Combating Deepfake Fraud in Insurance

A Strategic and Ethical AI-Driven Response for Monash Insurance

ZhouChi Wang | 29/04/2025

Introduction

What is a Deepfake?

A Deepfake is a synthetic image, video, or voice that's generated using artificial intelligence, often with an intention to mimic a real-life person or scene. As technology improves, Deepfakes has become more realistic and harder to detect (The Alan Turing Institute, 2023).

Recently, we were approached by a huge insurance company by the name of Monash Insurance. They were deeply concerned about the issue regarding customers that were being scammed by malicious parties using Deepfake AI tools to trick their customers into paying insurance towards the scammers rather than Monash Insurance themselves, on occasions even tricking Monash Insurance to pay for scammers instead. This incident highlights a broader societal issue: deepfakes are undermining personal and institutional security by exploiting the trust placed in visual media (Hirsch, 2023).

There are many such examples of incidents where Deepfake AI was used to negatively impact insurance companies. In 2024, UK insurers Alianz and Zurich reported a significant rise of claims on insurance using Deepfake technology to falsely edit images, making it appear as though that the vehicle was severely damaged, in hopes to scam an insurance claim from the companies. For instance, there was one particular case where an individual had posted a photo of his van on social media. Fraudsters saw this opportunity and digitally altered this image to include fake damages on the front bumpers and tried to claim a repair invoice exceeding £1,000. Upon closer inspection by the insurance company, it has been discovered that the submitted photos from fraudsters were identical to those of the one from social media, just with the added damage. This rise in image-based fraud, often referred to as "shallowfakes," led to a 300% increase in such incidents between 2021–2022 and 2022–2023 (The Guardian, 2024).

This initiative is especially crucial given Monash Insurance's current situation, which reflects systemic problems where malicious actors exploit gaps in both technological detection and regulatory oversight. A solution to this problem is of utmost importance, as scammers could cost Monash Insurance up to millions of dollars in revenue each year and potentially scare away regular customers.

Under such circumstances, we propose developing an AI application that is able to detect malicious parties and take immediate action against them. Our proposed solution integrates advanced deepfake detection algorithms with automated response protocols. Drawing from recent advancements in digital forensics, the system uses convolutional neural networks (CNNs) trained on datasets like FaceForensics++, which catalog manipulated facial features and artifacts (Rossler et al., 2019). Techniques such as analyzing eye-blinking patterns, which is often inconsistent in synthetic videos and identifying warping anomalies around facial contours improve detection accuracy (Li, Chang & Lyu, 2018). Additionally, the model employs Xception architecture, optimized to detect subtle pixel-level irregularities in images (Chollet, 2017). Beyond these

technical measures, the solution focuses on real-time analysis to flag suspicious claims immediately, reducing the delays associated with manual reviews. This report discusses the threat of Deepfake AI towards insurance companies, the ethical issues surrounding it and proposes a tailored solution for Monash Insurance company to tackle this problem.

Background and problem statement

As much as $308.6 billion is lost each year to insurance fraud — about a quarter of the size of the industry (Coalition Against Insurance Fraud, 2022). Why do these thieves have to make fake registration numbers (ZURICH, 2024) of "total loss" vehicles and get the insurance on it? These days, generative AI has also paved the way for fake paperwork: In seconds, bad actors can now forge entirely realistic-looking invoices or underwriting appraisals with real-seeming signatures and letterhead (Nicos Vekiarides, 2024).

Insurance companies have as their number one priority to protect its consumers from these Deepfake incidents to maintain its mutual trust and relationship, as failing to do so results in a huge financial loss for the companies, whereas mentioned above, it is estimated to be upwards of $308.6 billion (Coalition Against Insurance Fraud, 2022).

How is Deepfake generated?

Deepfake is mainly created using Generative Adversarial Networks (GANs) and Diffusion Models. GANs uses two AI models, where one is used as a generator that creates fake content and the other used as a discriminator that tries to spot fake content. As more interactions occur, it improves both AI models, resulting in a more realistic image or video. Diffusion models start with a noisy version of an image and learn to reconstruct the original. These are newer than GANs and are easier to train (The Alan Turing Institute, 2023).

How to detect Deepfakes?

