



## OPEN ACCESS

### EDITED BY

Xin Wang,  
University at Albany, United States

### REVIEWED BY

Gregory Gondwe,  
California State University, San Bernardino,  
United States  
Radhika Baskar,  
Saveetha University, India

### \*CORRESPONDENCE

Ibidun C. Obagbuwa  
✉ iobagbuwa@wsu.ac.za

RECEIVED 02 November 2025

REVISED 27 December 2025

ACCEPTED 19 January 2026

PUBLISHED 02 March 2026

### CITATION

Moyo BVC, Tuyikeze T, Matsebula F and  
Obagbuwa IC (2026) An AI-driven  
conceptual framework for detecting  
fake news and deepfake content: a  
systematic review.  
*Front. Artif. Intell.* 9:1737790.  
doi: 10.3389/frai.2026.1737790

### COPYRIGHT

© 2026 Moyo, Tuyikeze, Matsebula and  
Obagbuwa. This is an open-access  
article distributed under the terms of the  
[Creative Commons Attribution License  
\(CC BY\)](#). The use, distribution or  
reproduction in other forums is  
permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication  
in this journal is cited, in accordance  
with accepted academic practice. No  
use, distribution or reproduction is  
permitted which does not comply with  
these terms.

# An AI-driven conceptual framework for detecting fake news and deepfake content: a systematic review

Bravlyn VC. Moyo<sup>1</sup>, Tite Tuyikeze<sup>1</sup>, Fezile Matsebula<sup>1</sup> and  
Ibidun C. Obagbuwa<sup>1,2\*</sup>

<sup>1</sup>Department of Computer Science & Information Technology, Faculty of Natural and Applied Sciences, Sol Plaatje University, Kimberley, South Africa, <sup>2</sup>Department of Mathematical Sciences and Computing, Faculty of Natural Sciences, Walter Sisulu University, Mthatha, South Africa

The rapid advancement of generative artificial intelligence (AI) has enabled the creation of highly realistic synthetic media, commonly referred to as deepfakes, which are increasingly multimodal and difficult to detect. While these technologies offer creative and commercial potential, they also pose critical challenges related to misinformation, media trust, and societal harm. Despite the growing body of research, existing reviews remain fragmented, often separating technical detection advances from social and governance considerations. This study addresses this gap through a systematic review conducted in accordance with PRISMA guidelines across IEEE Xplore, Scopus, ACM Digital Library, and Web of Science. From an initial set of 120 database records, complemented by citation chaining, 34 studies published between 2014 and 2025 were included for analysis. Eighteen studies focused on deepfake generation and detection models, eight examined social and behavioural implications, and eight addressed ethical and regulatory frameworks. Thematic synthesis reveals a clear methodological shift from convolutional neural networks toward transformer- and CLIP-based architectures, alongside the emergence of large-scale benchmark datasets. However, persistent challenges remain in multimodal detection, cross-dataset generalization, explainability–robustness trade-offs, and the translation of governance principles into deployable systems. This review contributes an integrated conceptual framework that operationally connects detection technologies, explainable AI (XAI), and governance mechanisms through explicit feedback loops. Future research directions emphasize robust multimodal benchmarks, retrieval-augmented detection systems, and interdisciplinary approaches that align technical innovation with ethical and policy safeguards.

### KEYWORDS

deepfakes, explainable AI (XAI), generative artificial intelligence, media trust, misinformation, multimodal detection

## 1 Introduction

Generative AI has emerged as a transformative force in digital content creation. Among its most striking applications are deepfakes synthetically generated or manipulated videos, images, and audio that can convincingly imitate real individuals. Initially conceived as technical demonstrations, deepfakes have evolved into powerful tools with dual-use potential, supporting both creative innovation and malicious activities such as non-consensual sexual content, political misinformation, and

reputational harm. Recent studies have investigated deepfake creation and identification in visual, audio, and multimodal domains, as well as the social and cognitive impacts of misinformation (Chuk-Ke and Dong, 2024; Donahue et al., 2019; Gowrisankar and Thing, 2024; Green and Swets, 1966; Kumar et al., 2022; Loth et al., 2024; Nguyen et al., 2020, 2022; Pearson and Zinets, 2022; Rossler et al., 2019; Siarohin et al., 2019; Shu et al., 2020; Sweller, 1988; Zhang et al., 2024).

