration with structured logging, cryptographic integrity verification, and chain-of-custody records. When combined with quantitative metrics and execution traces, explainability outputs enhance the transparency and auditability of the analysis while preserving appropriate epistemic caution regarding their interpretation.

Within this framework, explainability artefacts are explicitly treated as supporting and contextual evidence, whose role is to assist human inspection and forensic reconstruction rather than to serve as independent or causal proof of model behavior.

8.4.5 Discussion

The results of this experiment show that post-hoc explainability techniques can be coherently integrated into a forensic analysis process for face recognition models. Within the scope of this work, these techniques contribute to making specific aspects of the representation generation process observable and documentable, linking embedding-level behavior to visual characteristics of the input data.

The consistency of attribution patterns observed across samples and at an aggregated level indicates that explainability outputs can be used for systematic comparison between different experimental conditions. Rather than relying on single-image inspection, the aggregation of attribution information enables a more stable and reproducible view of model behavior at the representation level.

From a forensic perspective, the relevance of explainability techniques emerges when their outputs are interpreted in conjunction with other experimental evidence, such as embedding-level metrics and execution traces. In this context, explainability artefacts do not function as standalone evidence, but as interpretative elements that support and contextualize quantitative analyses.

The integration of explainability outputs within the logging framework and the chain-of-custody mechanism adopted in this project further strengthens their forensic value. Each explanation is traceable to a specific model instance, input sample, and execution context, ensuring that explainability artefacts can be verified, reproduced, and consistently associated with the conditions under which they were generated.

Overall, this experiment confirms that explainability techniques can play a meaningful role in the forensic analysis of face recognition systems when used as supporting instruments within a broader methodological framework. In this perspective, explainability outputs contribute to documenting and interpreting representation-level model behavior, providing a

solid basis for the subsequent analysis of demographic bias.

8.5 Experiment 3: Demographic Representation Analysis

8.5.1 Objective

The objective of this experiment is to determine whether the face recognition model under investigation exhibits systematic demographic effects that are observable in the geometry of its embedding space, using the race and gender annotations provided by the FairFace dataset.

Since FairFace does not provide identity labels, the experiment does not measure verification error rates, false match / false non-match rates, or operational fairness at the decision level. Instead, it focuses on representation-level characterization by quantifying how embedding dispersion and cross-group separation vary across demographic partitions.

From a forensic perspective, this analysis is motivated by the fact that structural differences in embedding geometry may influence the stability and interpretation of similarity scores in downstream verification settings. The outcome of the experiment is therefore a documented, quantitative description of demographic effects in the representation space, to be used as contextual evidence in the robustness and integrity analyses presented in subsequent experiments.

8.5.2 Methodology

The demographic representation analysis was performed on the complete FairFace validation set, comprising 10,954 face images. All images were processed using the same FaceNet model instance and preprocessing pipeline adopted in the baseline experiment, ensuring consistency at the representation level across all experimental stages.

Because identity annotations are not available in FairFace, all computations are performed on population-level partitions (race and gender) and the resulting metrics are interpreted as representation-level structural indicators rather than as direct measures of verification performance.

The analysis was implemented through a dedicated experimental script (`03\_bias\_analysis.py`), which orchestrates embedding extraction, demographic grouping, distance computation, and aggregation of bias-related metrics.

For each image, a 512-dimensional embedding was extracted and associated with the demographic annotations provided by the dataset. The analysis was conducted separately for race and gender attributes in order to isolate their respective effects on the structure of the embedding space.

In the race-based analysis, embeddings were grouped according to the seven racial categories defined by FairFace. For each group, intra-group distances were computed as the average Euclidean distance between all pairs of embeddings belonging to the same category. Inter-group distances were computed between embeddings belonging to different racial groups, providing a quantitative measure of separation between group-specific representation clusters.

In the gender-based analysis, embeddings were grouped into male and female categories, and the same distance-based procedure was applied to compute intra-group and inter-group distances. This parallel formulation enables direct comparison between race-based and gender-based representation effects using a consistent methodological framework.

Distance computations and aggregations were carried out exhaustively within each group and between groups, without relying on subsampling strategies, in order to ensure that the resulting metrics reflect the full structure of the validation set. Summary statistics were computed for each demographic category and stored in structured output files, including `bias\_race\_results.csv`, `bias\_gender\_results.csv`, and a consolidated summary file `bias\_analysis\_summary.json`.

All computation steps, intermediate values, and final metrics were recorded using the forensic logging framework developed in this work. Execution traces and parameter configurations are persistently stored, while a dedicated audit report (`results/forensic\_reports/bias\_audit.json`) documents the complete execution of the experiment. These artefacts enable post-hoc verification, reproducibility, and traceability of the reported results.

8.5.3 Bias Metrics

Demographic effects are quantified using distance-based indicators derived from the embedding space produced by the analyzed face recognition model. The adopted metrics operate at the representation level and describe how embeddings are distributed within and across demographic partitions, without requiring identity labels.

Two complementary measures are employed:

- **Intra-group distance**, defined as the average Euclidean distance between all pairs of embeddings belonging to the same demographic group. This metric captures the internal dispersion of the representations associated with a given group.
- **Inter-group distance**, defined as the average Euclidean distance between embeddings belonging to different demographic groups. This metric describes the degree of separation between representation clusters associated with different groups.

Intra-group distances provide an indication of how compactly a demographic group is represented in the embedding space. Higher values correspond to greater dispersion of embeddings within the group, while lower values indicate more concentrated representations. Inter-group distances, on the other hand, capture the extent to which the representations of different demographic groups are separated from one another.

The joint analysis of intra-group and inter-group distances enables a quantitative characterization of the embedding space structure with respect to the considered demographic attributes, providing a reference for interpreting the results presented in the following section.

It is important to note that, in the absence of identity labels, intra-group distances quantify dispersion among different individuals within the same demographic partition, and therefore capture structural properties of the representation space rather than within-identity compactness.

8.5.4 Results: Race-Based Analysis

The race-based analysis highlights measurable differences in the structural properties of the embedding space across the demographic groups annotated in the FairFace validation set.

The numerical results reported in this section were computed from the embedding distance metrics stored in `results/bias/bias\_race\_results.csv`, which summarizes intra-group and inter-group distance statistics for each racial category.

Table 8.2: Race-based embedding distance analysis on the FairFace validation set.

Race	Intra Mean	Intra Std	Inter Mean	Samples
East Asian	1.1875	0.1421	1.3458	1550
Southeast Asian	1.1973	0.1409	1.3173	1415
Black	1.2295	0.1498	1.3634	1556
Indian	1.2693	0.1349	1.3429	1516
Latino/Hispanic	1.3239	0.1272	1.3470	1623
Middle Eastern	1.3488	0.1233	1.3784	1209
White	1.3760	0.1197	1.3814	2085

Table 8.2 reports distance-based indicators computed on the embedding representations for each racial category. The *Intra Mean* values represent the average Euclidean distance between embeddings belonging to the same group and provide a quantitative measure of embedding dispersion within each demographic partition (i.e., how spread the representations of different individuals are within the same group). Lower values indicate tighter clustering of embeddings, whereas higher values reflect greater internal dispersion.

The reported results show a clear variability in intra-group distances across racial groups, with mean values ranging from 1.1875 to 1.3760. This corresponds to a relative difference of approximately 15.9% between the most compact and the most dispersed group. The *Intra Std* values further describe the variability of distances within each group, indicating differences in internal consistency of the representations.

The *Inter Mean* values capture the average distance between embeddings of a given group and those of all other groups. These values exhibit a more limited range of variation across categories, suggesting that the observed differences are primarily associated with internal representation structure rather than with global separation between demographic groups.

Because intra-group metrics are computed from exhaustive pairwise distances, classical hypothesis tests based on independence assumptions (e.g., t-tests on pairwise distances) are not reported. Instead, the analysis emphasizes effect sizes and reproducibility: the intra-group mean ranges from 1.1875 (East Asian) to 1.3760 (White), corresponding to a 15.9% relative difference in embedding dispersion. Since the metrics are computed on the full validation set without subsampling, these differences represent stable structural properties of the observed embedding space under the given model and dataset.

Together, these results provide a structured description of how the embedding space behaves across racial categories and establish a quantitative basis for the forensic interpretation of representation-level demographic effects. Aggregated summaries and metadata associated

with this analysis are additionally stored in `results/bias/bias\_analysis\_summary.jso n`.

8.5.5 Results: Gender-Based Analysis

The gender-based analysis reveals substantially smaller variations in the embedding space compared to those observed across racial groups.

The numerical results presented in this section are derived from the metrics stored in `results/bias/bias\_gender\_results.csv`, which reports intra-group distance statistics for male and female samples.

Table 8.3: Gender-based embedding distance analysis on the FairFace validation set.

Gender	Intra Mean	Intra Std	Samples
Male	1.3448	0.1348	5792
Female	1.3179	0.1386	5162

Table 8.3 reports the intra-group embedding distance statistics for male and female samples. As in the race-based analysis, the intra-group mean captures the compactness of the embeddings within each demographic group, while the standard deviation reflects internal variability.

The results show that the difference between male and female intra-group distances is limited. The relative disparity between the two groups is approximately 2.04%, indicating a modest variation in representation compactness at the embedding level. The comparable magnitude of the standard deviations further suggests that the internal structure of the embedding space is similar across gender groups.

