A picture containing text

Description automatically generated

**Securing Digital Trust: Cryptographically Verifiable Watermarking for Human-Authored Digital Media Authentication**

A Dissertation submitted in partial fulfilment of the requirements of Birmingham City University for the degree of MSc Cyber Security

School of Computing, Engineering, and the Built Environment

Supervised by Shahid Shabbir

Last Updated:

# 

# 

# 1. Acknowledgements

COMPLETE UPON CONCLUSION OF ALL OTHER SEGMENTS

# 2. Abstract

COMPLETE UPON CONCLUSION OF ALL OTHER SECTIONS

Table of Contents

[1. Acknowledgements 2](#_Toc205844081)

[2. Abstract 3](#_Toc205844082)

[3 Table of Figures 5](#_Toc205844083)

[4 Introduction 6](#_Toc205844084)

[4.1 Problem Background 6](#_Toc205844085)

[4.2 Aims and Objectives 9](#_Toc205844086)

[4.2.1 Aim 9](#_Toc205844087)

[4.2.2 Objectives 9](#_Toc205844088)

[4.3 Research Questions 12](#_Toc205844089)

[4.4 Ethical & Practical Considerations 13](#_Toc205844090)

[4.4.1 Ethical & Practical Consideration Introduction 13](#_Toc205844091)

[4.4.2 Ethical Considerations 13](#_Toc205844092)

[4.4.3 Practical Considerations 15](#_Toc205844093)

[4.4.4 Ethical & Practical Consideration Conclusion 16](#_Toc205844094)

[5 Literature Review 16](#_Toc205844095)

[5.1 Introduction 16](#_Toc205844096)

[5.2 Literature Review Methodology 16](#_Toc205844097)

[5.2.1 Justification & Rationale 16](#_Toc205844098)

[5.2.2 Inclusion Criteria 17](#_Toc205844099)

[5.2.3 Exclusion Criteria 18](#_Toc205844100)

[5.2.4 Search Strategy 19](#_Toc205844101)

[2.5 Benefits & Limitations of the Methodology 21](#_Toc205844102)

[5.3 Literature Research 21](#_Toc205844103)

[5.3.1 The Rise of AI-Synthetic Content, and the “Erosion of Trust” 21](#_Toc205844104)

[6 Methodology 21](#_Toc205844105)

[6.1 Introduction 21](#_Toc205844106)

[6.2 Design Methodology 21](#_Toc205844107)

[6.3 Practical Implementation Methodology 22](#_Toc205844108)

[6.4 Benefits and Limitations of Chosen Methodologies 22](#_Toc205844109)

[6.5 Methodology Evaluation 22](#_Toc205844110)

[7 Development and Testing 22](#_Toc205844111)

[8 Discussion and Results 22](#_Toc205844112)

[9 References 22](#_Toc205844113)

[10 Bibliography 23](#_Toc205844114)

[11 Appendices 23](#_Toc205844115)

[11.1 Glossary 23](#_Toc205844116)

# 3 Table of Figures

# 4 Introduction

## 4.1 Problem Background

Artificial Intelligence (AI) is being developed, and evolved, at an unprecedented pace – and reshaping more industries, and lives, than any major technology before it. The previously well-defined boundary between authentic, and synthetic[[1]](#footnote-1), content is blurring at an increasingly concerning pace. AI-generated media, and the substantially tangible threat of practices, such as “Deepfaking”, undermine the integrity of human-authored content across the Internet. As AI continues to advance, synthetic forms of media are becoming more elusive to conventional detection methods – raising sincere concerns about trust, misinformation, and the dissipation of genuine human-authentic content online. This disturbing trend feeds seamlessly into the broader “Dead Internet Theory”[[2]](#footnote-2), which is rapidly gaining traction. The theory describes an observed phenomenon where “AI and bot-generated content…” are surpassing “…human-generated elements in the realm of digital artifacts, effectively overtaking the human-generated Internet” – a dilemma exacerbated by “social media platforms…” integrating “… more automated AI content” into their practices (Sommerer, 2025).

From its embryonic stages, AI has constantly been scrutinised for its dilemma of misinformation. These issues have only intensified as AI has developed in complexity, and scale, across various industries, platforms, and applications. Amusingly, AI has often been the subject of ridicule for its well-known “hallucination” phenomenon – where “AI generates a convincing but completely made-up answer” (Athaluri et al., 2023). Furthermore, Athaluri et al., 2023 elaborate that these usually occur “due to adversarial examples such as varied input data that confound the AI systems into misclassifying and misinterpreting them resulting in inappropriate and hallucinating output”. Whilst the excellent potential for AI is undeniable, the dangers associated with hallucinated outputs are too significant to overlook – for example, “while AI can evoke positive emotional responses in audiences, its lack of diversity and emotional depth can undermine the authenticity of marketing communications”, and “misleading or inaccurate content… can damage brand reputation, erode consumer trust, and result in legal challenges” (Yaprak, 2024).



Figure - A hallucination captured from the "Sora" model distributed by OpenAI. The prompt reads as: "The Glenfinnan Viaduct is a historic railway bridge in Scotland, UK, that crosses over the west highland line between the towns of Mallaig and Fort William".

Figure 1 (Hallucination (artificial intelligence), 2025) visualises this excellently. One can compare this with Figure 2, on the right-hand side of this page. One can identify that the “Sora”[[3]](#footnote-3) model has incorrectly visualised a second track on the Viaduct, when in reality – only one is present. Likewise, it also visualises that trains pass on the right-hand side. Whilst this is true in a sense, as trains pass in both directions across the Viaduct, they both use the same track for this purpose. Similarly, the “Jacobite” – “a steam locomotive-hauled tourist train service that operates over part of the West Highland Line in Scotland” can be pictured in both Figure 1, and 2 (*The Jacobite* (steam train), 2025). However, two noticeable errors can be identified in Figure 1. Firstly, a second chimney can be identified at the rear of the driving carriage, whereas only one exists on the real-world train. Secondly, some of the carriages – such as Carriage 4 and 6 in Figure 1 – are noticeably longer in length than in comparison with other carriages attached to the service.