Detecting Deepfakes has grown significantly difficult as time passes since it is always improving through AI models. There are several techniques that can be applied to detect Deepfakes. Some deepfaked images and videos contain inconsistencies like mismatched lighting, unnatural facial features and off sync lip movement which can be spotted easily. Besides that, GANs and diffusion models also leave subtle digital patterns that detection models can easily recognize. Moreover, Deepfakes used for malicious purposes often spread via fake or bot social media accounts, which can be flagged based on their metadata and activity, entirely avoiding the need to directly detect the Deepfakes themselves (The Alan Turing Institute, 2023).

Monash Insurance is facing a growing threat from Deepfake technology, where AI-generated images, videos, and documents are being used by fraudsters to submit false claims and mislead both customers and the company. This type of fraud contributes to the broader issue of insurance fraud, which costs the industry over $308.6 billion annually. Deepfakes are particularly dangerous because they are becoming increasingly realistic and harder to detect, leading to significant financial losses and damaging the trust between insurers and their customers. To address this issue, Monash Insurance company requires an AI- driven solution that can detect Deepfakes in real time that can immediately take action against any suspicious activities.

Methodology

This project uses ERF as a procedural ethical approach to computationally analyse the wicked problem of insurance claim fraud through the use of Deepfake technology and aids in creating an ethical AI system suitable for Monash Insurance. Going beyond merely cost-benefit or technical feasibility-driven solutions, the ERF allows for a deeper examination of moral responsibilities to the public, breach of trust present in Deepfake fraud, and the professional ethical duties of computing professionals (Gotterbarn et al., 2018). This approach ensures that the suggested AI application is not only solving the technical challenge, but also is robustly aligned with the computing ethics principles (ACM, 2018), thus creating societal value and guaranteeing fairness among the stakeholders. Our method focuses on Steps 1 and 2 of the ERF-Defining the Ethical Problem and Gleaning the Relevant Ethical Facts-to guide our analysis and to provide rationale for the recommended AI solution.

The Ethical Reasoning Framework defines a process for recognizing and evaluating ethical issues, determining facts, critically analyzing options, and justifying actions for addressing complex computer-based challenges in a formal and structured way (Gotterbarn et al., 2018). In this paper, the ERF is the framework under which we examine the moral wrongness created by insurance fraud using AI-generated synthetic media (Deepfakes). The backdrop describes what is happening — the trend of more convincing digital liars scamming insurers — but the ERF permits us to go further: to consider why it is ethically wrong to act this way and just what existing moral standards are being breached. Through the conscientious application of Step 1 (Defining the Ethical Problem) and Step 2 (Gleaning the Appropriate Ethical Facts) we establish a sound ethical basis to underpin the development and deployment of our proposed AI detection and response system for Monash Insurance.

Ethical reasoning framework

Step 1: Identifying the Ethical Problem and ACM Code Violations

The central ethical concern arises from its international use to mislead and defraud, in the first place by means of purposeful application of AI, namely the generation of deepfakes (Heidari et al. 2024; Rana et al. This involves Monash Insurance, its clients and others where fake proof (photo, videos, documents) are created for fake insurance claims. This application of AI is inherently deceptive, and overtly unethical (Li & Wan, 2023; Tuysuz & Kılıç, 2023). However, when we add levels of modern AI, the ethical challenge becomes increasingly complex. It is more than just "a betary of trust" between an insurer and a claimant; it is a serious blow to the trustworthiness of the digital evidence itself (Li & Wan, 2023). With deepfake tools growing more available and real-looking, a mismatch in knowledge and powers in favor of the malicious party can be engineered: the latter can counterfeit records that are too hard for insurers and even the human eye to tell apart from the real thing (Rössler et al., 2019).

This state of affairs is inconsistent with a number of foundational principles specified in the ACM Code of Ethics and Professional Conduct (ACM, 2018).

Principle 1.2: Avoid Harm. The use of deepfakes for insurance fraud causes significant, direct financial harm to Monash Insurance through fraudulent payouts. It also inflicts indirect harm on honest policyholders, who may face increased premiums due to the insurer's losses or become subject to heightened suspicion and more intrusive verification processes, potentially delaying legitimate claims. Furthermore, the proliferation of undetectable deepfakes harms society by eroding trust in digital media and institutions (ACM, 2018; Li & Wan, 2023). The AI system itself, if poorly designed (e.g., high false positive rate), could also cause harm by wrongly rejecting legitimate claims or flagging innocent customers, thus requiring careful consideration to minimize unintended harm (ACM, 2018).