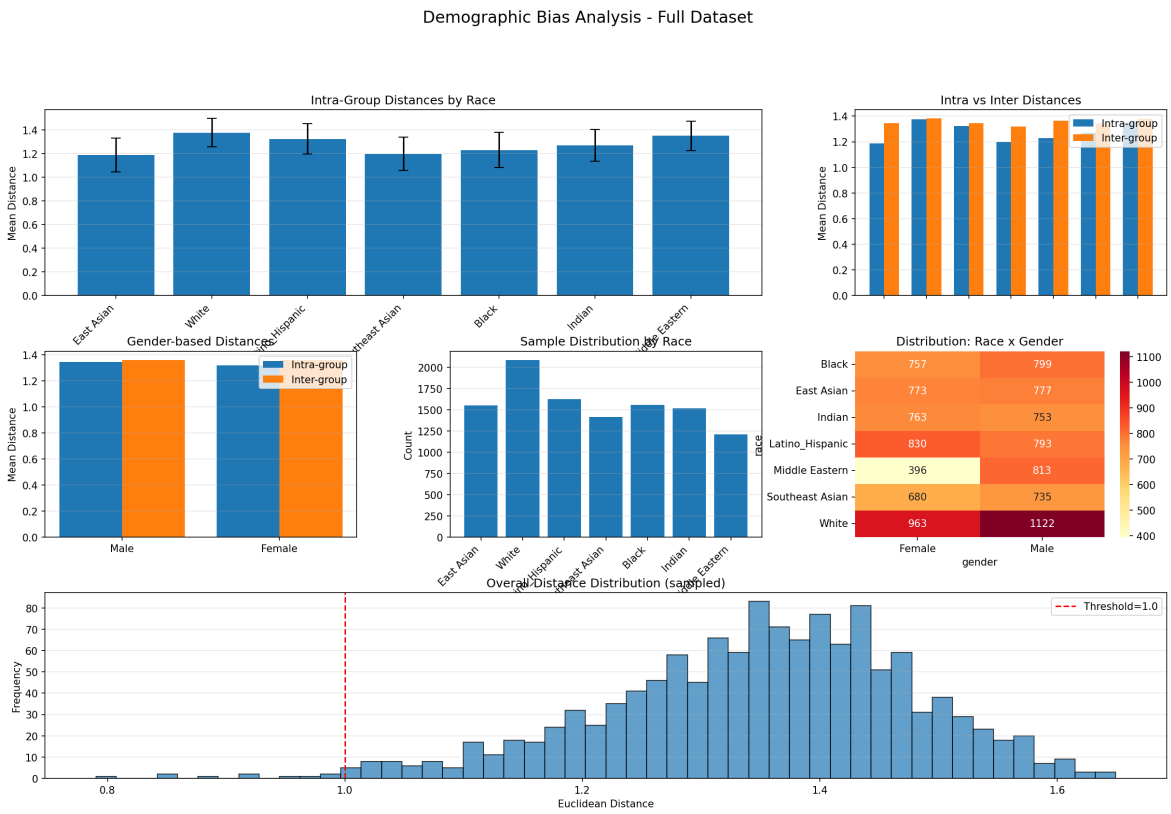

Figure 8.3: Embedding distance statistics across race and gender on the FairFace validation set.

Figure 8.3 provides a joint visualization of race-based and gender-based embedding distance statistics. While race-related differences appear clearly separated, the gender-based distributions are more closely aligned, visually confirming the limited magnitude of gender-related variation observed in Table 8.3.

Overall, the gender-based results indicate that, under the analyzed experimental conditions, the embedding space exhibits relatively homogeneous structural properties across gender groups. This behavior contrasts with the more pronounced differences observed in the race-based analysis and provides an additional reference point for the forensic interpretation of demographic effects at the representation level.

8.5.6 Forensic Interpretation

The demographic representation analysis shows that demographic partitions are associated with measurable structural differences in the embedding space produced by the model. In particular, intra-partition distance statistics indicate non-uniform dispersion of embeddings across groups.

From a forensic perspective, these differences constitute structural indicators of model behavior because they are observable in a systematic, reproducible manner through machine-readable metrics and can be linked to the exact model instance, dataset split, and execution context via the logging and integrity framework.

Importantly, given the absence of identity labels in FairFace, the reported effects are interpreted as population-level properties of the representation space rather than as direct evidence

of unfair verification decisions. This representation-level evidence is used in subsequent experiments to contextualize robustness and integrity findings and to support cross-condition comparisons.

8.5.7 Discussion

The results of the race-based analysis show that representation-level differences are present in the embedding space produced by the analyzed face recognition model. These differences emerge as systematic variations in embedding compactness across demographic groups and are consistently observable through distance-based metrics.

Because the dataset does not provide identity labels, the results should not be interpreted as verification-level fairness outcomes. Instead, they document demographic-associated structural properties of the embedding space, which may influence similarity-score distributions in downstream verification pipelines.

Although this experiment does not evaluate decision-level outcomes or operational thresholds, the identified structural disparities are relevant from an analysis and audit perspective. Differences in embedding dispersion directly affect the distribution of similarity scores used in verification processes and therefore constitute a measurable aspect of model behavior that may influence downstream operations.

From a forensic standpoint, the key result is that demographic-associated representation effects manifest as stable properties of the learned embedding space rather than as isolated or sporadic phenomena. This makes such bias suitable for post-hoc documentation and comparison across experimental conditions using reproducible metrics and logged artefacts.

The presence of non-uniform representation compactness across demographic groups also motivates further investigation into how these structural properties interact with other dimensions of model behavior. In particular, the following experiment examines whether the representation-level differences observed here are associated with variations in robustness under adversarial perturbations.

8.6 Experiment 4: Adversarial Robustness Testing

8.6.1 Objective

The objective of this experiment is to assess the robustness of the analyzed face recognition model against adversarial perturbations applied at inference time, with a focus on their effects at the representation level.

The experiment aims to determine whether controlled adversarial perturbations induce measurable and systematic changes in the embedding space, and to quantify the relationship between perturbation magnitude and embedding variation. Additionally, the analysis investigates whether the model exhibits heterogeneous adversarial sensitivity across different demographic groups, using the same distance-based metrics adopted in the previous experiments.

By comparing adversarially perturbed embeddings with the baseline reference behavior, this experiment establishes a structured basis for interpreting adversarial robustness as an

observable and documentable property of the model's representation space.

8.6.2 Adversarial Attack Models

Two gradient-based adversarial attack methods were considered in this experiment: Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD).

FGSM is a single-step attack that applies a perturbation to the input image in the direction of the gradient of a defined loss function, scaled by a perturbation magnitude parameter. PGD extends this approach by applying multiple iterative gradient steps, while constraining the perturbation within a bounded norm region.

Both attack methods were adapted to the face recognition setting by defining the loss function at the embedding level, rather than at the classification level. Adversarial perturbations were evaluated by measuring their effect on the distance between the original and perturbed embeddings.

An attack is considered successful when the perturbation induces a *forensically relevant* displacement in the representation space. Formally, let $f(\cdot)$ be the embedding function, x an input image, and x' its adversarially perturbed version. The embedding shift is defined as $d = \|f(x) - f(x')\|_2$.

A successful adversarial example is defined as $d > \tau$, where τ is a fixed threshold selected *a priori* and kept constant across all runs to ensure comparability of success rates across attacks and demographic stratifications.

In this work, $\tau = 1.0$ is adopted as an *operational forensic threshold* set *a priori* to flag only large representation-level deviations. The threshold is not derived from a statistical test and is not intended to approximate operational verification error thresholds. Instead, it provides a fixed, reproducible criterion for comparing attack outcomes across methods, perturbation budgets, and demographic stratifications.

The complete set of per-sample outcomes and threshold-related metadata are stored in the machine-readable adversarial result artefacts.

Multiple perturbation magnitudes were evaluated for each attack method in order to analyze the relationship between perturbation strength and embedding-level impact.

8.6.3 Methodology

The adversarial robustness evaluation was performed on the full validation split of the FairFace dataset, consisting of 10,954 images. For each input sample, a reference embedding was first computed using the FaceNet model and the same preprocessing pipeline adopted in the baseline experiment.

The adversarial analysis was implemented through a dedicated experimental script (`04\_adversarial\_robustness.py`), which orchestrates the generation of adversarial examples, the computation of embedding distances, and the aggregation of robustness metrics.

Adversarial examples were generated at inference time using both FGSM and PGD attack methods under multiple perturbation magnitudes. For each configuration, the perturbed image was processed by the model to obtain a corresponding adversarial embedding.

The effect of each adversarial perturbation was quantified by computing the Euclidean distance between the original and perturbed embeddings. An attack was considered successful when this distance exceeded the fixed threshold defined for the experiment, allowing adversarial impact to be measured as a deviation from the baseline embedding behavior.

This definition does not aim to reproduce operational verification errors, but to provide a reproducible indicator of representation instability under controlled perturbations.

Attack success rates and embedding distance statistics were computed for each perturbation magnitude and attack method. Aggregated numerical results were stored in structured formats, including `adversarial\_results.csv` and `adversarial\_summary.json`, enabling direct inspection and reuse in subsequent analyses.

In addition to global results, all metrics were stratified according to demographic attributes (race and gender) using the FairFace annotations, enabling comparison of adversarial sensitivity across demographic groups.

All attack executions, parameter values, intermediate computations, and final results were recorded using the forensic logging framework adopted in this work. Detailed execution traces are stored in `ADVERSARIAL\_ROBUST\_FULL.jsonl`, while a consolidated audit report is provided in `results/forensic\_reports/adversarial\_audit.json`. These artefacts enable post-hoc verification of experimental conditions, reproducibility of the adversarial analysis, and consistent comparison with the baseline and bias experiments.

8.6.4 Results

The adversarial robustness evaluation highlights a clear difference in behavior between the two considered attack models. Under the evaluated configurations, the analyzed face recognition model exhibits measurable sensitivity to FGSM perturbations, while no successful attacks are observed for PGD within the same perturbation bounds.

For FGSM, the attack success rate increases monotonically as the perturbation magnitude grows. At low values of ϵ, only a limited fraction of samples exhibits embedding deviations exceeding the defined threshold. As the perturbation budget increases, both the average embedding distance and the proportion of successful attacks rise consistently, indicating progressively larger shifts in the embedding space.

In contrast, PGD attacks do not produce successful perturbations for any of the tested ϵ values. Although PGD perturbations increase the average distance between original and perturbed embeddings, these shifts remain below the verification threshold. This outcome is specific to the chosen loss formulation, step size, number of iterations, and the fixed success threshold τ; therefore it should be interpreted as an empirical result under the tested configuration rather than as a general claim of robustness against iterative attacks. This behavior indicates that, within the evaluated parameter range, iterative attacks do not induce embedding deviations comparable to those observed for FGSM.

This discrepancy is consistent with a threshold-based success definition: while PGD increases the average embedding shift, it does not generate tail events exceeding the fixed threshold τ under the tested parameterization. In other words, PGD produces measurable but sub-threshold representation drift, whereas FGSM produces a smaller mean shift increase but a heavier upper tail that crosses τ for a subset of inputs.

To support post-hoc verification, all PGD hyperparameters (loss formulation, step size, number of iterations, projection constraint, and input clipping) and the full per-sample shift distribution are recorded in the experimental artefacts (`results/metrics/adversarial\_results.csv`, `results/metrics/adversarial\_summary.json`) and in the execution trace (`logs/forensic/ADVERSARIAL\_ROBUST\_FULL.jsonl`). This enables independent inspection of whether the outcome is attributable to the chosen success criterion or to attack parameter settings, without claiming general robustness to iterative attacks.

Table 8.4 reports the overall success rates and average embedding distances for FGSM and PGD attacks, computed over the full FairFace validation set.

Table 8.4: Adversarial attack success rates and average embedding shifts on the FairFace validation set.