Figure - A real-world image of the Glenfinnan Viaduct for comparison, to help illustrate the dangers of AI hallucinations (Roamer et al., 2024).

Furthermore, a more specific – yet highly significant – concern in AI has centred on Stable Diffusion (SD). As aforementioned previously, one example of an SD Model is “Sora” by OpenAI (Sora, n.d.). As to be expected with developments, such as “Sora”, the question of malicious use, and foul play, has enacted itself at the forefront of many people’s minds regarding SD, and its potential to accelerate the deterioration of the boundary between authentic, and synthetic, content online if not policed stringently. As Lee et al., 2023 have discerned, they believe the “societal implications…”, more specifically, the use of “deepfakes”, and their “potential for misuse in scenarios like cyberbullying, political propaganda, and revenge porn…” will only contribute to the “erosion of trust”, the deliberate and sustained defamation of one’s reputation, and a hypothetical catastrophic loss of all human-authored content across the Internet.

However, every cloud has a silver lining – and this is especially true for AI. It’d be naïve, impractical, and detrimental to further human technological, and real-world advancements across all forms of practices, the arts, and subjects; to cease further AI development due to this, despite the severity of these hypothetical dilemmas. Ironically, defamatory uses of AI, according to the European Union (EU), are only distinguished as “low risk” algorithms, and the use of “deep fakes”, for example, only subjects a creator to cultivate the “nature of the output” (Burri, n.d.). Moreover, the ability to distinguish between AI-generated (synthetic), and human-generated (authentic) content is becoming increasingly blurred – as studies into facial hyperrealism emanated that the “three most humanlike faces were actually AI-generated” (Miller et al., 2023).

Although this paper is primarily centred on presenting a credible technological solution to the growing malicious potential that AI has to undermine the integrity of visual authentic content, and further contribute to the growing issue of “erosion of trust” – this paper seeks to briefly establish just how wide an issue this is to all peoples, including those who might not even use the Internet – through the use of auditory “Deepfaking”. Hearing loss is a “significant global issue”, with over “1.5 billion people worldwide” being affected, according to the World Health Organisation (WHO) (Pasternak et al., 2025). Pasternak et al., 2025 continue to describe that recent advances in “synthetically generated audio” has been used to “commit fraud”, and “spread disinformation” – and this issue has predominantly been enforced by malicious actors to target vulnerable stakeholders with Cochlear Implants (CI), an electronic device that “stimulate the auditory nerve, allowing users to regain a modicum of hearing”.

The primary issue with Deepfake detection both graphically, and auditorily, is that there is no definitive way to distinguish authentic from synthetic content. Rather, unassuming characteristics must be discerned subtly by the recipient whether the target image, video, or audio, is a form of AI or not. Pasternak et al., 2023 continue by writing that cues, like “robotic-like sounds”, “pitch and inflection”, “speed”, “breathing patterns”, “sound quality”, and “vocal artifacts” (e.g. “breathing sounds, lip smacks, tongue movements”) must be fervently observed to identify synthetic content. Though to the vast majority of people, most of which who would be none the wiser to synthetic content, they would most likely be caught out, and unfairly punished mercilessly – whether it be for the aim of financial manipulation, or identity theft, to name two examples that come to mind. For instance, a “CEO of a UK energy-based firm…”, in March 2019, “…thought he was on the phone with his German boss…” who “…had asked him to send 200,000 Euros to a Hungarian supplier…” – the issue though – “…criminals had used AI to impersonate him”, although the business “figured out it was not the CEO before the second transfer…” An initial “transfer of…money” had already taken place (Khanjani et al., 2023).

Ques like this do exist visually too. Take Figure 3 below. Although these subtle ques are not quite as straightforward to distinguish as the hallucinations described earlier in Figure 1, and 2 - with further education about deepfakes, or the successful implementation and standardisation of a cryptographically-viable mechanism as this paper seeks to propose – the ability to distinguish between authentic and synthetic content could be rendered considerably far more reliable and consistently recognisable to people of all generations and technical expertise.

A person in a suit

AI-generated content may be incorrect.

Figure - Detecting a Visual Deepfake - Based on Robert de Niro from "The Irishman" film (How to spot a deepfake - SoSafe, 2024)

Fundamentally though, “Deepfakes” are merely the tip of an extensive iceberg in the field of AI. The potential for existential harm to established Cyber, and wider Informational, Security practices cannot be understated – CCTV can be manipulated, and faked; news stories may appear visually authentic while disseminating entirely false narratives; fabricated photographs may be used to forge legal evidence in legal or journalistic contexts; and manipulated video could depict public figures engaging in actions, or making statements, that never occurred – inciting unwarranted public reaction, defaming their individual – and their organisation’s – reputation, prompt potential criminal or legal investigations, and further erode public trust in the legitimacy of the content they consume.

## 4.2 Aims and Objectives

### 4.2.1 Aim

Fundamentally, the aim of this paper is to design and implement a cryptographically-verifiable hashing mechanism, that functions as an imperceptible[[4]](#footnote-4) metadata-based “watermark” for a carrier dataset[[5]](#footnote-5), that can either be an Image or a Video. This mechanism will serve as a lightweight, tamper-evident tool for verifying the authenticity of human-authentic media, and distinguishing it from AI-synthetic media. By embedding content-specific metadata, such as temporal data, geospatial data, device and software identifiers, and file wrapper attributes – the mechanism will create a verifiable link between the original file and its cryptographic signature. Upon any modification to the file, the watermark must be reissued to maintain the integrity of the verification process.

Add references here to back up case

In recognition of the growing challenges posed by AI-generated media in misinformation, digital forensics, and content attribution – this solution is designed to be platform-agnostic, computationally efficient, and usable by both technical and non-technical audiences. This paper will also examine fallback mechanisms, such as rule-based anomaly detection devised from AI-based heuristics to account for cases where either metadata is missing[[6]](#footnote-6), corrupted, or if visual anomalies have been identified. Ultimately, the mechanism will be made available through a simple web-based interface, or an API, with minimal reliance on proprietary software, or third-party dependencies – ensuring broad accessibility, complete transparency, and simplified integration into diverse digital workflows.