The implications of deepfakes extend well beyond technical domains, intersecting with media trust, democratic governance, and legal accountability. Social science research has examined misinformation dynamics (Zhou et al., 2021; Idiongo, 2024), while policymakers have begun formulating governance frameworks such as the EU Digital Services Act (European Parliament and Council of the European Union (EU DSA), 2022) and the EU AI Act (European Parliament and Council of the European Union (EU AI Act), 2024). In parallel, computer vision research has advanced rapidly, developing datasets and detection methods based on convolutional and transformer architectures (Verdoliva et al., 2019; Li et al., 2020).

Despite these parallel developments, current literature remains fragmented. Few studies systematically integrate technical detection methods with social, ethical, and policy perspectives. This review addresses that gap by synthesizing interdisciplinary research to provide a unified understanding of deepfake creation, detection, and governance. Specifically, it seeks to:

- i Identify dominant technical and social approaches of deepfake generation and detection.
- ii Evaluate how explainable AI and multimodal architectures enhance detection robustness.
- iii Examine how regulatory and ethical frameworks can inform the design of responsible detection systems.

By combining insights from computer vision, natural language processing, explainable AI, and social science, this review provides a cross-disciplinary taxonomy of deepfake research and outlines a conceptual framework for integrating detection, explainability, and governance. Through this synthesis, it aims to support the development of transparent, ethical, and technically resilient AI-based systems for mitigating the harms of synthetic media.

In line with contemporary academic consensus, this review avoids treating “fake news” as a standalone analytical category due to its politicized and ambiguous usage. Rather, the terms misinformation, disinformation, and AI-created misleading content are employed to differentiate unintentional errors, intentional distortion, and artificial media products, respectively. This terminological precision enables clearer alignment between technical detection methods, social impact studies, and regulatory frameworks.

## 2 Methodology

### 2.1 Review protocol and literature selection

This study adopted a systematic review approach to synthesize current research on deepfakes, generative Artificial Intelligence (AI), and misinformation. The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure methodological transparency, replicability, and rigor. The process involved six key stages: (i) defining the research scope and

questions, (ii) developing search strategies and selecting databases, (iii) applying inclusion and exclusion criteria, (iv) screening and extracting relevant data, (v) assessing methodological quality, and (vi) synthesizing and interpreting the findings.

The scope of this review covered the period 2014 to 2025, reflecting the evolution of deepfake research from its early technical demonstrations to contemporary multimodal and regulatory developments. Both qualitative and quantitative evidence were considered to capture the interdisciplinary nature of deepfake research across technical, social, and policy domains.

### 2.2 Search strategy and databases

A comprehensive multi-database search strategy was designed to capture the full breadth of deepfake-related research, including studies on generation, detection, misinformation, and regulation. The Boolean search string used (Scopus format) was:

“deepfake\*” OR “fake news” OR “synthetic media” OR “AI-generated content” OR “misinformation” OR “disinformation”) AND (“artificial intelligence” OR “machine learning” OR “deep learning” OR “neural network\*” OR “natural language processing” OR “computer vision”)

From an initial pool of 120 database records, complemented by backward and forward citation chaining, this query was executed and adapted for each database, including IEEE Xplore, ACM Digital Library, Scopus, SpringerLink, and Web of Science. To ensure inclusion of the most recent advances, arXiv preprints were also screened, recognizing that peer review often lags rapid developments in generative AI.

Preprints were included only when they were later cited in peer-reviewed venues or associated with publicly recognized datasets to mitigate quality concerns.

For regulatory and governance perspectives, official EU repositories and government sites were reviewed to obtain key policy texts such as the EU Digital Services Act (European Parliament and Council of the European Union (EU DSA), 2022) and the EU AI Act (European Parliament and Council of the European Union (EU AI Act), 2024). Reference lists of selected articles were also examined through backward and forward citation chasing to identify additional relevant studies not captured by the initial search.

arXiv preprints were included selectively due to the rapid pace of advances in generative AI and were cross-validated based on subsequent peer-reviewed adoption, dataset impact, or citation prominence.

### 2.3 Inclusion and exclusion criteria

**Inclusion criteria:**

Studies were included if they met the following criteria:

- i Published between 2014 and 2025;
- ii Peer-reviewed journal articles, conference papers, or preprints focusing on deepfake generation, detection, or governance;
- iii Addressed technical, social, or policy/regulatory dimensions of deepfakes;
- iv Investigated text, audio, image, or video modalities related to synthetic media;
- v Written in English and available in full-text form.