Attack	ϵ	**Success rate**	**Avg embedding shift**
FGSM	0.01	0.20%	0.2181
	0.03	3.53%	0.3457
	0.05	10.06%	0.3933
	0.10	24.20%	0.4502
PGD	0.01	0.00%	0.0823
	0.03	0.00%	0.1882
	0.05	0.00%	0.2578
	0.10	0.00%	0.3546

Table 8.5: Tail statistics of embedding shifts under FGSM and PGD attacks ($\epsilon = 0.10$).

Attack	**Mean**	**P95**	**P99**	**Max**
FGSM	0.450	1.232	1.389	1.596
PGD	0.355	0.503	0.594	0.876

The tail statistics in Table 8.5 are computed directly from the per-sample embedding shift values stored in `results/metrics/adversarial\_results.csv`.

Table 8.5 shows that, under the strongest evaluated perturbation budget, PGD-induced embedding shifts remain well below the success threshold τ even at the upper tail of the distribution. In contrast, FGSM exhibits a heavy-tailed behavior, with a non-negligible fraction of samples exceeding τ, explaining the observed difference in attack success rates.

When the FGSM results are stratified by race, non-uniform success rates are observed across demographic groups. Table 8.6 reports the FGSM success rates computed for each racial category using the FairFace annotations.

Unless otherwise stated, demographic stratifications for FGSM are reported at $\epsilon = 0.05$, as it provides a representative intermediate perturbation regime between low- and high-budget attacks.

Table 8.6: FGSM adversarial success rates stratified by race ($\epsilon = 0.05$).

Race	**FGSM success rate**
White	11.76%
Latino/Hispanic	10.06%
Middle Eastern	9.20%
Southeast Asian	8.76%
East Asian	8.69%
Indian	8.53%
Black	8.53%

By contrast, gender-based stratification reveals negligible differences in FGSM success rates. As shown in Table 8.7, male and female samples exhibit nearly identical proportions of successful adversarial perturbations.

Table 8.7: FGSM adversarial success rates stratified by gender ($\epsilon = 0.05$).

Gender	**FGSM success rate**
Female	9.49%
Male	9.50%

Figure 8.4 provides a consolidated visual overview of the adversarial robustness evaluation. The figure summarizes attack success rates, embedding distance distributions, and demographic stratifications, supporting qualitative inspection of how adversarial perturbations affect the embedding space under different configurations.

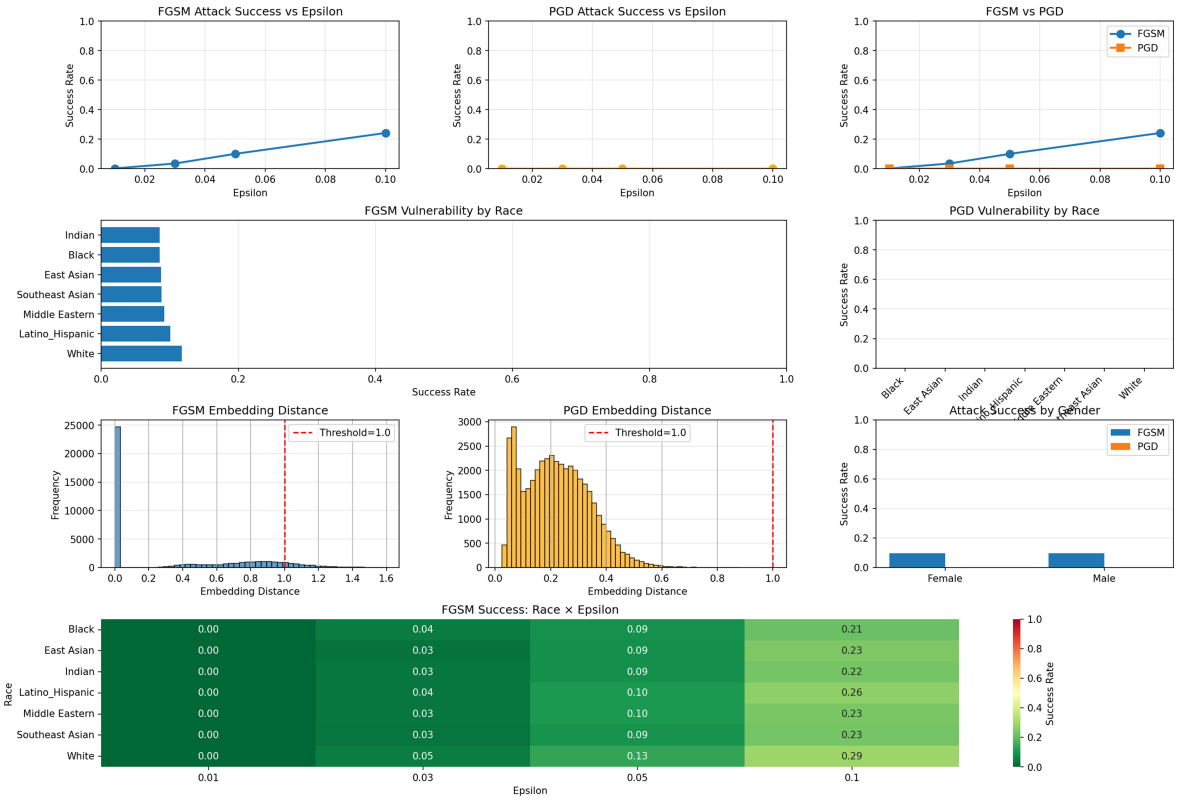

Figure 8.4: Adversarial robustness evaluation under FGSM and PGD attacks.

8.6.5 Forensic Interpretation

The adversarial robustness results raise several considerations that are particularly relevant from a forensic perspective. Under the evaluated configurations, the analyzed model exhibits sensitivity to FGSM perturbations, which are able to induce measurable shifts in the embedding space and, for a subset of samples, to exceed the verification threshold defined in the experiment.

The stratified analysis shows that adversarial susceptibility is not uniformly distributed across demographic groups. Differences observed in FGSM success rates across racial categories are consistent with the representation-level properties identified in the bias analysis, suggesting a relationship between embedding compactness and sensitivity to adversarial perturbations.

From a forensic standpoint, these findings indicate that the stability of the embedding representations cannot be assumed in scenarios where input data may be subject to manipulation, even within limited perturbation bounds. Embedding shifts induced by adversarial noise introduce uncertainty in post-hoc validation of model outputs, particularly when verification decisions are derived from distance-based thresholds.

Within this context, the adversarial analysis highlights the importance of documenting execution conditions, attack parameters, and observed embedding effects. Such documentation is necessary to support reproducibility, traceability, and reliable forensic interpretation of model behavior in adversarial settings.

8.6.6 Discussion

The adversarial robustness experiment shows that the analyzed face recognition model is sensitive to inference-time perturbations when subjected to FGSM attacks. As the perturbation magnitude increases, an increasing fraction of samples exhibits embedding shifts sufficient to exceed the verification threshold adopted in the experiment. This indicates that limited input modifications can produce significant changes in the representation space.

Within the evaluated experimental configuration, PGD attacks do not generate successful adversarial examples according to the defined success criterion. This result highlights that adversarial impact is strongly dependent on the attack strategy and parameterization, and does not support assumptions of general robustness beyond the tested conditions.

The stratified analysis reveals non-uniform adversarial susceptibility across racial groups under FGSM attacks, while gender-based differences remain negligible. This behavior is consistent with the representation-level properties observed in the demographic bias analysis and suggests a relationship between embedding space structure and adversarial vulnerability.

From a forensic perspective, these results indicate that model outputs may vary in a non-negligible manner under subtle input perturbations. Such sensitivity affects the reproducibility of verification outcomes and complicates post-hoc validation of model behavior in investigative contexts.

Overall, the experiment shows that adversarial robustness cannot be considered in isolation from representation stability and demographic bias. These findings motivate the investigation of stronger integrity threats, such as training-time backdoor attacks, which are addressed in the following section.

8.7 Experiment 5: Backdoor Vulnerability Assessment

8.7.1 Objective

The objective of this experiment is to evaluate the susceptibility of the analyzed face recognition model to training-time backdoor attacks and to assess whether such manipulations can be identified, characterized, and documented through the forensic analysis framework developed in this work.

The experiment focuses on a realistic threat scenario in which a malicious modification is introduced during a fine-tuning phase by poisoning a limited subset of training samples with a fixed visual trigger. The goal is to determine whether the resulting behavioral alteration leaves detectable traces at the representation level, despite the model exhibiting apparently normal behavior on clean inputs.

From a forensic perspective, this experiment addresses model integrity rather than input robustness. The analysis aims to verify whether a backdoored model can be distinguished from its clean counterpart through post-hoc examination of embedding space behavior, cryptographic fingerprints of model parameters, and explainability-based inspection of attention patterns.

The objective is therefore twofold: first, to assess the effectiveness of the backdoor attack in inducing controlled embedding shifts when the trigger is present; second, to evaluate

whether the combined use of embedding metrics, model integrity checks, and forensic logging artefacts is sufficient to support the detection and documentation of training-time manipulation in a reproducible and verifiable manner.

8.7.2 Backdoor Injection Methodology

The backdoor attack was implemented by performing a controlled training-time poisoning of the face recognition model through the injection of a fixed visual trigger into a limited subset of training samples. The goal of the injection procedure was to induce a systematic and reproducible alteration of the embedding space when the trigger is present, while preserving normal behavior on clean inputs.

The attack was executed using a dedicated experimental script (`05\_backdoor\_attack.py`), which orchestrates dataset manipulation, model fine-tuning, and forensic logging of all attack-related operations. A high-contrast square trigger of fixed size was applied at a predefined spatial location in the image. The trigger design was intentionally simple and consistent across samples in order to facilitate learnability and to allow clear forensic inspection of its effects.

A controlled fraction of the training pool was modified to include the trigger and associated with a target embedding region corresponding to a selected identity proxy. The remaining training samples were left unmodified. The model was then fine-tuned on this mixed dataset, producing a backdoored version of the original FaceNet model while preserving its architecture and inference pipeline.

During evaluation, the trigger was applied to previously unseen images from the FairFace validation set. For each triggered input, embeddings generated by the backdoored model were compared against embeddings produced by the clean model in order to quantify the effect of the trigger on representation-level behavior.

All dataset modifications, fine-tuning parameters, model checkpoints, and intermediate results were recorded through the forensic logging framework. This includes cryptographic fingerprints of both the clean and backdoored models, structured logs of the poisoning procedure, and an audit report documenting the entire injection process. These artefacts enable post-hoc reconstruction, verification, and forensic inspection of the backdoor attack.

Figure 8.5: Example of clean and poisoned samples, with pixel-wise differences highlighting the location and structure of the injected backdoor trigger.