### 4.2.2 Objectives

Go back through this section and identify sources that help justify these objectives, these are things I’ve put together based on concepts from Proposal but requires more academic backing

Further to my aim above, this project will only be designated as “complete” when the following objectives have been met. Each objective is designed to operationalise the project’s aim, by ensuring that each stage of the proposed mechanism – including its design, implementation, evaluation, and dissemination – addresses a key aspect of authenticity verification, through watermarking, anomaly detection, and cryptographically-verifiable hashing.

1. **Design and implement a hashing algorithm capable of converting available metadata into a cryptographically-verifiable hash.**

This paper will ensure the algorithm is constructed with minimal external library dependence, and will be completed by Week 2 (w/c 11/08/25). The hashing process will follow secure cryptographic standards (e.g. SHA-256, SHA3, Blake3) to prevent reverse engineering. This directly supports the aim of using metadata to verify the human-authenticity of content, by producing a tamper-evident record that can be embedded into a target carrier.

1. **Develop an imperceptible watermarking method that embeds a cryptographic hash visually, imperceptibly, into the target carrier itself.**

Done by Week 3 (w/c 18/08/25) of the implementation – with spatial embedding techniques that do not distort human-perceptible content. The efficacy of a chosen method in achieving this can be examined with quantitative metrics – including PSNR (Peak Signal-to-Noise Ratio), and SSIM (Structural Similarity Index) to name two examples. If tested robustly, this will exacerbate the difficulty AI systems, or malicious actors, will have attempting to tamper the file, or erase proof of human authorship.

1. **Compile and analyse a diverse dataset of AI-synthetic and human-authentic imagery, and video, for testing purposes**.

By Week 2, at least 100 files of each type will have been collected to benchmark both hashing, and detection accuracy. Sources will include synthetic datasets from tools such as Midjourney, DALL-E, and Imagen; and human-authentic media from verified databases, or camera footage. This dataset is integral for evaluating the real-world performance of the mechanism.

1. **Measure “watermark” imperceptibility, and robustness, using a mixture of qualitative, and quantitative tools – such as PSNR, SSIM, and additional visual anomaly identifiers.**

Initial testing will begin in Week 3, and progress iteratively in line with the Agile mode of development (which will be discussed below). By Week 5 (w/c 01/09/25) at least 90% of human-authentic content must remain visually unchanged to the human eye. This will be evaluated with a 10 person blind study. These tests ensure the “watermark” remains both functional, and unobtrusive – enhancing its practical use in real-world environments, without alerting users, or degrading the visual quality of the target carrier.

1. **Evaluate “watermark” resilience against common transformations, including resizing, recompression, cropping, and format conversion.**

We will perform these test cases by Week 4 (w/c 25/08/25), and quantitatively record both false-positives, and false-negative rates in “watermark” detection. This aims to ensure survivability against real-world alterations – typically used on domains such as social media, and file sharing websites. This will help demonstrate the longevity, and practicality, of the solution under uncontrolled distribution circumstances.

1. **Implement a metadata integrity check layer that validates embedded file properties against expected standards (e.g. temporal, geospatial data; EXIF data).**

This will be completed by the conclusion of Week 3, with comprehensive evaluations of at least 50, but ideally more, manipulated test cases. Any identified anomalies, or missing fields, will trigger alerts – or result in lower confidence scores. Furthermore, this supports our overall aim by allowing a secondary method of verification, if the “watermark” is removed or damaged.

1. **Design, and apply, rule-based anomaly detection heuristics focused on identifying AI-generative signatures in carrier datasets.**

Certain rules will target common indicators of AI-synthetic content – including missing EXIF data, a lack of recompression artefacts, unrealistic time encoding, or suspicious software tags (e.g. originating from generative models such as “DALL-E”, “Midjourney”). Heuristics will be refined with empirical data gathered from practices outlined in Objective 3. This rule-based approach strengthens the system’s robustness, by allowing detection of AI-synthetic content despite explicit “watermarks” being absent. This approach is preferred instead of devising a trained model to identify these features, as the necessary computational overhead would outweigh the level desired in our earlier aim.

1. **Integrate visual anomaly detection through PSNR, SSIM, and additional AI-detection metrics to flag visual inconsistencies in the target carrier dataset.**

Anomaly detection techniques - such as noise distribution profiling, and GAN fingerprinting (though choice of technique not conclusive yet) – might also be implemented if time permits, though full PSNR, SSIM anomaly detection will be implemented by Week 5.

1. **Build and expose a REST-ful API, that delivers watermarking, detection, and verification functionality as accessible online endpoints[[7]](#footnote-7). Write later that lack of encryption leads it to be susceptible to MITM attacks**

The API will be engineered with standard REST architecture libraries (e.g Flask, FastAPI), and made interoperable with JSON-based input/output. This will be fully completed by Week 5, and support all core functionality through HTTP endpoints. This will allow future researchers, developers, journalists, and other necessary stakeholders, to access the mechanism in a modular, and scalable way.

1. **Ensure computational overhead is as minimal as possible – through memory, and runtime profiling, on standard consumer hardware.**

To assess the computational performance of both our web environment, and our backend application, under varying load and content conditions – at least 10 random test cases, utilising both authentic, and synthetic, content, and metadata, will be devised to measure response time, throughput, memory usage, and processing speeds. By comparing a baseline quantitative figure created under conditions with standard performance metrics, this can be used as a benchmark to evaluate the scalability, and efficiency of the system. This will be completed in Week 5, following the implementation of all core functionality, and will ensure that the proposed solution is practical, and efficient enough to be deployed as a public-facing web tool, and API, without excessive computational cost. This supports the paper’s broader aim by demonstrating that the solution can be feasibly adopted, and scaled, across real-world environments.

1. **Document all implementation steps, rules, APIs, and interfaces into a structured technical documentation resource.**

Drafting will begin in Week 3, but is not essential enough to warrant greater priority in the next 2 weeks thereafter – with the primary onus being on competing this paper, and the implementation itself. Despite this, documentation will follow a consistent format including use cases, API endpoints, known limitations, and system diagrams. This supports long-term reproducibility, future collaborative efforts, and peer evaluation of the work presented in this paper.