**Exclusion criteria:**

- i Studies published in languages other than English;
- ii Duplicate records across multiple databases;

- iii Non-peer-reviewed sources such as blog posts, news articles, and unverified reports (except official institutional or legal documents);
- iv Studies focusing solely on unrelated AI applications (e.g., generative art) without relevance to misinformation or deepfake detection;
- v Theoretical discussions lacking empirical or methodological depth.

These criteria ensured that the final selection consisted of studies with direct relevance, methodological rigor, and conceptual clarity.

## 2.4 Search strategy

The search strategy combined structured database keyword searches with both backward citations chasing (examining the references of included studies) and forward citation chasing (tracking newer works that cited them). Initial queries employed broad terms such as *deepfake detection*, *generative adversarial networks (GANs)*, *synthetic media*, *misinformation*, and *fake news*. Throughout the review, the strategy was refined repeatedly to incorporate new methods and ideas. Other keywords comprised transformers, Vision Transformer (ViT), CLIP, voice generation, multimodal recognition, Xplainable AI (XAI), retrieval-augmented production (RAG), bias, equity, governance, and regulation. Boolean operators (e.g., “deepfake AND identification”, “synthetic media AND false information”, “transformer OR CLIP AND deepfake”) were employed to enhance recall and precision. This iterative approach ensured that both foundational studies and the most recent developments in technical, social, and regulatory domains were systematically captured.

## 2.5 Screening process

The screening process followed a three-stage procedure to ensure systematic inclusion of high-quality and relevant studies:

- 1 Title and abstract screening: all retrieved records were reviewed to exclude irrelevant topics, such as unrelated computer vision tasks or non-AI media analyses.
- 2 Full-text review: remaining articles were evaluated for methodological soundness, empirical contribution, and relevance to the research objectives.
- 3 Thematic categorization: eligible studies were coded into thematic clusters representing:
  - (i) Dataset creation and benchmark development.
  - (ii) Detection models and architectures.
  - (iii) Explainability and adversarial robustness.
  - (iv) Social trust and misinformation.
  - (v) Policy and regulatory frameworks.

At the full-text screening stage, studies were excluded for the following predefined reasons:

Reason 1: Primary focus on AI applications unrelated to misinformation or deceptive synthetic media ( $n = 6$ ).

Reason 2: Lack of empirical evaluation, methodological transparency, or reproducible analysis ( $n = 5$ ).

Reason 3: Conceptual or opinion-based papers without sufficient analytical depth or evidence synthesis ( $n = 7$ ).

## 2.5.1 Data extraction process

Data extraction was conducted using a structured data extraction form developed in Microsoft Excel. Two reviewers independently extracted data to minimize bias, with discrepancies resolved through discussion and consensus.

Extracted fields included:

- i Author(s), year, and publication type.
- ii Research domain and methodology.
- iii AI model or framework (e.g., GAN, transformer, CLIP).
- iv Dataset characteristics and evaluation metrics.
- v Key findings and thematic relevance.

This approach ensured consistency and completeness in capturing both technical and contextual details.

Inter-rater reliability between reviewers was assessed using Cohen’s kappa coefficient ( $\kappa$ ), which indicated strong agreement during the screening and data extraction phases.

## 2.6 Number of studies included

The systematic review included a total of 34 studies published between 2014 and 2025, capturing both the historical development and the latest advances in deepfake research. About 18 of them examined deepfake production, detection methods, and model architectures with an emphasis on computer vision and technical approaches. About eight studies from the social sciences looked at how deepfakes affect user behaviour, media trust, and disinformation in society. The other 8 studies focused on policy and regulatory frameworks, emphasizing ethical issues, governance strategies, and legal actions. Research was chosen for its capability to offer a broad viewpoint across various media formats text, images, video, and audio and for its role in enhancing knowledge of the technical, social, and policy aspects of deepfakes. This selection ensures a holistic view of the field, combining insights into both innovation and the societal challenges posed by synthetic media.

Although the ultimate set of 34 studies may seem narrow compared to the vastness of generative AI research, this demonstrates the careful use of stringent inclusion standards that prioritize interdisciplinary significance, methodological soundness, and clear involvement with detection, social consequences, or governance. This trade-off prioritizes analytical depth and coherence over exhaustive coverage and is acknowledged as a limitation of the review.