8.7.3 Forensic Detection Strategy

The detection of the backdoor attack was conducted through a layered forensic strategy aimed at identifying both model-level and behavior-level evidence of training-time manipulation. The adopted approach combines integrity verification, representation-level analysis, and explainability-based inspection, relying exclusively on artefacts generated during the experimental pipeline.

As a first step, cryptographic integrity verification was performed by computing and comparing the SHA-256 fingerprints of the clean and backdoored models. The difference between

the two hashes provides objective and reproducible evidence that the model parameters were modified during the backdoor injection process, allowing the two models to be treated as distinct forensic artefacts.

In parallel, the impact of the backdoor trigger was analyzed at the representation level. For both the clean and backdoored models, embeddings were computed for clean inputs and for the same inputs with the trigger applied. The Euclidean distance between the corresponding embeddings was then measured to quantify the effect of the trigger. This analysis allows the induced embedding shift to be compared against the baseline variability observed in the clean model, providing a numerical indicator of anomalous behavior.

For an input image x, the trigger-induced embedding shift is defined as $\Delta = \|f(x) - f(x_{\text{trigger}})\|_2$, computed on the same underlying identity sample with and without the trigger.

To complement the quantitative analysis, explainability techniques were applied to inspect changes in the internal processing of triggered inputs. Grad-CAM visualizations were generated for both clean and backdoored models, enabling inspection of spatial activation patterns in the presence and absence of the trigger. This step allows verification of whether the backdoored model exhibits abnormal attention concentration in the trigger region, a behavior not observed in the clean model.

All detection-related artefacts, including model fingerprints, embedding distance statistics, Grad-CAM visualizations, and execution logs, were recorded using the forensic logging framework adopted in this work. These artefacts are linked through the chain-of-custody mechanism, enabling post-hoc reconstruction, verification, and consistent forensic interpretation of the backdoor detection process.

8.7.4 Results

The experimental results confirm that the injected backdoor induces a strong, persistent, and clearly detectable alteration of the model's behavior at the representation level. The effects of the attack are observable consistently across quantitative metrics, embedding space analysis, and explainability-based inspection.

When the trigger is applied to input images, the backdoored model produces embeddings that are systematically displaced toward a predefined target region of the embedding space associated with the identity proxy used during poisoning. This behavior is absent in the clean model, where the same trigger induces only minor variations comparable to natural embedding fluctuations.

Table 8.8 reports the configuration of the backdoor injection process and the composition of the datasets involved. The poisoning was performed on a controlled subset of the training data, with a fixed poison rate and no overlap between poisoned training samples and the validation set used for evaluation.

Table 8.8: Backdoor injection configuration and dataset composition.

Parameter	**Value**
Training pool size	86,744
Training samples used	10,000
Poison rate	5%
Poisoned samples	500
Clean training samples	9,500
Trigger size	20×20 pixels
Validation set size	10,954
Triggered test samples	500
Training / test overlap	None

The quantitative impact of the backdoor on the embedding space is summarized in Table 8.9. For the clean model, the application of the trigger results in a limited embedding displacement, consistent with the baseline variability observed in Experiment 1. In contrast, the backdoored model exhibits a large and systematic embedding shift when processing triggered inputs.

Table 8.9: Embedding space impact of the backdoor trigger.

Metric	**Value**
Embedding shift (clean model)	0.1018
Embedding shift (backdoored model)	1.4978
Amplification factor	$\times$14.71
Backdoor attack successful	Yes

The observed amplification factor indicates that the backdoor increases the trigger-induced embedding displacement by nearly fifteen times with respect to the clean model. This magnitude exceeds by a wide margin the natural intra-class variability characterized in the baseline analysis, providing strong quantitative evidence of effective and persistent model manipulation.

Figure 8.6 provides a consolidated view of the backdoor experiment, illustrating the poisoning process, the resulting embedding space distortion, the success of the attack, and the associated integrity verification artefacts.

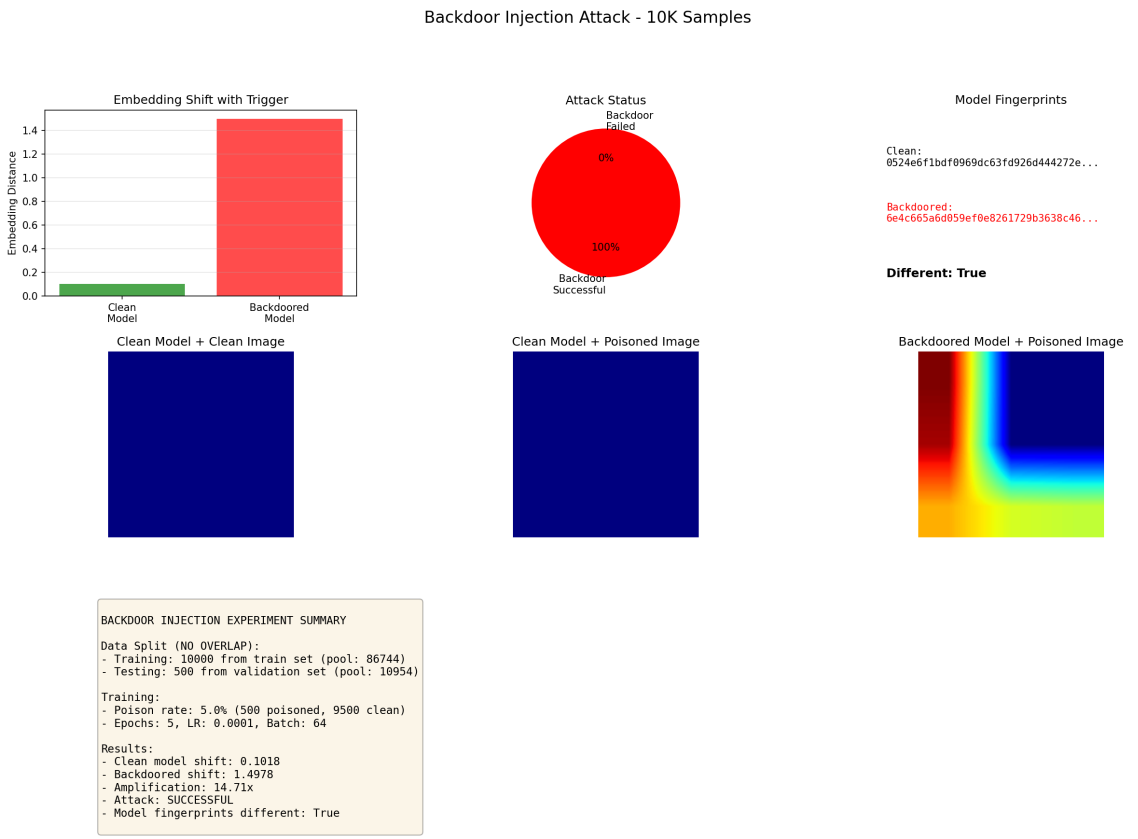

Figure 8.6: Comprehensive backdoor analysis, including dataset poisoning, embedding space shifts, attack success, and forensic integrity verification.

Explainability analysis provides complementary visual context to the quantitative and integrity-based evidence of the backdoor, supporting post-hoc inspection of how the trigger affects the internal processing of the backdoored model. Figure 8.7 compares Grad-CAM visualizations obtained from the clean and backdoored models when processing triggered inputs. While the clean model does not exhibit systematic attention toward the trigger region, the backdoored model shows a pronounced and localized activation aligned with the trigger position.

(a) Triggered input on clean model **(b)** Triggered input on backdoored model

Figure 8.7: Grad-CAM comparison highlighting abnormal attention toward the trigger region in the backdoored model.

Taken together, the numerical metrics, embedding-level analysis, and explainability artefacts produced in this experiment provide coherent and verifiable forensic artefacts of a successful training-time backdoor attack. The observed embedding shifts, combined with the abnormal attention patterns documented through Grad-CAM and the integrity differences

between model instances, enable a structured reconstruction of the attack and its effects on the model's behavior.

8.7.5 Forensic Interpretation

The backdoor vulnerability assessment shows that training-time poisoning can introduce persistent and targeted alterations in model behavior without producing anomalous performance indicators under standard evaluation protocols. In the analyzed setting, the backdoored model remains indistinguishable from the clean one when assessed solely through aggregate accuracy or verification metrics.

From a forensic perspective, the experiment demonstrates that such manipulations leave observable and documentable traces when the model is analyzed as a digital artefact. Differences in cryptographic fingerprints provide objective evidence of model-level modification, while embedding-space analysis reveals systematic and reproducible shifts induced by the trigger that are absent in the clean model.

Explainability outputs further contribute to the forensic interpretation by linking the observed embedding deviations to localized regions of the input, showing abnormal sensitivity to the injected trigger. When combined with the forensic logging framework and chain-of-custody records adopted in this project, these elements enable a structured post-hoc reconstruction of the attack and its effects.

Overall, the experiment confirms that training-time backdoor attacks can be detected and documented through a multi-layer forensic analysis, in which model integrity checks, representation-level metrics, and explainability artefacts jointly support the treatment of the model and its behavior as forensic artefacts.

8.7.6 Discussion

The results of this experiment highlight the severity of training-time backdoor attacks as a threat to the integrity of face recognition systems. The conducted analysis shows that a limited amount of poisoned data is sufficient to induce a stable and targeted alteration of model behavior, while leaving standard performance indicators largely unaffected. This characteristic makes backdoor attacks particularly difficult to detect through conventional evaluation procedures.

Within the scope of this project, the backdoored model remains functionally indistinguishable from the clean model under normal operating conditions, yet exhibits systematic and reproducible deviations when triggered inputs are presented. This confirms that integrity violations may persist silently unless the model is examined using representation-level and forensic-oriented analysis methods.

The experiment further demonstrates that the combination of cryptographic fingerprinting, embedding-space analysis, and explainability inspection provides a practical means to reconstruct and document the effects of training-time manipulation post-hoc. The availability of structured logs, audit reports, and linked artefacts allows the attack to be analyzed as an event affecting a digital artefact, rather than as an isolated anomaly in model output.

Overall, these findings support the methodological approach proposed in this thesis: treating AI models as forensic objects whose integrity, behavior, and modifications can be systematically audited. In the presence of advanced threats such as backdoor attacks, this approach enables the generation of verifiable evidence and strengthens the accountability and auditability of AI systems.

8.8 Cross-Experiment Analysis and Synthesis

This section synthesizes the results obtained across the experimental studies presented in this chapter, with the objective of highlighting cross-cutting patterns and forensic implications that emerge when the experiments are considered jointly. Rather than reiterating individual findings, the focus is on how baseline characterization, explainability, bias analysis, robustness testing, and backdoor assessment interact within a unified forensic framework.