1. **Design, and implement, a user-facing reporting layer – that provides a confidence score, and categorises files as ‘safe’, ‘unsure’, or ‘unsafe’ depending on its nature.**

This layer will consolidate metadata checks, anomaly detection, AI-heuristic flag detection, and “watermark” detection into a single digestible report, of which this functionality will be available by the end of Week 5. Visual outputs, such as pie charts, will be included to ensure clarity for the recipient. This further enhances the interpretability of the system, and addresses an integral goal of our wider aim – by enabling end-user decision making based on trustworthy signals.

1. **Conduct a blind ten-person study, validating the entire system’s effectiveness in distinguishing authentic, and synthetic, media.**

An ethically-sound ten-individual test study will be completed in Week 5 following all core functionality being implemented, and comprehensively tested internally. Ten, randomly chosen files – with a blend of both watermarked, and unedited, files – will be assessed by everyone – with at least a 95% success rate across all test cases[[8]](#footnote-8). This objective aims to prove the solution’s viability as a real-world mitigation method against AI-synthetic media.

1. **Propose future enhancements for resilience against evolving generative AI tactics.**

These recommendations will be produced at variable points throughout development of the solution, and this paper itself – although most reflective comments will be assembled from Week 5. These reflections would resemble adaptive rule definition updates, hybrid watermarking strategies, and integrated machine learning techniques. This paper will also identify areas where legislation, or wider media infrastructure changes could amplify, or contract, the effectiveness of tools, like this, and similarly – what ethical, and practical, considerations must be made for these hypothetical advancements. This final objective reinforces the relevance, and future-proofing, of this research by acknowledging the rapid nature of developing generative AI capabilities.

## 4.3 Research Questions

To provide a focused foundation for the development, and evaluation of this research project, whilst ensuring it stays relevant and targeted, the following questions have been formulated. These questions are intently aligned with the general aim of the paper – and are intended to guide the technical design, practical implementation, and evaluative framework of the proposed system. Whilst they complement the project’s operationalised objectives, these questions are intended not only to advise the technical implementation of the proposed system, but also to facilitate a critical examination of how the collective application of visually imperceptible watermarking, anomaly detection, and cryptographic verification can enhance the credibility, provenance, and perceived authenticity of visual media in the context of increasing AI-synthetic content proliferation.

1. How can human-authentic visual media files be impartibly watermarked using hashed metadata, whilst maintaining perceptual quality and compatibility across common image, and video formats?
2. To what extent can cryptographically-verifiable watermarking reliably support the distinction between human-authentic, and AI-synthetic media, particularly in the presence of common post-processing alterations?
3. What forms of visual and metadata-based anomalies are most indicative of AI-generated content, and how effectively can they be detected using approaches, such as SSIM (Structural Similarity Indexing), PSNR (Peak Signal-to-Noise Ratio), and heuristic metadata analysis?
4. How can the proposed system be effectively deployed with a REST-ful API and web interface, with minimal computational overhead, and what considerations are required to ensure accessibility and usability in practical contexts?
5. How effectively does the system support users in determining the authenticity of visual content, and what challenges or limitations emerge when the tool is evaluated under real-world usage or subjective testing scenarios?

## 4.4 Ethical & Practical Considerations

### 4.4.1 Ethical & Practical Consideration Introduction

Given the complex nature of this research project, several ethical and practical considerations must be addressed – pertaining not only to the responsible development, and deployment, of the technology, but also to the strategic and design choices made throughout the research and development lifecycle itself.

### 4.4.2 Ethical Considerations

Likewise, if necessary – add forms of sources here (not too much so it buffs out) – but to add weight and academic foundation to my updated Proposal points, giving it more credibility.

The application of watermarking and AI-detection technologies to publicly distributed media presents significant ethical implications – particularly in regions such as privacy, consent, attribution, and the risk of misuse. Without the indispensable integration of clear mitigative safeguards, and ethically grounded design features – such tools could inadvertently contribute, disastrously, to the continued “erosion of trust”, the mishandling of user data, or the misclassification of content. To proactively address these concerns, the following principles have been adopted to direct the development, and deployment, of this project in an ethically conscientious manner.

1. **Minimisation of Data Risk.**

All components of both the watermarking, and anomaly detection, systems will be designed to minimise the extraction, storage, and exposure of sensitive metadata. Albeit metadata is of the utmost significance for verifying authenticity, no personal data will be retained, or processed beyond what is necessary for the generation of hashed artefacts in the system.

1. **Transparency and User Awareness.**

This mechanism is designed with transparency entirely in mind. Stakeholders interacting with the system – regardless of whether this is via the web interface, or the API – must be clearly informed of what the watermarking procedure entails, what elements of metadata will be extracted, and how this information will be used. A future production-level deployment iteration of the mechanism would require definitive stakeholder consent, along with a visually discernible option to opt-out at all times. Similarly, all stakeholders involved in our blind study will undergo all required ethical procedures as dictated by Birmingham City University’s regulations.

1. **Open Standards, and Cryptographic Integrity.**

As aforementioned, proprietary cryptographic methods will not be employed. Instead, standardised, and openly available algorithms and libraries, such as SHA-256, “cryptography”, and “OpenCV” will be used – ensuring that the mechanism remains auditable, and verifiable by external reviewers. This open approach welcomes security through transparency rather than obscurity.

1. **Misuse Prevention.**

To mitigate the risk of this mechanism being exploited – to falsely label AI-synthetic content as authentic, and vice versa, to name one example – detective features will be limited to informative reporting, rather than automated enforcement. This pragmatic approach ensures no legal, or conclusive, status is assigned to any of its findings. As such, its results should be interpreted with human oversight.

1. **Regulatory, and Legal, Compliance.**

The proposed system will comply with relevant national, and international, legislation, including GDPR (General Data Protection Register) in the European Union, the Data Protection Act 2018, the Computer Misuse Act 1990, the ISO/IEC 27001:2022 standard on Information Security, and the Product Security and Telecommunications Act 2022. No data handled, or processed, by the proposed mechanism will violate existing data protection, or cyber security, legislation and regulatory frameworks.

1. **Free, and Fair, Access**.