Principle 1.3: Be Honest and Trustworthy. Deepfake fraud is a fundamental violation of honesty and trustworthiness (ACM, 2018). Malicious actors deliberately use AI to create deceptive media, misrepresenting reality to gain undue financial advantage. This undermines the foundational trust necessary for the insurance system to function effectively. The increasing sophistication of Deepfakes makes it difficult for Monash Insurance to trust the evidence submitted, damaging the relationship with all claimants. An ethical response requires not only detecting dishonesty but also implementing systems that are themselves transparent and trustworthy, providing clear information about their limitations and potential problems (ACM, 2018).

Principle 1.4: Be Fair and Take Action Not to Discriminate. While the primary act of fraud is unfair, the response to Deepfake threats must also be fair (ACM, 2018). Increased scrutiny applied broadly could disproportionately affect certain customer groups or lead to biases in how claims are processed if detection tools are not carefully designed and audited for fairness. The solution must ensure that the burden of proof does not unfairly shift to honest customers and that detection mechanisms do not introduce new forms of discrimination. This principle also calls for creating inclusive environments, which, in this context, means ensuring that security measures do not alienate or unduly burden legitimate users. Harassment and cyberbullying, sometimes facilitated by Deepfakes, are also explicit violations.

Principle 1.6: Respect Privacy. The creation of Deepfakes often involves the unauthorized use of individuals' biometric data (facial images, voices) scraped from public sources or obtained illicitly. While the proposed AI solution detects Deepfakes rather than creates them, its operation involves processing potentially sensitive claimant data (images, videos, documents). Therefore, the solution must adhere strictly to privacy principles, ensuring data is collected and used only for legitimate fraud detection purposes, with appropriate informed consent mechanisms, security safeguards, and data retention policies (ACM, 2018; Li & Wan, 2023). Compliance with regulations like GDPR is paramount (Gotterbarn et al., 2018).

Principle 2.5: Give Comprehensive and Thorough Evaluations of Computer Systems and Their Impacts, Including Analysis of Possible Risks. Insurers relying on outdated or purely manual verification processes in the face of known Deepfake risks may fail to meet this principle (ACM, 2018). A thorough evaluation of the current claims verification system at Monash Insurance would likely reveal its inadequacy against AI-generated fakes. This principle mandates a proactive assessment of system capabilities and risks, necessitating the adoption of more advanced detection methods. The proposed AI solution itself must also undergo continuous, thorough evaluation regarding its accuracy, potential biases, security vulnerabilities, and societal impacts (ACM, 2018).

Principle 2.8: Access Computing and Communication Resources Only When Authorized. The creation of deepfakes for fraud often involves unauthorized access to and use of source images/videos and potentially compromised accounts or systems for dissemination (ACM, 2018). The proposed AI solution must operate within authorized boundaries, accessing only the necessary data for claim verification with proper permissions.

Principle 2.9: Design and Implement Systems That Are Robustly and Usably Secure. The deepfake problem highlights the vulnerability of systems reliant on visual verification. An ethical response demands a system that is not only capable of detection but is itself robustly secure against attacks (e.g., adversarial attacks designed to fool the detector) and is usable by Monash Insurance personnel (ACM, 2018). Security features must be practical and not overly burdensome to ensure they are consistently used.

Step 2: An Ethical AI Solution Addressing ACM Code Violations

In ERF we are forced to discuss the ethical implications surrounding the descriptions of the factual situation (Gotterbarn et al., 2018). The difficulty is not merely the technical task of recognizing Deepfakes, but the ethical failing when insurers are powerless to combat this fraud. Isn't leaving the door open to fraud (for lack of detection) also effectively punishing honest customers? They end up paying more or doing more work. This is an unfair allocation of harm: the fraudster is the winner, and the honest claimant the victim, placed on hold, subject to an intrusive investigation, and perhaps denied even a legitimate claim. A college's failure to properly authenticate digital submissions is a failure in its duty of care to its communities. This situation is already great pressure for AI developers and insurers to predict the future misuse and develop system that supports accountability, transparency, and fairness (ACM, 2018; Tuysuz & Kılıç, 2023). As other scholars and professionals have pointed out, the use of tools like deepfakes in high-stakes fields like insurance requires responsible AI solutions – and new Security measures to back them up (Li & Wan, 2023). The ERF further requires that we do more than design a technical fix; we must design a solution that actually limits additional moral harm (Gotterbarn et al., 2018).