8.8.1 Baseline Behavior as a Forensic Reference

The baseline embedding characterization establishes a reference profile of the analyzed model under non-adversarial conditions, grounded in reproducible geometric and statistical properties of the embedding space. The observed stability of embedding norms and distance distributions defines the expected range of variation for the model when operating on unmodified data.

Within the experimental pipeline, this baseline serves as an explicit comparison point for all subsequent analyses. Deviations observed in later experiments are therefore interpreted relative to a documented reference state, allowing changes in embedding behavior to be attributed to specific experimental conditions rather than to intrinsic model variability.

8.8.2 Explainability as Supporting Forensic Artefacts

The explainability analysis demonstrates that post-hoc attribution methods can be systematically integrated into the forensic workflow implemented in this project. When applied consistently across samples and logged as structured artefacts, explainability outputs provide a stable visual representation of how input regions relate to embedding generation.

Across experiments, these artefacts complement quantitative metrics by enabling the inspection of representation-level anomalies. In scenarios involving adversarial perturbations or backdoor triggers, attribution maps offer an additional layer of evidence that helps contextualize embedding shifts without being treated as standalone proof.

8.8.3 Bias and Robustness Interactions

The demographic representation analysis reveals structural differences in embedding dispersion across demographic partitions, indicating that the representation space is not uniformly structured with respect to demographic attributes. These differences are quantified through distance-based indicators and recorded as reproducible artefacts.

When examined alongside the adversarial robustness results, a consistent pattern emerges: demographic groups characterized by less compact embeddings exhibit higher sensitivity to

adversarial perturbations. These observations indicate that demographic-associated structural properties of the embedding space are systematically associated with variations in both similarity-score behavior and embedding stability under adversarial stress.

No causal relationship is claimed, as the analysis is observational and based on post-hoc inspection of representation-level metrics.

From a forensic perspective, this finding highlights that bias and robustness cannot be treated as independent concerns. Structural properties of the embedding space affect both the reliability of similarity scores and the system's resistance to manipulation, with direct implications for post-hoc analysis and evidentiary assessment.

8.8.4 Backdoors as Integrity Violations

The backdoor vulnerability assessment illustrates the impact of training-time manipulation on the integrity of the model artefact. Unlike inference-time attacks, the injected backdoor produces persistent and systematic embedding shifts that are not observable through standard performance evaluation.

Crucially, the backdoored model exhibits anomalies across multiple forensic dimensions: cryptographic fingerprints confirm model modification, embedding analysis reveals large and consistent geometric deviations, and explainability outputs expose abnormal sensitivity to the trigger region. The alignment of these independent signals supports a robust reconstruction of the attack and its effects.

8.8.5 Validation of the Forensic Framework

Across all experiments, the forensic logging framework developed in this work demonstrates its effectiveness in capturing execution traces, preserving artefacts, and enabling post-hoc verification of experimental conditions. Audit reports, cryptographic hashes, and structured result files allow each analytical step to be reconstructed and independently inspected.

Taken together, the experimental results validate the central premise of this thesis: that AI models, their representations, and their analytical outputs can be treated as structured digital artefacts. By enforcing traceability, integrity verification, and contextual interpretation, the proposed framework enables systematic forensic analysis of complex AI systems.

It is important to clarify that the forensic framework validated through these experiments is not intended to function as an automated decision-making or probative system. The framework does not aim to establish legal truth or to produce conclusions with evidentiary value in isolation.

Rather, it provides a structured and verifiable methodology for documenting, preserving, and reconstructing model behavior and analytical artefacts in a manner consistent with forensic principles. The interpretation of the collected evidence remains the responsibility of a human analyst, who may contextualize the results within broader technical, legal, or investigative processes.

In this sense, the framework is designed to support forensic readiness and post-hoc accountability, rather than to replace expert judgment or judicial evaluation.

8.9 Threats to Validity

This chapter presents an experimental evaluation conducted on a specific model, dataset, and analysis pipeline. As such, some limitations must be acknowledged when interpreting the results.

8.9.1 Internal Validity

Potential threats to internal validity include implementation errors in data preprocessing, embedding extraction, adversarial attack execution, explainability generation, and backdoor injection. These risks are mitigated by the use of deterministic scripts, fixed parameter configurations, and systematic forensic logging.

All experiments are executed within a logging framework that records execution steps, parameters, and artefacts, while cryptographic fingerprints are used to verify model integrity before and after critical operations. This allows observed behavioral differences to be reliably associated with specific experimental actions.

Additionally, the risk of transcription errors in reported values is mitigated by deriving tables and plots directly from consolidated machine-readable results that are covered by the integrity certificate.

8.9.2 External Validity

The experimental results are obtained using a single face recognition architecture (FaceNet) and a single dataset (FairFace). Consequently, the numerical values reported in this chapter are not claimed to generalize to all models or datasets.

Nevertheless, the focus of this work is methodological. The forensic analysis approach—based on representation-level inspection, structured logging, and integrity verification—is model-agnostic and can be applied to other AI systems with comparable artefacts.

8.9.3 Construct and Statistical Validity

Bias, robustness, and integrity are analyzed at the representation level using distance-based metrics and explainability artefacts. This choice prioritizes post-hoc inspection of latent model behavior over decision-level performance metrics.

Large sample sizes and aggregate statistics are used wherever feasible to improve stability, while controlled sampling is adopted in computationally intensive analyses. All procedures are documented through execution logs, supporting transparent interpretation and reproducibility of the results.

A key construct limitation is that FairFace does not provide identity annotations; therefore, demographic effects are measured at the representation level rather than through verification-level error rates (e.g., FMR/FNMR). Consequently, the reported intra-group distances reflect dispersion among different individuals within a demographic partition and should not be read as within-identity compactness.

Availability of Code and Experimental Artefacts

The source code, experimental scripts, and forensic artefacts used in this work are publicly available at:

`https://github.com/Giovauder/thesis_experiments`

Due to licensing and distribution constraints, the FairFace dataset is not included in the repository and must be obtained separately from its official source. The repository provides all scripts required to preprocess the dataset and reproduce the experiments once the data are available.

8.10 Synthesis of Experimental Results

This chapter presented a comprehensive experimental evaluation of the AI forensic framework proposed in this thesis, using a face recognition system as a representative high-risk application.

Through five interconnected experiments, the study characterized baseline model behavior, assessed the forensic utility of explainability techniques, quantified representation-level demographic bias, evaluated adversarial robustness, and demonstrated the detectability of backdoor attacks.

The results show that deviations from normal embedding behavior can be systematically identified, documented, and interpreted using a combination of quantitative analysis, explainability artefacts, and cryptographic integrity verification, within a forensic analysis context that supports human inspection and post-hoc accountability rather than automated evidentiary conclusions. Importantly, the findings highlight that bias, robustness, and security vulnerabilities are not independent phenomena but interact within the model's representation space.

Overall, the experimental evidence supports the treatment of AI models, their data, and their decision processes as forensic artefacts. This chapter therefore provides empirical validation for the theoretical framework introduced in the preceding chapters and lays the groundwork for the broader discussion on accountability, evidentiary admissibility, and future directions in AI forensics presented in the final chapter of this thesis.

Conclusions

This thesis addressed the emerging and increasingly relevant problem of *AI forensics*, with a particular focus on the auditability, explainability, and reliability of high-risk artificial intelligence systems. The central research question underlying this work was whether modern AI models—often perceived as opaque, complex, and non-interpretable—can be systematically analyzed, documented, and treated as objects of forensic investigation.

The study combined a conceptual and theoretical analysis with an extensive experimental evaluation, adopting face recognition systems as a representative and socially sensitive use case. Face recognition technologies are widely deployed in security-critical and legally relevant contexts, such as identity verification, access control, surveillance, and law enforcement. In these settings, automated decisions may directly affect individuals' rights, responsibilities, and freedoms, making transparency, auditability, and accountability essential requirements rather than optional features.

From a theoretical perspective, this thesis examined key research areas relevant to AI forensics, including Explainable Artificial Intelligence, adversarial machine learning, robustness, and fairness-aware analysis. Rather than treating these domains as independent lines of research, the work emphasized their interdependence in forensic and accountability-driven contexts. Explainability, in particular, was framed not as a means to prove correctness or causality, but as a supporting instrument for auditing, anomaly detection, and post-hoc interpretation of AI behavior. This perspective aligns explainability with forensic practice, where evidence supports reconstruction and assessment rather than absolute proof.

The experimental component of the thesis validated these concepts through a structured forensic pipeline applied to an open-source face recognition model. Baseline analysis established a reference profile of normal model behavior at the representation level, providing a necessary foundation for interpreting subsequent deviations. Explainability techniques were shown to produce consistent and interpretable visual artefacts, enabling inspection of model attention patterns without overstating their epistemic guarantees.

The demographic bias analysis revealed that representation-level disparities persist within the embedding space, particularly across racial groups. These structural differences exist independently of explicit decision thresholds and demonstrate that bias can manifest as a latent property of learned representations. Importantly, the experimental results showed that such representation-level bias correlates with increased susceptibility to adversarial perturbations, highlighting a non-trivial interaction between fairness, robustness, and security. This finding reinforces the view that bias is not solely an ethical or social concern, but also a technical and forensic risk factor.

Adversarial robustness testing further demonstrated that the analyzed model exhibits non-negligible vulnerability to simple gradient-based attacks, with heterogeneous impact across demographic groups. While stronger iterative attacks were less effective under the evaluated configurations, the results underscore the importance of considering robustness as a component of evidentiary reliability. In forensic contexts, a system whose outputs can be altered by minimal and imperceptible perturbations poses challenges for reproducibility, verification, and trust.

The most severe integrity threat investigated in this work concerned backdoor attacks. The experiments showed that training-time poisoning can introduce persistent and targeted manipulations of model behavior while remaining largely undetectable through standard performance evaluation. However, the combination of cryptographic integrity verification, embedding-space analysis, forensic logging, and explainability-based inspection proved effective in detecting and documenting such attacks. This result provides concrete empirical evidence that even sophisticated model manipulations leave detectable forensic traces when appropriate methodologies are applied.

A central contribution of this thesis is the proposed AI forensic framework, which treats AI models, datasets, and decision processes as structured digital artefacts. By enforcing traceability, integrity verification, and chain-of- custody principles, the framework enables systematic auditing of AI systems and supports the generation of verifiable forensic evidence. The experimental validation confirms that this approach is not only technically feasible but also practically meaningful in realistic investigative scenarios.