The tool will be released under a permissive licence, most likely the MIT License, instead of the GNU GPLv3 License, due to the former’s “permissive” nature, and their regards towards “derivative works”. The issue with GNU GPLv3 is that it expects a derivative work to also inherit the open-source nature of this project, hypothetically restricting its capacity to develop in the future (Open Source Licensing Simplified: A Comparative Overview of Popular Licenses | Blog | Endor Labs, n.d.).

This will ensure accessibility to institutions, journalists, educators, and individuals; critically, without being behind a paywall, subscription model, or platform exclusivity. Although, as aforementioned, it may be forked into a commercialised distribution, the rudimentary version – developed in this paper – will always remain open-source. This reflects the paper’s broader objective of contributing to open, and verifiable, information ecosystems.

1. **Accountability in Design, Implementation, and Evaluation.**

All major design choices, limitations, and known risks – and their associated vulnerabilities – will be documented in full. This, in turn, will ensure that all future iterations of this project – should they advance towards professional deployment – are open to constructive scrutiny, and further collaborative development.

### 4.4.3 Practical Considerations

In a similar vein to the Ethical Considerations above, this project is similarly subject to several practical constraints, most notably of which – are temporal, and computational, resources. The following restrictions, and their associated decisions, have been acknowledged to scope the project realistically.

1. **Temporal Scale Constraints.**

The entire research, design, implementation, and further evaluation cycle for this project is constrained into a five-week time period, concluding in early September 2025. Given this limited scale, as will be discussed further in the “Methodology” section, this project will adopt an Agile methodology – enabling iterative development through incremental prototypes. Instead of pursuing a fully realised production-grade system, each developmental incrementation will fixate on delivering core aspects of the proposed system’s functionality – including watermark embedding, anomaly detection, and the implementation and integration of an API – in isolated, demonstratable stages.

This approach permits the incremental assessment of key technical decisions, ensuring that foundational design choices – such as the efficacy of watermark embedding, or the accuracy of anomaly detection heuristics – can be evaluated rigorously, and refined promptly in the implementation process.

1. **Computational Efficiency.**

Regarding the paper’s goal of wide accessibility, the system will be implemented utilising tools, and libraries, available in a standard Python installation on a mainstream personal computer. A fundamental level of care will be taken to minimise computational overhead, simultaneously ensuring that all core functionalities remain lightweight, performant, and most importantly – robust – on both consumer-level hardware, and enterprise-level (i.e. server) hardware.

1. **Scope of Evaluation.**

Whilst the system aims to, imperceptibly, embed watermarks, and include primitive anomaly detection heuristic functionality, it will not incorporate sophisticated Machine Learning (ML) classifiers for deepfake detection, due to temporal, and training, constraints. Instead, metrics, such as PSNR and SSIM – or a similar one after the initial Methodology evaluation process has concluded – alongside rule-based metadata heuristics, will serve as the principal evaluative tools.

1. **Prototype-Level Limitations.**

Due to development constraints, some features – such as user authentication (i.e. Tokens), upload encryption, or tamper-resistant logging – will be proposed but not fully implemented. These limitations will be clearly evoked, and hypothetical suggestions for future development of this project will be defined in the “Discussion and Results” section below.

1. **Participant Testing Feasibility.**

In a similar vein, given the arduous temporal constraints of this project – participant testing – such as the planned blind watermark perceptibility study - will be conducted on a small scale, as aforementioned previously. Whilst this does inadvertently limit the breadth of statistical generalisability, the analysis will still engage quantifiable metrics to evaluate our system’s performance in practical, everyday scenarios. The results, therefore, are best understood as applied insights into the system’s effectiveness, offering a meaningful indication of its potential impact, and usability, when deployed in relevant domains.

### 4.4.4 Ethical & Practical Consideration Conclusion

All things considered, these ethical and practical considerations are responsible for three main outcomes. They shape this research to ensure that its development remains responsible, technically feasible, and – most fundamentally – always aligned with the original aim. By acknowledging limitations, and embedding accountability at each stage, this paper seeks to constructively contribute to the broader conversation around digital authenticity, and the novel role of cryptography in safeguarding truth in visual media.

# 5 Literature Review

(Make sure to include visual models from sources) – not just one of them!

## 5.1 Introduction

Could probably tie this a little more to the Aims and Objectives mentioned earlier.

This chapter provides a critical review of the current academic, and technical, landscape surrounding digital media authenticity – with a central focus on imperceptible watermarking, metadata analysis, cryptographic hashing, and anomaly-based detection methods. Its purpose is to evaluate how these areas support the development of a system capable of verifying the human authorship of visual media.

To ensure relevance to the rapidly evolving fields of AI-synthetic content, and Cyber Security in its entirety – this Review intends to draw upon literature published exclusively from 2024 onwards. Sources include peer-reviewed articles, whitepapers, technical documentation, documented standards emerging from both academia, and industry; and finally, legal and regulatory frameworks (e.g. ISO/IEC 27001:2022, GDPR).

Why is the Lit. Review structured “thematically”, why is this an optimal procedure? References will be gathered once “blueprint” *Lit. Review Methodology* has been drafted.

This Review will be structured thematically, in order to address integral system components, relevant technologies, developmental or theoretical gaps, and considerations regarding the mechanism’s impending implementation. This analysis will, in turn, directly inform the design, justification, and evaluation of the proposed system’s different components – as outlined in the project’s aim and objectives earlier in the paper.

## 5.2 Literature Review Methodology

### 5.2.1 Justification & Rationale

Intrinsically, this Review’s rationale originates from the highly dynamic, and rapidly-evolving nature of technologies in relation to AI-synthetic media, and forms of digital content authentication. As the sophistication of generative models (e.g. “Sora”, “Midjourney”, “Runway”) increases exponentially, established techniques for validating media authenticity are becoming increasingly redundant {Perhaps note what kind of techniques?}. Therefore, the Review is intentionally structured to explore more recent advancements in metadata-based watermarking, cryptographic verification, anomaly detection in both image, and video, files; and in the broader sphere of AI-content discrimination.