Building on the lessons learned from ERF and the analysis of ACM code violations, we propose an AI solution to this problem as a means of deploying a prototype Deepfake Detection and Response System (DDRS) for Monash Insurance. A solution that has been designed having both practicality in mind as well as ethical concerns at its heart, a solution that answers the identified breaches and takes into account the impact on the variety of audiences and affected parties.

Addressing ACM Code Violations:

1. Mitigating Harm (Principle 1.2): The primary function of the DDRS is to prevent the financial harm caused by fraudulent Deepfake claims. By accurately identifying manipulated media in real-time (Rössler et al., 2019), it aims to stop fraudulent payouts before they occur. Crucially, the system design prioritizes minimizing harm to honest policyholders (ACM, 2018). This involves:

High Accuracy & Low False Positives: Training the detection models (e.g., CNNs like Xception (Chollet, 2017), potentially combined with RNNs for temporal analysis (Güera & Delp, 2018)) on diverse, large-scale datasets (like FaceForensics++ (Rössler et al., 2019), DFDC (Dolhansky et al., 2020), Celeb-DF (Li et al., 2019), and potentially proprietary Monash data) to achieve high accuracy and, critically, a very low false positive rate. Techniques like analyzing eye-blinking inconsistencies (Li et al., 2018), GAN-generated artifacts (Guarnera et al., 2020; Wang et al., 2020), spatial and temporal features (Rana et al., 2022), and frequency domain analysis (Frank et al., 2020) will be employed. Ensemble methods might be used to improve robustness (Rana & Sung, 2020).

Risk-Based Triage: Instead of outright rejection, flagged claims are routed for prioritized human review by specialized fraud investigators, ensuring a human-in-the-loop for borderline cases.

Continuous Monitoring: Regularly evaluating model performance against new deepfake techniques and real-world data to retrain and update the system (ACM, 2018), mitigating the risk of outdated detection capabilities causing harm.

2. Restoring Honesty and Trustworthiness (Principle 1.3): The DDRS directly combats the dishonesty inherent in Deepfake fraud. By providing a reliable mechanism for verifying digital evidence, it helps restore trust between Monash Insurance and its genuine customers. To embody trustworthiness itself, the system will incorporate:

Explainability: Implementing methods (e.g., attention mechanisms, LIME, SHAP) to provide investigators with insights into why a particular piece of media was flagged as suspicious, facilitating informed human review rather than blind reliance on the AI (Balasubramaniam et al., 2022).

Transparency in Limitations: Clearly documenting the system's capabilities, limitations, known vulnerabilities (e.g., susceptibility to certain adversarial attacks (Hussain et al., 2020)), and performance metrics (accuracy, precision, recall, AUC) for internal stakeholders (ACM, 2018).

3. Ensuring Fairness and Non-Discrimination (Principle 1.4): The DDRS is designed to be a tool for precise detection, not broad suspicion. Fairness is embedded by:

Targeted Application: The system analyzes submitted claim evidence, not policyholders directly, reducing the risk of profiling.

Bias Auditing: Regularly auditing the training data and model predictions for potential biases related to demographics (age, ethnicity, gender, etc.) reflected in the visual data, ensuring the system does not unfairly target specific groups (ACM, 2018). Utilizing diverse datasets like DFDC, which used paid actors from various backgrounds (Dolhansky et al., 2020), can help mitigate initial bias.

Standardized Process: Ensuring that flagged claims follow a standardized, fair review process, preventing arbitrary treatment of claimants.

4. Respecting Privacy (Principle 1.6): Data privacy is a cornerstone of the DDRS design (ACM, 2018).

Data Minimization: The system only accesses and processes data strictly necessary for claim verification and fraud detection.

Purpose Limitation: Data processed for fraud detection is not used for other purposes (e.g., marketing, underwriting) without explicit consent.

Secure Handling: Implementing robust security measures (encryption, access controls) to protect claimant data during processing and storage.

Compliance: Designing the system to comply with relevant data protection laws (e.g., GDPR), including provisions for data access, correction, and deletion where applicable (Gotterbarn et al., 2018). Informed consent procedures for data use in model training and operation will be carefully implemented.