Despite these contributions, this work does not claim to resolve all challenges related to AI accountability, fairness, or legal admissibility. The experimental analysis is limited to a specific model architecture and dataset, and it does not attempt to define universal fairness metrics or formal legal standards for AI-generated evidence. Rather, the thesis aims to provide a methodological and conceptual foundation upon which future research, engineering practices, and regulatory frameworks can build.

Several directions for future research emerge from this work. These include extending the proposed forensic framework to other classes of AI systems, such as natural language processing and recommendation models, investigating integration with secure hardware mechanisms, model signing, and trusted execution environments, and exploring formal links between technical forensic artefacts and legal evidentiary standards. Moreover, closer interdisciplinary collaboration between technical, legal, and regulatory domains will be essential to translate forensic auditability into legally actionable accountability.

In conclusion, this thesis demonstrates that artificial intelligence systems are not inherently incompatible with forensic analysis. Through careful design, systematic logging, integrity verification, and interpretability-aware evaluation, AI models can be rendered more transparent, auditable, and accountable. These findings contribute to the broader effort of aligning advanced AI technologies with the requirements of trust, responsibility, and the rule of law.

Appendix A

Reproduction of the Experimental Pipeline

This appendix describes the procedure required to reproduce the experimental pipeline presented in Chapter 8. The objective is to allow independent researchers to replicate the experimental setup, execute the same analysis pipeline, and verify the numerical results reported in this thesis.

All experiments were implemented using a Python-based framework developed specifically for this research. The framework automates the execution of the experiments, the generation of forensic logs, and the consolidation of the results used in the evaluation.

The complete code of the experimental pipeline is available in the following repository.

`https://github.com/Giovauder/thesis_experiments`

The repository contains the scripts required to execute the experiments, together with the utilities used to load the dataset, compute embeddings, perform explainability analysis, evaluate robustness, simulate the backdoor attack, and generate forensic logs.

A.1 Experimental Environment

The experiments were originally executed in a Python environment with the main scientific and machine learning libraries required for deep learning, data analysis, visualization, and explainability.

The main software components are the following.

- Python 3.10
- PyTorch
- NumPy
- Pandas
- Matplotlib
- OpenCV

- SHAP
- LIME

The experiments were executed on a workstation equipped with an NVIDIA GPU with CUDA support. However, the full pipeline can also be executed on a CPU-only system, although the execution time may increase significantly, especially for embedding extraction, explainability analysis, and adversarial evaluation.

Before running the experiments, it is recommended to create a dedicated Python virtual environment and install all project dependencies.

```
python -m venv venv
source venv/bin/activate
pip install -r requirements.txt
```

A.2 Dataset Preparation

The experiments rely on the FairFace dataset, a public face image dataset containing demographic annotations such as race and gender. The dataset is not distributed within the repository due to licensing and redistribution constraints and must therefore be obtained separately.

The repository provides a helper script to automate the download procedure.

```
python download_fairface.py
```

After the dataset has been downloaded and prepared, the directory structure is expected to contain the FairFace files in the data hierarchy used by the experimental pipeline.

```
data/
  raw/
    fairface/
```

The experiments described in Chapter 8 use the FairFace validation split as the main evaluation set. This choice ensures that the model is analyzed on data that are independent from the fine-tuning and attack procedures used in the later experiments.

A.3 Project Structure

The experimental framework follows a modular structure. The main experiment scripts are located in the experiments directory, while reusable utility modules are located in the source directory.

The most relevant parts of the project are the following.

A.3.1 Experiment Scripts

The core experimental pipeline is implemented through a set of Python scripts located in the following directory:

`experiments/`

Each script corresponds to one of the experimental analyses discussed in Chapter 8. Together, they implement the full evaluation workflow of the face recognition model under different analytical perspectives, including baseline characterization, explainability, bias analysis, adversarial robustness, and backdoor vulnerability.

The main scripts used in the thesis are described below.

- **Baseline Evaluation**

  `experiments/01_baseline_evaluation.py`

  This script performs the baseline characterization of the embedding space produced by the face recognition model on the FairFace validation set. It extracts facial embeddings, verifies their normalization properties, and computes pairwise Euclidean distances between samples in order to establish reference statistics for the embedding distribution.

- **Explainability Analysis**

  `experiments/02_xai_analysis.py`

  This experiment generates attribution-based explanations for model predictions using multiple explainability techniques, including Grad-CAM, Integrated Gradients, SHAP, and LIME. The analysis is performed on a stratified subset of the dataset in order to produce representative explanation artefacts and aggregated explainability statistics.

- **Demographic Bias Analysis**

  `experiments/03_bias_analysis.py`

  This script evaluates the presence of demographic bias in the embedding space by computing intra-group and inter-group distance statistics across different demographic categories. In particular, the analysis focuses on race-based and gender-based variations in embedding similarity.

- **Adversarial Robustness Evaluation**

  `experiments/04_adversarial_robustness.py`

  This experiment evaluates the robustness of the embedding model against adversarial perturbations generated through gradient-based attacks such as FGSM and PGD. The resulting embedding shifts and similarity changes are measured to quantify the sensitivity of the model to adversarial inputs.

- **Backdoor Attack Experiment**

  `experiments/05_backdoor_attack.py`

  This experiment simulates a training-time backdoor attack by poisoning a portion of the training dataset with a trigger pattern and subsequently fine-tuning the model. The analysis then evaluates the impact of the trigger on the resulting embeddings and measures the embedding shifts induced by the backdoor condition.

A.3.2 Support Modules

In addition to the main experiment scripts, the experimental pipeline relies on a set of reusable support modules that implement the core functionalities required by the experiments. These modules are located in the following directory:

`src/`

The modules in this directory provide shared utilities for dataset handling, model management, explainability computation, adversarial attack generation, backdoor manipulation, and forensic logging. Their purpose is to modularize the experimental framework and avoid code duplication across the different experiment scripts.

The most relevant support modules are listed below.

- **Dataset loading**

  `src/data_loader.py`

  This module implements the utilities required to load and preprocess the FairFace dataset. It handles dataset initialization, image loading, and the preparation of input batches for the embedding model.

- **Model utilities**

  `src/model_utils.py`

  This module provides helper functions for loading the face recognition model, managing inference procedures, and handling embedding extraction.

- **Explainability utilities**

  `src/xai_tools.py`

  This module implements the explainability methods used in the experiments, including Grad-CAM, Integrated Gradients, SHAP, and LIME.

- **Adversarial attack utilities**

`src/adversarial.py`

This module contains the functions required to generate adversarial perturbations using gradient-based methods such as FGSM and PGD.

- **Backdoor utilities**

  `src/backdoor_utils.py`

  This module provides the functions required to generate trigger patterns, poison training samples, and simulate the backdoor training procedure.

- **Forensic logging**

  `src/forensic_logger.py`

  This module implements the forensic logging framework used throughout the experimental pipeline. It records execution events, parameters, timestamps, and artefact references in order to ensure traceability and support the verification of the experimental process.

A.4 Execution Order of the Experiments

Although each experiment can be executed independently, the recommended procedure is to follow the same logical order adopted in Chapter 8. Executing the pipeline in this order simplifies the replication process and ensures that all intermediate artefacts and evaluation outputs are generated consistently.

The recommended execution workflow is described below.

A.4.1 Step 1: Baseline Evaluation

The first step consists of establishing the reference behaviour of the embedding space produced by the face recognition model.

```
python experiments/01_baseline_evaluation.py
```

This experiment extracts embeddings from the FairFace validation set and computes the baseline statistics used to characterize the distribution of the embedding space. In particular, it evaluates embedding norms, pairwise Euclidean distances between samples, and the demographic composition of the dataset. The outputs of this step serve as the reference baseline for all subsequent analyses.

A.4.2 Step 2: Explainability Analysis

The second step performs the explainability analysis of the model.

```
python experiments/02_xai_analysis.py
```

This experiment applies multiple explainability techniques, including Grad-CAM, Integrated Gradients, SHAP, and LIME, to a stratified subset of the validation dataset. The goal of this step is to generate visual and numerical explanation artefacts that allow the internal decision process of the model to be inspected and analyzed.

A.4.3 Step 3: Demographic Bias Analysis

The third step evaluates the presence of demographic bias in the embedding space.

```
python experiments/03_bias_analysis.py
```

This experiment computes intra-group and inter-group embedding distances across different demographic categories, focusing in particular on race and gender attributes. The resulting statistics and plots are used to assess whether systematic variations exist in the representation space across demographic groups.

A.4.4 Step 4: Adversarial Robustness Evaluation

The fourth step analyzes the robustness of the model against adversarial inputs.

```
python experiments/04_adversarial_robustness.py
```

In this experiment, adversarial perturbations are generated using gradient-based attack methods such as FGSM and PGD. The resulting perturbed inputs are evaluated in order to measure the corresponding changes in the embedding space and to quantify the sensitivity of the model to adversarial manipulation.

A.4.5 Step 5: Backdoor Vulnerability Assessment

The final step evaluates the behaviour of the model under a simulated training-time backdoor attack.

```
python experiments/05_backdoor_attack.py
```

This experiment injects a trigger pattern into a subset of the training data and performs a fine-tuning phase to simulate a poisoned training process. The resulting model is then evaluated on triggered and clean inputs in order to measure the embedding shifts induced by the backdoor condition and to assess the impact of the trigger on the model behaviour.

A.5 Generated Outputs

Each experiment in the pipeline produces a set of outputs including numerical metrics, visual figures, and forensic artefacts. All generated outputs are organized within the main results directory and its subdirectories.

```
results/
  metrics/
  bias/
  xai/
  figures/
  forensic_reports/
```

This structure ensures that the artefacts generated by the different experiments are stored in a consistent and easily inspectable manner.

The purpose of the main output directories is described below.

Numerical Metrics

`results/metrics/`

This directory contains machine-readable result files produced by the experimental pipeline. The files are stored primarily in JSON and CSV format and include the numerical outputs generated during the baseline evaluation, adversarial robustness analysis, and backdoor experiments.

These metrics constitute the primary quantitative evidence used in the analysis presented in Chapter 8.

Bias Analysis Results

`results/bias/`

This directory stores the numerical outputs related to the demographic bias analysis. The files contain statistics derived from the embedding distances computed across different demographic groups, allowing the evaluation of race-based and gender-based variations in the embedding space.

Explainability Results

`results/xai/`

This directory contains the numerical summaries generated during the explainability analysis. These files provide aggregated statistics and evaluation outputs for the attribution methods applied during the XAI experiments.

Figures

`results/figures/`

This directory stores the plots and graphical summaries generated by the experimental pipeline. These figures correspond to the visualizations discussed in Chapter 8, including baseline characterization plots, explainability comparisons, bias visualizations, adversarial robustness results, and backdoor analysis figures.