Moreover, the Review – at a wider perspective – serves a dual purpose. Firstly, to establish a foundation of proven concepts, that could be adapted into our proposed mechanism; and secondly, to identify conceptual, or technical gaps, that research might directly address or work around. Furthermore, this ensures that our project remains grounded, notably, in current technological understanding – and robust to tangible implementation concerns – such as computational overhead, legal and regulatory compliance, and system transparency; three key cornerstones addressed earlier.

The methodology, presented below, has been chosen to ensure the Review only infers sources with direct, recent relevance – maintaining a close linkage between literature observations, and both the aim and objectives presented earlier in the paper.

### 5.2.2 Inclusion Criteria

This, and the Exclusion criteria, feel like they cover a lot of the same bases. Should they get integrated into a single section? Would probably save a lot on the word count.

Literature has only been included in the Review if it meets the following conditions – ensuring relevance to all aspects of the proposed mechanism. Furthermore, this necessitates selecting sources that are both technically relevant, and contextually aligned, with the challenges posed by synthetic media.

1. **Sources Published in 2024 or later.**

This criterion ensures that chosen literature reflects the most up-to-date developments in AI, watermarking techniques, cryptographic verification, and anomaly detection. Given the accelerated pace of innovation in these specific domains, especially, incorporating only the most current research inherently supports the technical feasibility, and relevance, of the proposed system.

However, this temporal restriction may result in foundational work, or highly influential studies, published earlier years – that helped shape the field – be inadvertently excluded. Whilst efforts will be made to include this type of imperative work where it’s referenced in more modern sources – there does remain a genuine risk of overlooking valuable historical context, or foundational methodologies.

1. **Peer-Reviewed Academic Articles, Technical White Papers, and Regulatory Frameworks (e.g. ISO/IEC 27001:2022).**

The addition of only peer-reviewed, or industry-validated material, will positively ensure high standards of technical accuracy, comprehensive methodology, and academic credibility. Consequently, these sources are more likely to have undergone professional validation, and are thus more reliable for informing mechanism design, and evaluation.

Nonetheless, this criterion might limit the incorporation of developing concepts, or practical insights, from industry practitioners that have not yet entered the peer-reviewed domain as of yet. Some non-traditional sources, such as developer blogs, or GitHub[[9]](#footnote-9) documentation, may offer practical relevance that is unfortunately excluded.

1. **Literature Focused on Image, and Video-based Media Authentication, Anomaly Detection, and Cryptographic Watermarking.**

This criterion, specifically, narrows the research scope to visual media – including images, and videos – as these constitute the primary “carrier datasets” of the system; ensuring that the literature directly supports the technical implementation, and evaluation, of the watermarking, and verification, mechanism proposed in this paper.

Conversely, a drawback of this narrowing is the exclusion of transferable methodologies from adjacent domains – including the likes of auditorial, and biometric, forensics. Though cross-domain insights could potentially enrich this project, these form of domains will be deprioritised in favour of technical specificity.

1. **Observed Studies Employing Metrics Such as PSNR, SSIM, and Other Quantitative Scoring Methodologies.**

Emphasising empirical validation ensures that techniques included in the Review are not only theoretically grounded, but also practically assessed. Such metrics are essential, not just for evaluating the imperceptibility of watermarks in a target file, but also for assessing the accuracy of synthetic content detection.

Despite this, a reliance on quantitative evaluation could lead to the underrepresentation of conceptual, or theoretical, contributions. Therefore, some cutting-edge approaches might still be in embryonic stages of development, and lack the sufficient level of empirical validation to meet this criterion.

1. **Explore Addressing Ethical, Legal, and Regulatory Aspects of Digital Provenance, and Synthetic Media Detection**

Inclusion of literature that explores legal compliance (e.g. GDPR, 2018 Data Protection Act), ethical concerns, and the extent of the proposed mechanism’s social impact will ensure that the mechanism is developed within a responsible framework. This supports the paper’s aim of producing a trustworthy, transparent, and accountable tool.

However, a predominant limitation with this is that such literature may overemphasise established policy, and neglect speculative or interdisciplinary techniques – which could, in turn, raise significant doubt regarding the broader consequences of detection technologies that are not fully understood at this present time.

### 5.2.3 Exclusion Criteria

Literature has only been excluded from the Review if it meets the following conditions – ensuring that retained sources directly contribute to the technical, methodological, and contextual requirements of the proposed mechanism. These exclusions aren’t intended to just be a simple inversion of the Inclusion Criteria – rather, they exist to protect the Review’s evidential strength, and alignment, with the mechanism’s intended purpose. Each criterion is applied deliberately, with careful consideration of their respective limitations.

1. **Outdated, or Deprecated, Technical Methods**

Sources that focus on technologies, algorithms, or approaches that are now considered obsolete – including watermarking tools vulnerable to modern removal tools, or pre-2015 hashing methods – have been excluded. The rationale to which this is based on is that only technically viable methods, capable of withstanding contemporary modern AI techniques, are considered in this paper.

A drawback is that certain older approaches might contain foundational concepts, or insights, that could inform modernised forms of hybrid, or updated systems. Unfortunately, potentially valuable theoretical underpinnings might be excluded.

1. **Irrelevant Application Contexts**

Literature that inspects technologies in unrelated domains, including – but not limited to – watermarking for biological datasets, or anomaly detection for Internet of Things (IoT) networks, are excluded. This will ensure that the Review remains focused on image, and video authenticity challenges imperative to the proposed mechanism. A notable drawback, that should be mentioned, is that cross-domain innovations, which could theoretically be adapted for synthetic media detection, could be mistakeably omitted due to these strict contextual regulations.

1. **Proprietary, or Non-Transparent Implementation**

Studies that rely on closed-source, proprietary mechanisms with a lack of publicly available technical details have been excluded. This certifies that all considered methods can be replicated, validated, and integrated into an open, and universally accessible, framework – in line with an integral cornerstone of this paper’s philosophy – complete commitment to transparency. Cutting-edge proprietary methods, which are often at the forefront of commercial AI detection – might be overlooked despite their potential relevance, simply due to their lack of technical disclosure.