5. Providing Thorough Assessments (Principle 2.5): The DDRS represents a necessary upgrade to provide a more thorough assessment of claim validity in the modern threat landscape. It automates the analysis of visual and potentially other data points at a scale and depth infeasible through manual review alone. Continuous risk assessment of the DDRS itself is also integral (ACM, 2018).
6. Authorised Access (Principle 2.8): The system operates within Monash Insurance's infrastructure, accessing claim data through authorized channels and APIs, adhering to internal data governance policies (ACM, 2018).
7. Robust and Usable Security (Principle 2.9): The DDRS incorporates security by design (ACM, 2018).

Model Security: Investigating defenses against adversarial attacks aimed at fooling the detection models (Carlini & Farid, 2020).

System Security: Hardening the application infrastructure against unauthorized access or tampering.

Usability: Designing the user interface for claims handlers and investigators to be intuitive, integrating seamlessly into existing workflows and clearly presenting detection results and confidence scores without requiring deep technical expertise.

Incorporating Users:

1. Direct Users (Monash Insurance):

Claims Handlers: Receive initial flags/scores from the DDRS integrated into their claims processing system, enabling them to quickly identify low-risk claims and escalate high-risk ones. The interface must be clear and provide confidence levels.

Fraud Investigators: Use the DDRS outputs, including explainability features (Balasubramaniam et al., 2022), as a primary tool for in-depth investigation of flagged claims. They provide feedback on model performance, contributing to retraining cycles.

IT/Security Teams: Responsible for system deployment, maintenance, security monitoring, and managing updates and model retraining processes.

Leadership/Management: Utilize aggregated data and reports from the DDRS to understand fraud trends, assess system ROI, ensure regulatory compliance, and fulfill their leadership responsibilities under Principle 3 (e.g., 3.1 Public Good, 3.2 Social Responsibilities, 3.4 Supporting Ethical Policies, 3.7 Infrastructure Stewardship) (ACM, 2018).

2. Indirect Users and Impacted Parties:

Honest Policyholders: Benefit from more efficient processing of legitimate claims (as resources are not wasted on fraud) and potentially more stable premiums. The system aims to minimize their burden during verification.

Fraudulent Actors/Scammers: Face a higher probability of detection and failure, deterring future attempts. The system's automated response protocols (e.g., immediate flagging, potential account holds pending investigation) aim to act swiftly against them.

Regulatory Bodies/Government: Monash Insurance can use the DDRS implementation to demonstrate proactive compliance with financial regulations, anti-fraud measures, and data protection laws. Aggregated, anonymized data might inform regulatory understanding of deepfake threats (Tuysuz & Kılıç, 2023).

Wider Society: By effectively combating a significant vector of digital deception, the DDRS contributes modestly to maintaining trust in digital interactions (Li & Wan, 2023) and potentially deterring the use of deepfakes for other malicious purposes.

The Originality Of Our System:

While the underlying deep learning techniques (CNNs (Heidari et al., 2024), specific architectures like Xception (Chollet, 2017), analysis of features like eye-blinking (Li et al., 2018) or GAN artifacts (Wang et al., 2020)) are established in the research literature (Rana et al., 2022), the originality of the proposed DDRS lies in:

Domain-Specific Application And Integration: Tailoring and integrating these techniques specifically for the nuanced requirements of insurance claim fraud detection at Monash Insurance. This involves training models on relevant data types (claim photos, potentially video evidence, scanned documents) and integrating the system into specific insurance workflows.

Ethical Framework Integration: The explicit and rigorous grounding of the system's design and operation within the ERF and ACM Code of Ethics (ACM, 2018; Gotterbarn et al., 2018), ensuring ethical considerations are primary, not secondary.

Automated Response Protocols: Moving beyond mere detection to include automated response protocols (as mentioned in the introduction). While high-risk cases require human review, the system can potentially automate initial actions like flagging, prioritizing queues, or even requesting specific additional verification from claimants based on detection confidence and type, streamlining the response to suspicious activities.

Real-Time Analysis Focus: Emphasizing real-time or near-real-time processing to enable immediate flagging during the claim intake or initial review process, reducing the window for fraudulent actors.

This tailored, ethically grounded, and action-oriented application of deepfake detection technology constitutes the novel contribution of this project for Monash Insurance. The system aims not just to detect fakes but to provide a trustworthy, fair, and effective defense against a significant and evolving threat (Mirsky & Lee, 2021), upholding both the business integrity of Monash Insurance and the ethical principles of the computing profession (ACM, 2018).