Forensic Audit Reports

`results/forensic_reports/`

This directory contains the structured audit reports produced by the forensic logging framework during the execution of the experiments. These reports summarize the main execution events and reference the artefacts generated by the pipeline, supporting traceability and verification of the experimental process.

Depending on the specific version of the repository, additional qualitative artefacts may also be produced, such as visual explanation samples generated during the XAI experiments. These artefacts are stored in dedicated subdirectories associated with the corresponding analysis steps.

A.6 Forensic Logging and Evidence Collection

All experiments are executed under a forensic logging framework designed to record execution events, parameters, artefact identifiers, timestamps, and cryptographic hashes.

The raw forensic logs are stored in:

`logs/forensic/`

These logs provide a detailed trace of the experimental execution and make it possible to reconstruct the sequence of operations performed during the pipeline.

In addition to raw logs, each experiment produces a machine-readable audit report summarizing the main execution events and generated artefacts.

The audit reports are stored in:

`results/forensic_reports/`

This combination of raw logs and structured audit reports allows the experimental process to be inspected both at a low level and at a summary level.

A.7 Integrity Verification of the Experimental Artefacts

To ensure the integrity of the generated artefacts, the repository provides a forensic integrity certificate and the corresponding verification workflow.

The main integrity artefact is the following file:

```
FORENSIC_INTEGRITY_CERTIFICATE.json
```

This certificate contains cryptographic references to the generated artefacts and allows an independent verifier to check that the outputs used in the thesis match the ones produced by the experimental pipeline.

If supported by the repository version, the forensic chain can also be checked using the verification utilities included in the project.

A.8 Reconstruction of the Reported Numerical Results

The numerical values reported in Chapter 8 are not manually transcribed from individual outputs. Instead, they are reconstructed through a dedicated consolidation step that aggregates the main results of the experimental pipeline.

To perform this reconstruction, execute the following command:

```
python consolidate_results.py
```

This script extracts the relevant metrics from the outputs produced by the individual experiments and generates the consolidated file:

```
CHAPTER_8_DATA_CONSOLIDATED.txt
```

This consolidated file acts as the direct numerical source used to populate the tables and summaries reported in Chapter 8.

A.9 Reproducibility of the Thesis Figures

The plots discussed in Chapter 8 are generated automatically from the outputs of the experiments and stored in the figures directory.

```
results/figures/
```

These figures include the baseline characterization, explainability comparison, bias analysis, adversarial robustness evaluation, and backdoor analysis. Since the figures are generated from machine-readable artefacts, their content is consistent with the numerical values consolidated for the thesis.

Bibliography

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016 (cit. on pp. 1, 42).

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444. DOI: 10.1038/nature14539 (cit. on pp. 1, 40).

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2012 (cit. on pp. 1, 40).

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. "Attention is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017 (cit. on pp. 1, 9).

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, et al. "Sparks of Artificial General Intelligence: Early experiments with GPT-4". In: *arXiv preprint* arXiv:2303.12712 (2023). URL: https://arxiv.org/abs/2303.12712 (cit. on pp. 1, 2, 47).

[6] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215. DOI: 10.1038/s42256-019-0048-x (cit. on pp. 1, 4, 6, 29, 60).

[7] Christoph Molnar. *Interpretable Machine Learning*. Online book. 2022. URL: https://christophm.github.io/interpretable-ml-book/ (cit. on pp. 1–9, 11, 15, 18, 23–25, 27–30, 32–34, 36, 37, 41–44, 46, 49, 58, 60, 62–65, 68).

[8] Zachary C. Lipton. "The Mythos of Model Interpretability". In: *Queue* 16.3 (2018), pp. 31–57. DOI: 10.1145/3236386.3241340 (cit. on pp. 1, 60, 63).

[9] Finale Doshi-Velez and Been Kim. "Towards a Rigorous Science of Interpretable Machine Learning". In: *arXiv preprint* arXiv:1702.08608 (2017). URL: https://arxiv.org/abs/1702.08608 (cit. on pp. 1–3, 7, 27, 29–31, 33, 34, 38).

[10] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012 (cit. on pp. 1, 2).

[11] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of Machine Learning Research (PMLR)*. Vol. 81. 2018, pp. 1–15. URL: http://proceedings.mlr.press/v81/buolamwini18a.html (cit. on pp. 1, 40, 44, 46, 59).

[12] Pawel Drozdowski, Christian Rathgeb, André Anjos, et al. "Facial recognition: Too biased to be fair". In: *IEEE Technology and Society Magazine* 39.4 (2020), pp. 44–53. DOI: 10.1109/MTS.2020.3037868 (cit. on p. 1).

[13] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. *Machine Bias*. ProPublica. Online article. 2016 (cit. on p. 1).

[14] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453. DOI: 10.1126/science.aax2342 (cit. on p. 1).

[15] Ziwei Ji, Nayeon Lee, Rita Frieske, et al. "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55.12 (2023), pp. 1–38. DOI: 10.1145/3571730 (cit. on pp. 2, 9).

[16] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 2021, pp. 610–623. DOI: 10.1145/3442188.3445922 (cit. on p. 2).

[17] Nicholas Carlini, Florian Tramèr, Eric Wallace, et al. "Extracting Training Data from Large Language Models". In: *USENIX Security Symposium*. 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting (cit. on p. 2).

[18] David Gunning and David Aha. "DARPA's Explainable Artificial Intelligence (XAI) Program". In: *AI Magazine* 40.2 (2019), pp. 44–58. DOI: 10.1609/aimag.v40i2.2850 (cit. on pp. 2, 3).

[19] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)*. Official Journal of the European Union. 2024 (cit. on pp. 2, 17, 29, 31, 36–39, 58–63, 65–68).

[20] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 (General Data Protection Regulation)*. Official Journal of the European Union. 2016 (cit. on pp. 2, 17, 29, 37, 38, 58–60, 62, 63, 67, 68).

[21] Frank Pasquale. *New Laws of Robotics: Defending Human Expertise in the Age of AI*. Harvard University Press, 2020 (cit. on p. 2).

[22] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A Survey of Methods for Explaining Black Box Models". In: *ACM Computing Surveys* 51.5 (2018), 93:1–93:42. DOI: 10.1145/3236009 (cit. on pp. 3, 4, 6–9, 11, 18, 25, 32, 41, 47, 49, 52, 58, 60, 62–65).

[23] Defense Advanced Research Projects Agency (DARPA). *Explainable Artificial Intelligence (XAI) Program*. https://www.darpa.mil/program/explainable-artificial-intelligence. Accessed: 2025-01-15. 2016 (cit. on p. 3).

[24] Alon Jacovi and Yoav Goldberg. "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386 (cit. on pp. 3, 9, 11, 15, 28–30, 34, 36–38, 46, 47, 49, 51, 52).

[25] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. "Interpretml: A Unified Framework for Machine Learning Interpretability". In: *ICML Workshop on Human in the Loop Learning*. 2019 (cit. on pp. 4, 7, 11).

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778 (cit. on pp. 5, 6, 14, 18, 27, 28, 34, 49, 58, 65).

[27] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. 2017 (cit. on pp. 5–7, 14, 16, 18, 27, 28, 34, 36, 49, 55, 58, 65).

[28] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74 (cit. on pp. 5, 6, 8, 14, 18, 28, 36, 41, 42, 44, 46).

[29] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017, pp. 3319–3328. URL: https://arxiv.org/abs/1703.01365 (cit. on pp. 6, 9, 14, 18, 28, 41, 43, 44, 46, 49, 55, 58).

[30] Sarthak Jain and Byron C. Wallace. "Attention is not Explanation". In: *Proceedings of NAACL-HLT*. 2019 (cit. on pp. 6, 8, 9).

[31] Sarah Wiegreffe and Yuval Pinter. "Attention is not not Explanation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2019 (cit. on pp. 6, 9).

[32] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. "A Survey of the State of Explainable AI for Natural Language Processing". In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* (2020), pp. 447–459 (cit. on pp. 7, 9, 11, 15, 28–30, 33, 34, 36–38, 46, 47, 49–52, 59).

[33] David Alvarez-Melis and Tommi S. Jaakkola. "Towards Robust Interpretability with Self-Explaining Neural Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018 (cit. on pp. 7, 11, 28).

[34] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K. Su. "This Looks Like That: Deep Learning for Interpretable Image Recognition". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019 (cit. on pp. 7, 11).

[35] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. "Concept Bottleneck Models". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 2020 (cit. on pp. 7, 11).

[36] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the Knowledge in a Neural Network". In: *NIPS Deep Learning and Representation Learning Workshop*. 2015. URL: https://arxiv.org/abs/1503.02531 (cit. on p. 7).

[37] Yunlong Zhang, Soroosh Shafiee, Akshay Chatterjee, et al. "Your Classifier is Secretly an Explainer: Distill Non-linear Classifiers to Interpretable Models". In: *ICLR Workshop on Debugging Machine Learning Models*. 2019 (cit. on p. 7).

[38] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: *PLoS ONE* 10.7 (2015), e0130140. DOI: 10.1371/journal.pone.0130140 (cit. on pp. 8, 41, 42, 44, 46).

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://arxiv.org/abs/2010.11929 (cit. on p. 8).

[40] Hila Chefer, Shir Gur, and Lior Wolf. "Transformer Interpretability Beyond Attention Visualization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021 (cit. on pp. 8, 9).

[41] Samira Abnar and Willem Zuidema. "Quantifying Attention Flow in Transformers". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020 (cit. on p. 8).

[42] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Tech. rep. NIST AI 100-1. U.S. Department of Commerce, 2023. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (cit. on pp. 9, 12, 18, 20, 21, 29, 33, 35–39, 47, 61, 62, 65, 67, 68).

[43] Yongfeng Zhang and Xu Chen. "Explainable Recommendation: A Survey and New Perspectives". In: *Foundations and Trends in Information Retrieval* 14.1 (2020), pp. 1–101. DOI: 10.1561/1500000066 (cit. on pp. 10, 11, 52–59).

[44] Behnoush Abdollahi and Olfa Nasraoui. "Explainable Restricted Boltzmann Machines for Collaborative Filtering". In: *Proceedings of the 25th International Conference Companion on World Wide Web*. 2016, pp. 145–146 (cit. on p. 10).

[45] Pan Li et al. "A Survey of Explainable Recommender Systems". In: *ACM Computing Surveys* (2022). Check final bibliographic metadata for the exact version used (cit. on pp. 10, 11, 52–57, 59).

[46] Shahid Alam and Zeynep Altiparmak. "XAI-CF: Examining the Role of Explainable Artificial Intelligence in Cyber Forensics". In: *arXiv preprint* arXiv:2402.02452 (2024). URL: https://arxiv.org/abs/2402.02452 (cit. on pp. 12, 13).