1. **Lack of Empirical Validation**

Certain papers, that make use of theoretical claims without reinforcing them via empirical testing, benchmark datasets, or measurable evaluation metrics have been avoided – ensuring that retained literature is, therefore, grounded in attestable evidence; thus, enabling reliable comparisons, and informed system design. On the other hand, highly innovative – but early-stage – research, that has not yet undergone a rigorous regime of testing – might be disregarded despite its potential to inspire innovative new design choices.

1. **Limited Accessibility or Excessive Resource Demands**

Certain forms of literature will be excluded if their described methods rely on proprietary, hardware-specific, or excessively resource-intensive tools. These include, but aren’t limited to – high-end GPU clustering, large memory environments, or closed frameworks. These can’t be adapted to a lightweight, broadly accessible Pythonic-based implementation. This restriction will support the administration of an integral goal of this paper being that the proposed mechanism will have as low a computational overhead as possible – and likewise, avoiding unnecessary external dependencies where possible.

However, it must be pointed out that because of this filter – a collection of high-performance approaches might be neglected if a similar adaptation would require significant redevelopment, or compromise to an extent. Despite this, methods that can be feasibly re-engineered into a conventional Python environment, without significant overhead will still be studied.

### 5.2.4 Search Strategy

To promote a rigorous, comprehensive, and most significantly – robust – Review relevant to our proposed mechanism – this paper intends to employ a structured search strategy, with a distinct purpose to identify, and critically, evaluate, the sources that directly inform the technical, contextual, and methodological components of the research. Searches were conducted between June, and August, 2025 – with meticulous adherence to both the “Inclusion”, and “Exclusion” criteria’s in the last section.

1. **Databases, and Search Engines**

Primary searches were conducted with several reputable academic, and technical databases – including, but not limited to:

* IEEE Xplore
* ACM Digital Library
* ScienceDirect
* SpringerLink
* Google Scholar
* arXiv

1. **Keywords, and Search Strings**

Search queries were individually refined cyclically to balance specificity with coverage, therefore compounding core technical concepts with contextual terms. Forms of Boolean logic, and database-specific syntax, were both used in conjunction with one another to evoke optimal results. Several examples of these include:

- “cryptographic watermarking” AND ‘image authentication” AND “metadata integrity”

- “AI media detection” AND (“PSNR” OR “SSIM” OR “visual anomaly detection”)

- “content provenance”  
- “digital watermark”

- “metadata forgery” AND “detection methods” and (“image” OR “video”)

- “lightweight implementation” AND “Python” AND “cryptographic verification”

- “metadata analysis” AND (“provenance verification” OR “digital content authenticity”)

- “deepfake detection” AND (“forensic watermarking” OR “media integrity”)

- “hash embedding” AND (“imperceptible watermark” OR “digital media verification”)

- “geospatial metadata”  
- “image authenticity” OR “video authenticity”

- “rule-based AI detection” OR “heuristic-based AI detection”

- “digital provenance” AND (“GDPR compliance “OR “Data Protection Act” OR “Computer Misuse Act”)

- “video codec analysis”

- “image file analysis”

- “metadata forgery detection” AND (“hash verification” OR “tamper-proof watermarking”)

1. **Source Selection and Verification**

All search results underwent an initial screening process against the Inclusion Criteria to determine their suitability. This vital process ensured that sources were published later than 2024 – to maintain relevance in this rapidly-developing domain; presented a distinct contribution to at least one aspect of the proposed mechanism’s aim; and originated from either a peer-reviewed, industry-accredited, or high-standard source – to endorse technical accuracy, and academic credibility.

Likewise, relevant work was also excluded if any segments of the Exclusion Criteria were also met. For example, if obsolete methodologies were employed, proprietary dependencies were introduced, or a lack of technical transparency was apparent.

Furthermore, in certain cases where ambiguity arose regarding a source’s methodological, or technical relevance – a systematic two-pass evaluation process was undertaken. The initial pass constituted an abstract-level review to determine preliminary suitability against both the Inclusion, and Exclusion, criteria’s. Only if a source progressed, would they then endure a second, full-text review to ascertain their technical validity, methodological stringency, and contextual alignment with the project’s aims before their final inclusion.

This approach ensures that research time is used efficiently, by filtering out unsuitable material at an early stage. Despite this, the strategy does permit a thorough examination of sources, that may appear initially ambiguous – but hold potential value to this paper’s proposed mechanism’s development.

### 2.5 Benefits & Limitations of the Methodology

Should I modify this so that a comparison should be made with other forms of Review? (i.e. Scope Review, Systematic Review)

This methodology ensures that all literature included in the Review directly supports the project’s distinct aim of creating an accessible, verifiable authenticity mechanism for all types of digital media. By restricting sources to 2024, and onwards, inherently secures this paper alignment with the latest developments in imperceptible watermarking, anomaly detection, and cryptographic verification – reducing the risk of relying on obsolete, precarious methods. A combination of broad conceptual terms, with precise technical keywords in the methodical search strategy enabled the capture of both foundational concepts, and emerging techniques – with a two-phased screening process – ensuring efficiency, and relevance, throughout. Correspondingly, the prioritisation of sources evoking open, non-proprietary methods, and those with minimal resource requirements – ensures that any proposed approach can be feasibly implemented within the project’s five-week timeframe.

Despite this, compromises have been made. Excluding pre-2024 literature has inadvertently omitted historically influential work – but this has been since mitigated, to a certain extent – by fastidiously referencing foundational research where theoretically necessary. Comparably, the avoidance of proprietary, or high-dependency methods increase accessibility – but risks excluding theoretically high-performance solutions – in such cases, conceptual understandings are incorporated without any reliance on closed implementation. However, it must be noted that these constraints, though restrictive – reinforce the feasibility, reproducibility, and accessibility of the final product – whilst also ensuring that it remains practical for wide adoption by all forms of demographic. Disregarded convoluted methods will, nevertheless, be acknowledged in future discussion below for future exploration – under fewer deployment constraints.

## 5.3 Literature Research

### 5.3.1 The Rise of AI-Synthetic Content, and the “Erosion of Trust”

# 6 Methodology

(Make sure Design (e.g. Agile), Implementation choices (e.g. PSNR/SSIM, SHA-256, is questioned comprehensively and compared with similar technologies, advantages and disadvantages -> do this in 6.4/6.5 Sections.