## 2.7 Quality assessment

The methodological quality and potential risk of bias for each study were evaluated using an adapted version of the Critical Appraisal Skills Program (CASP) checklist. Technical studies were assessed based on criteria such as dataset transparency, model reproducibility, and evaluation robustness, while social science and policy studies were reviewed for methodological clarity, validity of interpretation, and evidence linkage. Each study received a quality rating (high, moderate, or low) based on these criteria, which informed the weight given during synthesis of the 34 studies evaluated, 21 were rated high quality, 9 moderate, and 4 low according to the adapted CASP checklist.

## 2.8 Rationale for selection

The studies included in this review were selected for several key reasons. They first offer a historical basis for comprehending the evolution of generative AI and the rise of misinformation by illustrating the progression of deepfakes from early technical demos to highly advanced synthetic media. Second, they showcase state-of-the-art technological detection techniques, such as multimodal analysis, explainable AI, and machine learning model advancements. Third, the research investigates sociological, ethical, and legal aspects, analysing how deepfakes influence media credibility, public opinion, privacy, and regulatory systems. Finally, the selection deliberately encompasses a range of media modalities text, images, video, and audio to ensure a comprehensive understanding of both the technical challenges and societal implications of deepfakes (see Figure 1).

## 3 Thematic review of literature

### 3.1 Overview of thematic analysis

A thematic review approach was used to organize and interpret the selected studies according to recurring patterns, concepts, and

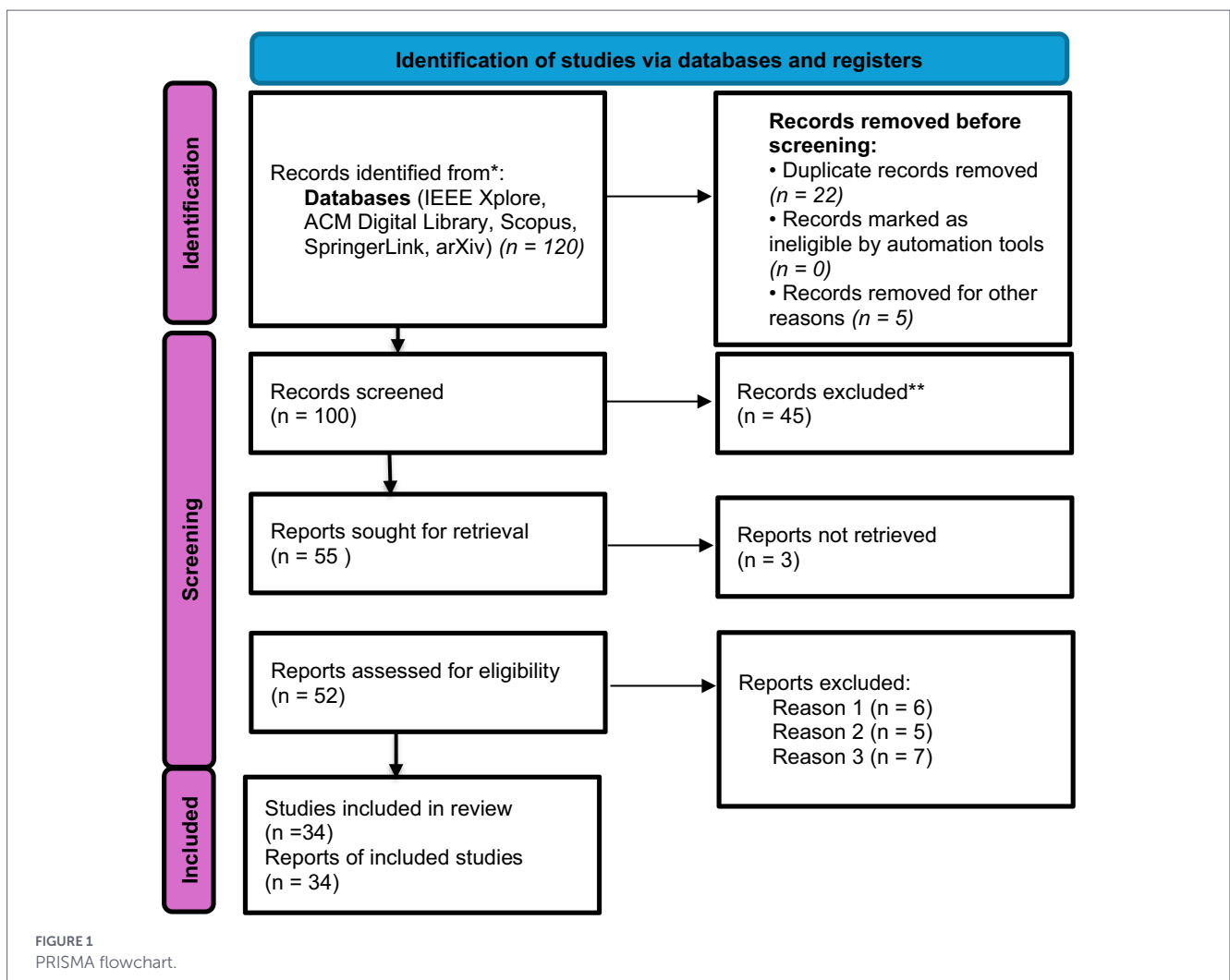
research priorities. Finding commonly discussed subjects, assembling related concepts into clusters, and honing these clusters into cohesive themes were the steps in the analysis process. This approach facilitated the emergence of both chronological and conceptual links, underscoring the progression of research on deepfakes and synthetic media from initial technical trials to wider societal and policy-focused conversations.