Proposed solution

In response to the significant challenges posed by deepfake technology as identified in the problem statement — the financial loss and lack of trust experienced by Monash Insurance as a result of fraudulent claims (Coalition Against Insurance Fraud, 2022; Hirsch, 2023) — we propose that an AI-driven Deepfake Detection and Response System (DDRS) be developed and adopted. This solution directly addresses the requirement for a strong, time-critical approach to detecting and responding to fakes contributed to the claims, and has the advantage that it is agnostic to the nature of the media.

The DDRS is grounded in the ethical analysis of the method section above and the design of it is anchored in the standards of the ACM Code of Ethics and Professional Conduct (ACM, 2018). It leverages state-of-the-art deep learning which may include CNNs such as potentially using architectures such as Xception (Chollet, 2017; Rössler et al., 2019) and integrates the analyses of different types of manipulation artefacts such as eye-blinking inconsistency (Li et al., 2018), GAN-generated patterns (Guarnera et al., 2020; Wang et al., 2020) and spatio- temporal features (Rana et al., 2022), and can be trained on large-scale, real-world, challenges like large-scale, diverse datasets (Dolhansky et al., 2020; Rössler et al., 2019).

Importantly, these reasons are directly linked to the narrowed down ethical breaches and stakeholder concerns. In pursuing highly precise detection but with low occurrence of false positives our system itself targets ACM Principle 1.2 (Avoid harm), where we want to avoid the financial loss through fraud but also the risk of punishing honest customers wrongly (ACM, 2018). By acting as a trusted verifier of digital evidence, it aims to counteract the breach of Principle 1.3 (Be Honest and Trustworthy), which in turn can help reestablish trust in the claims domain (ACM, 2018; Li & Wan, 2023). By integrating bias auditing, explainability features and a human-in-the-loop to review flagged cases, is guaranteed the compliance with Principle 1.4 (Be Fair and Take Action Not to Discriminate) (ACM, 2018). Moreover, Principle 1.6 of privacy implicit into the architecture of the system (ACM, 2018, Gotterbarn et al., 2018) relates to the profound ethical considerations obtained from methodology, as well as authorized data access policies (Principle 2.8), and strong security design in this case (Principle 2.9).

As such, the proposed solution does not only offer a technical remedy. It provides Monash Insurance with a process on how to implement the deep assessments that are mandated by Principle 2.5 (ACM, 2018) in an era of contemporary digital dangers. Real-time detection combined with automated response procedures (e.g., flagging claims for immediate human review) make the DDRS a proactive, performant, and ethical defense mechanism. It recognizes the challenging cast of user-types, including internal claims handlers and investigators, external policyholders, and regulators, it strives to improve security and trust, while ensuring individual rights and fairness are not compromised. Monash Insurance can now leverage this ethically conscious AI to combat Deepfake fraud responsibly.

Conclusion

Ultimately, Deepfake technology poses a significant threat to the insurance industry, particularly to companies like Monash Insurance. With more realistic images, video, and printed materials created by artificial methods made more readily available to malicious actors, threats of fraud, economic loss, and erosion of customers' trust have increased tremendously. Through using a rigorous ethical analysis aided by the Ethical Reasoning Framework (ERF) and referencing the ACM Code of Ethics, our analysis not only identifies the seriousness of these challenges but also emphasizes that a response that is both ethically sound and technically solid is needed.

The proposed Deepfake Detection and Response System (DDRS) addresses these concerns in an effective way by incorporating cutting-edge artificial intelligence methods coupled with a strong commitment to fairness, transparency, respect for privacy, and harm avoidance. Through real-time analysis, interpretability, auditing, and strong data management, the DDRS aims to promote consumer trust, protect Monash Insurance from fraud, and ensure ethical practice in using artificial intelligence technology.

Ultimately, this project demonstrates how, through strategic technology utilization and ethical compliance, potential vulnerabilities can be turned into means of innovation, cooperation, and lasting resiliency.