## 6.1 Introduction

## 6.2 Design Methodology

Make sure to consider all major vulnerabilities associated with different components of the system.

## 6.3 Practical Implementation Methodology

## 6.4 Benefits and Limitations of Chosen Methodologies

## 6.5 Methodology Evaluation

# 7 Development and Testing

# 8 Discussion and Results

# 9 References

Sommerer, T. (2025) Baudrillard and the Dead Internet Theory. Revisiting Baudrillard’s (dis)trust in Artificial Intelligence. *Philosophy & Technology*, 38 (2): 54. doi:[10.1007/s13347-025-00878-5](https://doi.org/10.1007/s13347-025-00878-5).

Athaluri, S.A., Manthena, S.V., Kesapragada, V.S.R.K.M., et al. (2023) Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus*. doi:[10.7759/cureus.37432](https://doi.org/10.7759/cureus.37432).

Yaprak, B. (2024) Generative Artificial Intelligence in Marketing: The Invisible Danger of AI Hallucinations. *Journal of Economy Business and Management*, 8 (2): 133–158. doi:[10.7596/jebm.1588897](https://doi.org/10.7596/jebm.1588897).

Hallucination (artificial intelligence) (2025). *Wikipedia*. Available at: <https://en.wikipedia.org/w/index.php?title=Hallucination_(artificial_intelligence)&oldid=1303077795> (Accessed: 29 July 2025).

Roamer, S.S. is the founder of C., Adventure, A.T.W.F. on, adventure, the outdoors W. a love for, et al. (2024) *How To Find Glenfinnan Viaduct Viewpoint In 2025?* Available at: <https://continuousroamer.com/glenfinnan-viaduct-viewpoint/> (Accessed: 29 July 2025).

*The Jacobite* (steam train) (2025). *Wikipedia*. Available at: <https://en.wikipedia.org/w/index.php?title=The_Jacobite_(steam_train)&oldid=1290249391> (Accessed: 29 July 2025).

*Sora* (n.d.). Available at: <https://openai.com/sora/> (Accessed: 29 July 2025).

Lee, B., Martinez, A., Thompson, P., et al. (2023) *AI-Generated Imagery Landscape: A Journey from Apprehension to Insightful Exploration*. doi:[10.13140/RG.2.2.34039.39843](https://doi.org/10.13140/RG.2.2.34039.39843).

Burri, T. (n.d.) The New EU Legislation on Artificial Intelligence: A Primer.

Miller, E.J., Steward, B.A., Witkower, Z., et al. (2023) AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychological Science*, 34 (12): 1390–1403. doi:[10.1177/09567976231207095](https://doi.org/10.1177/09567976231207095).

Pasternak, M., Warren, K., Olszewski, D., et al. (2025) “Characterizing the Impact of Audio Deepfakes in the Presence of Cochlear Implant.” In *Proceedings 2025 Network and Distributed System Security Symposium*. San Diego, CA, USA, 2025. Internet Society. doi:[10.14722/ndss.2025.241117](https://doi.org/10.14722/ndss.2025.241117).

Khanjani, Z., Watson, G. and Janeja, V.P. (2023) Audio deepfakes: A survey. *Frontiers in Big Data*, 5: 1001063. doi:[10.3389/fdata.2022.1001063](https://doi.org/10.3389/fdata.2022.1001063).

*How to spot a deepfake - SoSafe* (2024). Available at: <https://sosafe-awareness.com/blog/how-to-spot-a-deepfake/> (Accessed: 29 July 2025).

*Open Source Licensing Simplified: A Comparative Overview of Popular Licenses | Blog | Endor Labs* (n.d.). Available at: <https://www.endorlabs.com/learn/open-source-licensing-simplified-a-comparative-overview-of-popular-licenses> (Accessed: 4 August 2025).

*About GitHub and Git* (n.d.). Available at: <https://docs-internal.github.com/en/get-started/start-your-journey/about-github-and-git> (Accessed: 8 August 2025).

# 10 Bibliography

# 11 Appendices

## 11.1 Glossary

1. For the purposes of this paper, “Authentic” denotes media content that has been created entirely by human means – without any form of intervention by Artificial Intelligence (AI). Conversely, “Synthetic” media connotes content that has been produced – either in its entirety, or partly – using AI-based generative models, or forms of algorithms. [↑](#footnote-ref-1)
2. The “Dead Internet Theory” is a fringe hypothesis asserting that much of today’s web content is no longer produced by humans, but by bots and generative AI models {Insert reference to support}. [↑](#footnote-ref-2)
3. The term “hallucination” here refers to Sora’s capacity to generate visually plausible, but entirely fabricated image, and video, outputs. [↑](#footnote-ref-3)
4. Whilst the watermark is designed to be imperceptible – the empirical reality is that certain modes of compression, editing practices, and file manipulation procedures may unintentionally degrade, or expose, embedded artefacts. [↑](#footnote-ref-4)
5. The term “carrier dataset” is used here to refer to image, and video, files – that serve as the medium for embedded watermark and metadata artefacts, as per specialised guidance from my supervisor – Shahid Shabbir. [↑](#footnote-ref-5)
6. Metadata fields, such as EXIF data, are often stripped, or altered (to an extent), by platforms and editing tools. As such, their absence is not an obvious sign of AI-synthetic media. [↑](#footnote-ref-6)
7. During implementation, REST API endpoints will not be implemented with any form of TLS/SSL encryption for reasons of simplicity, performance testing, and temporal constraints. Functionality, of this kind, may hypothetically be implemented in a future iteration of this mechanism. [↑](#footnote-ref-7)
8. The 95% rate prescribed represents a target benchmark for evaluating real-world classification efficacy, based on internal performance goals. It must be explicitly stated that it does not reflect an established industry threshold of any form. [↑](#footnote-ref-8)
9. “GitHub” is a “cloud-based platform where” uses can “store, share, and work together with others to write code” (About GitHub and Git, n.d.). It stores “repositories”, with forum pages that users can contribute their work to via the “Git” protocol. [↑](#footnote-ref-9)