From this analysis, four overarching themes were identified:

- i Evolution of Deepfake Technologies and Detection Research;
- ii Technical developments in deepfake generation and detection methods;
- iii Explainability, robustness, and evaluation challenges in AI-based detection;
- iv Social and psychological impacts, including misinformation, media trust, and user behaviour;
- v Governance, ethics, and policy frameworks addressing regulation and accountability.

### 3.2 Evolution of deepfake technologies and detection research

The trajectory of deepfake development mirrors broader advances in generative AI. Early methods, notably Generative Adversarial



Networks (GANs) (Goodfellow et al., 2014) established the adversarial paradigm that underpins synthetic image and video generation. Subsequent variants such as StyleGAN2/3 (Lehtinen & Aila NVIDIA) dramatically improved realism and controllability. Meanwhile, transformer architectures (Vaswani et al., 2017) and large-scale language models (Brown et al., 2020) expanded synthetic generation to text and audio, creating a multimodal threat landscape.

These architectural advances not only enabled high-fidelity media synthesis but also reshaped detection research. As deepfakes grew more realistic, early CNN-based detectors emphasizing pixel-level inconsistencies gave way to transformer and CLIP-based frameworks capable of modelling contextual and semantic relationships. This shift reflects an evolution from surface-level artifact detection toward cross-modal, meaning-aware analysis, mirroring the progression of misinformation itself from isolated falsifications to integrated multimodal narratives.

This architectural transition reflects more than incremental performance improvement. Detectors based on CNNs chiefly leverage low-level visual artifacts, which modern generative models increasingly reduce. Transformer and CLIP-based approaches instead model long-range spatial, temporal, and semantic dependencies, enabling improved robustness to compression and post-processing. Nonetheless, these benefits come with trade-offs in computational expense, data reliance, and interpretability. The literature rarely addresses how such models can be deployed in real-time or resource-constrained environments, revealing a gap between benchmark success and operational feasibility (see Table 1).

### 3.3 Comparative analysis

To synthesize and contrast the findings across the selected studies, a comparative analysis was performed focusing on methodological design, evaluation performance, and disciplinary orientation. To find similarities, contrasts, and latest trends, this analysis combines data from computer vision, multimodal AI, social science, and policy research.

A comparative analysis identified three primary methodological clusters: CNN-based, Transformer-based, and CLIP-based/multimodal frameworks each presenting unique advantages and disadvantages. CNN architectures (e.g., Section 3: Thematic Review/Detection Models, EfficientNet) exhibit high accuracy on established

benchmarks such as FaceForensics++ and DFDC but demonstrate limited generalization to unseen datasets or manipulation techniques. Although they require more data and have greater computing costs, Vision Transformers (ViT, Swin-T) and hybrid models provide better contextual modeling and modest improvements in cross-dataset generalization. CLIP-based and multimodal methods utilize vision-language pretraining to enable zero-shot and few-shot detection, indicating scalability potential across modalities, yet they continue to be vulnerable to adversarial perturbations and domain shifts.