References

Association for Computing Machinery (ACM). (2018). ACM Code of Ethics and Professional Conduct. https://ethics.acm.org/code-of-ethics/

Balasubramaniam, N., Kauppinen, M., Hiekkanen, K., & Kujala, S. (2022). Transparency and explainability of AI Systems: ethical guidelines in practice. Springer. https://www.sciencedirect.com/science/article/pii/S0950584923000514

Carlini, N., & Farid, H. (2020). Evading deepfake-image detectors with white- and black-box attacks. arXiv. https://doi.org/10.48550/arxiv.2004.00622

Coalition Against Insurance Fraud (2022). The impact of insurance fraud on the U.S. economy 2022. https://insurancefraud.org/wp-content/uploads/The-Impact-of-Insurance-Fraud-on-the-U.S.-Economy-Report-2022-8.26.2022-1.pdf

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The DeepFake Detection Challenge (DFDC) Dataset. arXiv https://doi.org/10.48550/arXiv.2006.07397

Frank, J., Eisenhofer, T., Schonherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. arXiv. https://doi.org/10.48550/arXiv.2003.08685

Gotterbarn, D., Bruckman, A., Flick, C., Miller, K., & Wolf, M. J. (2018). ACM Code of Ethics: A guide for positive action. Communications of the ACM, 61(1), 121–128. https://doi.org/10.1145/3173016

Guarnera, L., Giudice, O., & Battiato, S. (2020). Fighting deepfake by exposing the convolutional traces on images. arXiv. https://doi.org/10.48550/arXiv.2008.04095

Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. IEEE. https://doi.org/10.1109/AVSS.2018.8639163

Heidari, A., Navimipour, N. J., Dag, H., & Unal, M. (2023). Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 14(2). https://doi.org/10.1002/widm.1520

Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., & McAuley, J. (2020). Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. arXiv. https://doi.org/10.48550/arXiv.2002.12749

Jones, R. (2024, May 2). Fraudsters editing vehicle photos to add fake damage in UK insurance scam. The Guardian. https://www.theguardian.com/business/article/2024/may/02/car-insurance-scam-fake-damaged-added-photos-manipulated

Lalchand, S., Srinivas, V., Maggiore, B., & Henderson, J. (2024, May 29). Generative AI Is Expected to Magnify the Risk of Deepfakes and Other Fraud in Banking. Deloitte Insights; Deloitte. https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-predictions/2024/deepfake-banking-fraud-risk-on-the-rise.html

Li, Y., Chang, M.-C., & Lyu, S. (2018, December 1). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. IEEE Xplore. https://doi.org/10.1109/WIFS.2018.8630787

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr42600.2020.00327

Li, M. and Wan, Y. (2023), "Norms or fun? The influence of ethical concerns and perceived enjoyment on the regulation of deepfake information", Internet Research, Vol. 33 No. 5, pp. 1750-1773. https://doi.org/10.1108/INTR-07-2022-0561

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. arXiv. https://doi.org/10.48550/arXiv.2004.11138

Needham, R., Mulligan, C., & Hamilton-Wood, F. (2024, November 5). Deepfakes in the insurance market – a personal injury perspective. Kennedys Law; Kennedys Law LLP. https://kennedyslaw.com/en/thought-leadership/article/2024/deepfakes-in-the-insurance-market-a-personal-injury-perspective/

Nicos Vekiarides. (2024, July 17). Viewpoint: Deepfake Fraud Is On the Rise. Here's How Insurers Can Respond. Insurance Journal. https://www.insurancejournal.com/news/national/2024/07/17/784226.htm

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. IEEE. https://doi.org/10.1109/ACCESS.2022.3154404

Rana, M. S., & Sung, A. H. (2020). DeepfakeStack: A deep ensemble-based learning technique for deepfake detection. IEEE. https://doi.org/10.1109/CSCloud-EdgeCom49738.2020.00021

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision, 1–11. https://doi.org/10.1109/ICCV.2019.00009

Satariano, A. (2023). Insurance firms struggle to keep up with AI-generated fraud. The New York Times. https://www.nytimes.com

Swatton, P., & Leblanc, M. (2024, June 7). What are deepfakes and how can we detect them? The Alan Turing Institute. https://www.turing.ac.uk/blog/what-are-deepfakes-and-how-can-we-detect-them

Tuysuz, M. K., & Kılıç, A. (2023). Analyzing the legal and ethical considerations of deepfake technology. Interdisciplinary Studies in Society, Law, and Politics, 2(2), 4–10. https://doi.org/10.61838/kman.isslp.2.2.2

Wang, S., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. https://doi.org/10.48550/arXiv.1912.11035

Zurich. (2024, January 22). Insurance must prepare for a rise in deepfake AI fraud. Zurich.co.uk; Zurich. https://www.zurich.co.uk/news-and-insight/insurance-must-prepare-for-a-rise-in-deepfake-ai-fraud