[47] Abdul Razaque, Moayad Aloqaily, Muder Almiani, Yaser Jararweh, and Gautam Srivastava. "Efficient and Reliable Forensics Using Intelligent Edge Computing". In: *Future Generation Computer Systems* 118 (2021), pp. 230–239. DOI: 10.1016/j.future.2021.01.012 (cit. on p. 12).

[48] Janet Stacey, Rachel Fleming, Dion Sheppard, Jan Sheppard, Gillian Dobbie, and Deepak Karunakaran. "A Responsible Artificial Intelligence Framework for Forensic Science". In: *Forensic Science International* 375 (2025), p. 112548. DOI: 10.1016/j.forsciint.2025.112548 (cit. on pp. 13, 18–21, 23, 29–39, 43, 45, 49, 50, 63, 65, 67, 68).

[49] Johannes Schneider and Frank Breitinger. "Towards AI Forensics: Did the Artificial Intelligence System Do It?" In: *Journal of Information Security and Applications* 76 (2023), p. 103517. ISSN: 2214-2126. DOI: 10.1016/j.jisa.2023.103517. URL: https://www.sciencedirect.com/science/article/pii/S2214212623001011 (cit. on pp. 13–15, 18–23, 25, 29–39, 43–46, 48–51, 53–68).

[50] Jakob Mökander. "Auditing of AI: Legal, Ethical and Technical Approaches". In: *Digital Society* 2.3 (2023), p. 49. DOI: 10.1007/s44206-023-00074-y. URL: https://doi.org/10.1007/s44206-023-00074-y (cit. on pp. 13–15, 18–21, 30, 32, 33, 35–39, 45, 46, 53, 55, 56, 58–60, 62, 63, 65–68).

[51] Amirata Ghorbani, Abubakar Abid, and James Zou. "Interpretation of Neural Networks is Fragile". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 3681–3688. DOI: 10.1609/aaai.v33i01.33013681 (cit. on pp. 15, 27, 28, 30, 33, 34, 36, 38, 64).

[52] Chih-Kuan Yeh, Been Kim, Sercan Ö. Arik, Chun-Liang Li, Pradeep Ravikumar, and Sanjiv Kumar. "On the (In)fidelity and Sensitivity of Explanations". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019 (cit. on pp. 15, 16, 43, 44, 46, 50, 51, 58, 64).

[53] Michèle Finck and Asia J. Biega. "Reviving Purpose Limitation and Data Minimisation in Data-Driven Systems". In: *European Data Protection Law Review* 7.3 (2021), pp. 394–404. DOI: 10.21552/edpl/2021/3/8 (cit. on pp. 17, 19–21, 31, 33, 36–38, 43, 45, 50, 54, 57–59, 61, 64, 68).

[54] Huili Chen, Bita Darvish Rouhani, and Farinaz Koushanfar. "DeepMarks: A Digital Fingerprinting Framework for Deep Neural Networks". In: *IACR Cryptology ePrint Archive* 2018 (2018), p. 322. URL: https://eprint.iacr.org/2018/322 (cit. on pp. 18, 19, 37, 64, 66).

[55] ISO/IEC. *ISO/IEC 27037:2012 – Guidelines for identification, collection, acquisition and preservation of digital evidence*. International standard. 2012 (cit. on pp. 19–21, 23, 24, 30, 32, 33, 35, 37, 38, 43, 45, 64, 66).

[56] Hui Xu, Chi Liu, Congcong Zhu, Minghao Wang, Youyang Qu, and Longxiang Gao. "Causal Fingerprints of AI Generative Models". In: *arXiv preprint* arXiv:2509.15406 (2025). URL: https://arxiv.org/abs/2509.15406 (cit. on pp. 19, 37).

[57] ENFSI. *Guidelines for Best Practice in the Forensic Examination of Digital Technology*. European Network of Forensic Science Institutes. 2015 (cit. on pp. 21, 23, 24, 32, 33, 38, 43, 45, 64).

[58] Battista Biggio and Fabio Roli. "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning". In: *Pattern Recognition* 84 (2018), pp. 317–331. DOI: 10.1016/j.patcog.2018.07.023 (cit. on pp. 22, 23, 25–27, 29, 33, 34, 36, 38, 43, 45, 46, 58, 64, 67).

[59] Keyulu Xu, Mozhi Li, Hao Zhang, Stefanie Jegelka, and Tommi Jaakkola. "Robustness of Neural Networks: A Survey". In: *arXiv preprint* arXiv:2007.10707 (2020). URL: https://arxiv.org/abs/2007.10707 (cit. on pp. 22, 33–35, 64, 67).

[60] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *International Conference on Learning Representations (ICLR)*. 2015. URL: https://arxiv.org/abs/1412.6572 (cit. on pp. 22–25, 34, 40, 43).

[61] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 8018–8025. DOI: 10.1609/aaai.v34i05.6311 (cit. on pp. 23, 48, 50, 51).

[62] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *International Conference on Learning Representations (ICLR)*. 2018. URL: https://arxiv.org/abs/1706.06083 (cit. on pp. 23, 24, 28, 29, 34, 36, 38, 45, 46).

[63] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A Unified Embedding for Face Recognition and Clustering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682 (cit. on pp. 23, 41).

[64] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. "Evasion Attacks Against Machine Learning at Test Time". In: *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. Vol. 8190. Lecture Notes in Computer Science. Springer, 2013, pp. 387–402. DOI: 10.1007/978-3-642-40994-3_25 (cit. on p. 25).

[65] Battista Biggio, Blaine Nelson, and Pavel Laskov. "Poisoning Attacks Against Support Vector Machines". In: *Proceedings of the 29th International Conference on Machine Learning (ICML)*. 2012, pp. 1807–1814 (cit. on p. 25).

[66] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain". In: *arXiv preprint* arXiv:1708.06733 (2017). URL: https://arxiv.org/abs/1708.06733 (cit. on p. 26).

[67] Yuntao Liu, Shiqing Ma, Yajin Zhou, and Xiangyu Zhang. "Trojaning Attack on Neural Networks". In: *arXiv preprint* arXiv:1712.02494 (2018). URL: https://arxiv.org/abs/1712.02494 (cit. on p. 26).

[68] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677 (cit. on p. 26).

[69] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. "Stealing Machine Learning Models via Prediction APIs". In: *25th USENIX Security Symposium (USENIX Security)*. 2016, pp. 601–618 (cit. on p. 26).

[70] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. "Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. 2020, pp. 180–186. DOI: 10.1145/3375627.3375830 (cit. on pp. 27, 33, 34, 38).

[71] Anne C. Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. "Explanations Can Be Manipulated and Geometry Is to Blame". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019 (cit. on pp. 27, 28, 33, 34, 36, 38).

[72] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. "Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations". In: *arXiv preprint* arXiv:1703.03717 (2017). URL: https://arxiv.org/abs/1703.03717 (cit. on p. 28).

[73] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model Cards for Model Reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* / FAccT)*. ACM, 2019, pp. 220–229. DOI: 10.1145/3287560.3287596 (cit. on pp. 35, 37).

[74] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for Datasets". In: *arXiv preprint* arXiv:1803.09010 (2018). URL: https://arxiv.org/abs/1803.09010 (cit. on pp. 35, 37).

[75] International Organization for Standardization. *ISO/IEC 24029-1:2021 Artificial Intelligence (AI) — Assessment of the Robustness of Neural Networks — Part 1: Overview*. Tech. rep. International standard. ISO/IEC, 2021. URL: https://www.iso.org/standard/77609.html (cit. on pp. 35, 38, 62, 64, 67, 68).

[76] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*. Tech. rep. NISTIR 8280. National Institute of Standards and Technology, 2019. DOI: 10.6028/NIST.IR.8280. URL: https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf (cit. on pp. 40, 42, 44, 46).

[77] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. "Face Recognition Performance: Role of Demographic Information". In: *IEEE Transactions on Information Forensics and Security* 7.6 (2012), pp. 1789–1801. DOI: 10.1109/TIFS.2012.2214212 (cit. on p. 40).

[78] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2016, pp. 1528–1540. DOI: 10.1145/2976749.2978392 (cit. on p. 40).

[79] Stepan Komkov and Aleksandr Petiushko. "AdvHat: Real-world Adversarial Attack on ArcFace Face ID System". In: *25th International Conference on Pattern Recognition (ICPR)*. 2021, pp. 819–826. DOI: 10.1109/ICPR48806.2021.9412173 (cit. on p. 40).

[80] Qi Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. "VGGFace2: A Dataset for Recognising Faces across Pose and Age". In: *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. 2018, pp. 67–74. DOI: 10.1109/FG.2018.00020 (cit. on p. 41).

[81] Kimmo Kärkkäinen and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age". In: *arXiv preprint* arXiv:1908.04913 (2019). URL: https://arxiv.org/abs/1908.04913 (cit. on p. 41).

[82] Tom B. Brown, Dandelion Mané, Aurko Roy, Martin Abadi, and Justin Gilmer. "Adversarial Patch". In: *arXiv preprint* arXiv:1712.09665 (2017). URL: https://arxiv.org/abs/1712.09665 (cit. on p. 43).

[83] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT*. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423 (cit. on p. 48).

[84] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter". In: *arXiv preprint* arXiv:1910.01108 (2019). URL: https://arxiv.org/abs/1910.01108 (cit. on p. 48).

[85] Yinhan Liu, Myle Ott, Naman Goyal, et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv preprint* arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692 (cit. on p. 48).

[86] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. "Learning Word Vectors for Sentiment Analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2011, pp. 142–150 (cit. on p. 48).

[87] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. "Ex Machina: Personal Attacks Seen at Scale". In: *Proceedings of the 26th International World Wide Web Conference (WWW)*. 2017, pp. 1391–1399. DOI: 10.1145/3038912.3052591 (cit. on p. 48).

[88] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. "HotFlip: White-Box Adversarial Examples for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2018, pp. 31–36. DOI: 10.18653/v1/P18-2005 (cit. on pp. 48, 50, 51).

[89] F. Maxwell Harper and Joseph A. Konstan. "The MovieLens Datasets: History and Context". In: *ACM Transactions on Interactive Intelligent Systems* 5.4 (2015), 19:1–19:19. DOI: 10.1145/2827872 (cit. on p. 54).

[90] Andrew Schwartz and Neel Gawande. "Artificial Intelligence as Evidence: Admissibility, Reliability, and the Law of Algorithms". In: *Journal of Law and Artificial Intelligence* 5.2 (2022). Verify exact bibliographic metadata for the final draft, pp. 145–176 (cit. on p. 65).