Recent research has delved deeper into multimodal fusion and extensive pretraining techniques to enhance generalization and resilience in deepfake detection, especially in cross-dataset and real-world scenarios (Chen et al., 2024a; Chen et al., 2024b).

Foundational deep learning architectures underpinning contemporary detection systems include Xception networks (Chollet, 2017), transformer-based language models such as BERT (Devlin et al., 2019), and generative adversarial networks including StyleGAN (Karras et al., 2021). These architectures have directly impacted the generation pipelines for deepfakes and the detection methods assessed in benchmark datasets like FaceForensics++ (Dolhansky et al., 2019).

From a disciplinary standpoint, methodological research use diverges:

- i Computer vision studies prioritize quantitative metrics such as AUC, accuracy, and F1-score, emphasizing model robustness and scalability.
- ii Social science research focuses on user perception, misinformation spread, and trust restoration, prioritizing ecological validity over computational precision.
- iii Policy and governance studies emphasize legal accountability, transparency mechanisms, and platform obligations, focusing less on algorithms and more on institutional enforcement.

#### 3.3.1 Quantitative overview of methodological clusters

Among the 34 studies analysed, approximately 59% ( $n = 20$ ) employed CNN-based detection models, 26% ( $n = 9$ ) used transformer or hybrid architectures, and 15% ( $n = 5$ ) adopted multimodal or CLIP-based approaches. Social science and policy-focused studies

TABLE 1 Summary of emerging themes in the current state of knowledge.

Theme	Focus area	Key findings	Representative studies/datasets
1. Advances in datasets and detection models	Development of datasets and model architectures for detection	Expansion from CNNs to Transformer-based and CLIP-integrated models; improved generalization via multimodal benchmarks	FaceForensics++, DFDC, Celeb-DF, FakeVoices, WaveFake, Deepfake-Eval-2024
2. Explainability and adversarial robustness	Interpretability and resilience of detection systems	Use of Grad-CAM and similar methods; emerging risks from explainability-based attacks	Selvaraju et al. (2017)
3. Social, ethical, and policy responses	Societal impacts and governance frameworks	Rising misinformation, erosion of trust, regulatory responses via EU AI and DSA Acts	Idiongo (2024), National Sexual Violence Resource Center (2024), and European Parliament and Council of the European Union (EU DSA) (2022, 2024)

comprised ~30% of the total dataset, underscoring the growing interdisciplinary scope of deepfake research.

The predominance of CNN-based approaches (59%) reflects both historical inertia and dataset availability, as many widely used benchmarks were designed to expose CNN-detectable artifacts. While this has accelerated short-term performance gains, it may also constrain innovation by incentivizing dataset-specific optimization rather than real-world generalization. This methodological concentration highlights the need for evaluation protocols that reward robustness, multimodal reasoning, and cross-domain adaptability (see Table 2).

### 3.3.2 Interpretation and implications

The comparative synthesis highlights that while technical advancements in AI detection (e.g., transformers, CLIP) show promise, cross-domain generalization and ethical integration remain unresolved. Complementary insights on trust and governance are offered by social and policy studies, indicating the necessity of interdisciplinary frameworks that combine societal resilience, regulatory oversight, and technical detection accuracy. These comparative insights form the foundation for the conceptual framework presented in the next section, integrating strengths across disciplines to advance comprehensive deepfake mitigation strategies.

## 3.4 Conceptual framework

The framework integrates technical, social, and governance dimensions through feedback loops that connect model transparency, user trust, and regulatory compliance to build accountable and resilient AI detection systems (see Figure 2).

The framework is organized in a three-tier system:

- 1 Technical layer – encompassing detection models, detection outcomes and robust datasets;

- 2 Social layer – focusing on media literacy, trust perception, and behavioural impact;
- 3 Policy & governance layer – covering regulation, ethics, and platform accountability;
- 4 Explainability mechanisms – dealing with model transparency, user feedback, perception data, joint influence on responsible AI deployment and societal resilience.

These components are interconnected through bidirectional feedback loops. Technical detection systems generate transparent outputs that inform social understanding and policy decisions, while governance structures provide ethical and compliance feedback to guide model development and user interaction.

This framework extends prior conceptual models (Mirsky and Lee, 2021; Zhou et al., 2019) by explicitly incorporating explainability and feedback dynamics that connect machine learning performance with human trust and policy accountability. Unlike earlier models focusing solely on detection or social impacts, this integrated design emphasizes *mutual reinforcement* between technology, transparency, and governance to create a resilient AI ecosystem for mitigating deepfake threats.

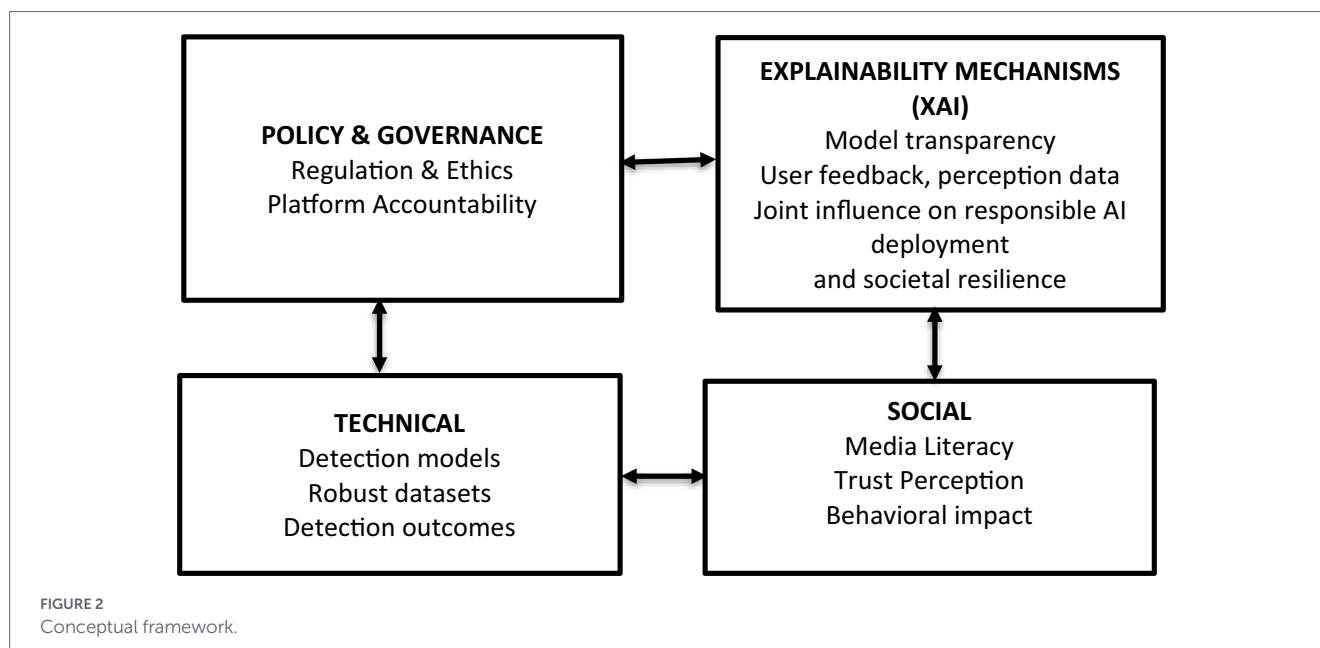
### 3.4.1 Operationalization of the conceptual framework

The framework is designed to function as an operational model rather than a static taxonomy. In the technical layer, detection models analyze multimodal content and produce classification results along with explainability indicators like confidence scores or attention maps. These outputs directly influence the social layer by shaping moderation decisions, media literacy interventions, and user trust calibration.

User engagement, such as requests, modifications, and sharing actions, generates feedback data that returns to the technical layer,

TABLE 2 Comparative summary of detection approaches and research perspectives.

Approach domain	Representative studies	Core techniques	Strengths	Limitations	Evaluation focus metrics
CNN-based detection	Li et al., 2020 and Verdoliva et al. (2019)	XceptionNet, EfficientNet, ResNet	High benchmark accuracy, efficient training	Poor cross-dataset generalization; sensitive to new manipulations	AUC, F1-score, accuracy
Transformer-based models	Chen et al. (2024a, 2024b) and Wang et al. (2024)	ViT, Swin Transformer, hybrid CNN-Transformer	Captures long-range dependencies, strong contextual modelling	Data-hungry, computationally expensive	Accuracy, cross-dataset robustness
CLIP multimodal approaches	Yermakov et al. (2025) and Brown et al., 2025	Vision-language pretraining, frozen CLIP features	Zero-shot and multimodal generalization; scalable	Adversarial vulnerability, distribution shift	Zero-shot accuracy, multimodal retrieval performance
Explainable & robust AI	Kozik et al. (2024) and Selvaraju et al. (2017)	Grad-CAM, feature attribution, robustness testing	Transparency, user interpretability	Risk of adversarial exploitation	Interpretability quality, robustness metrics
Social behavioural studies	Idiongo (2024) and Zhou et al. (2021)	Surveys, experiments, discourse analysis	Insights into trust, misinformation spread	Lack of quantitative precision	User trust, perception, misinformation spread
Policy & regulatory studies	European Parliament and Council of the European Union (EU DSA) (2022) and European Parliament and Council of the European Union (EU AI Act) (2024)	Legislative review, policy analysis	Governance, accountability frameworks	Limited enforcement mechanisms, lag behind tech	Policy compliance, transparency obligations



guiding dataset improvement, bias assessment, and model updates. Simultaneously, the policy and governance layer constrains system behaviour through transparency obligations, accountability requirements, and platform-level enforcement mechanisms, which shape both technical design and social response strategies.

In practical scenarios such as election integrity or non-consensual deepfake mitigation, this closed-loop structure enables continuous alignment between detection accuracy, societal trust, and regulatory compliance.

In contrast to earlier multi-layer frameworks that mainly classify stakeholders or research areas, this framework specifically represents feedback loops among technical outputs, human understanding, and governance limitations. This enables evaluation of how design choices in detection and explainability propagate through social trust and regulatory accountability, making the framework applicable to system design, policy assessment, and interdisciplinary research planning.

### 3.5 Knowledge gaps

Despite significant advances in deepfake detection, several key gaps and contentious issues remain. Generalization is a major challenge, as many detectors perform well on benchmark datasets but struggle to maintain accuracy across new or unseen datasets, limiting their real-world applicability. Multimodal deepfakes, which combine video, audio, and text, are understudied compared to single-modality cases, leaving detection methods less prepared for increasingly sophisticated manipulations. The balance between explainability and security also presents a dilemma: while explainable AI tools improve transparency, they can potentially be exploited by adversaries to bypass detection. Additionally, ethical and legal protections remain insufficient, and harmful such as non-consensual deepfake pornography and political disinformation highlight gaps in current laws and enforcement. Finally, the lack of standardized, adversarial resilient evaluation benchmarks complicates the comparison of detection methods across studies, making it difficult to assess progress and deploy robust solutions in practice.

## 4 Critical discussion

The literature reveals a field advancing rapidly on the technical front while struggling to ensure adequate societal protection. Although detection models have improved significantly, the adversarial dynamics of the problem mean that innovations in synthesis frequently outpace detection capabilities. This ongoing arms race is intensified by dataset biases, inconsistent reporting standards, and fragmented evaluation practices.

In addition, the interdisciplinary gaps persist until today. Social scientists extensively document the social and behavioural consequences of misinformation but often neglect the technical constraints of detection systems. Conversely, many computer vision studies validate models using synthetic benchmarks without evaluating real-world or societal implications. Regulatory initiatives such as the EU AI Act and Digital Services Act introduce frameworks for accountability, yet translating these high-level principles into enforceable technical standards and platform practices remains an open challenge.

Previous surveys and forensic studies established the foundation for contemporary detection research by systematizing manipulation categories and assessment methods (Verdoliva et al., 2019; Zhou et al., 2021).

Deployment challenges also include scalability, privacy protection when handling sensitive data, and the ethical management of false positives that could affect legitimate content. These challenges directly relate to the conceptual framework proposed in this review: weaknesses in the *technical domain* (e.g., bias, overfitting) affect *social trust and media literacy*, while insufficient *policy enforcement* weakens governance feedback loops. A more integrated research agenda that unites robust technical evaluation, human-centred design, and co-developed regulatory mechanisms is necessary to ensure that progress in detection contributes meaningfully to societal protection.

Recent regulatory initiatives further underscore the need for accountable AI-driven detection systems. The Digital Services Act (European Parliament and Council of the European Union (EU DSA), 2022) of the European Union establishes requirements for platform transparency and risk reduction concerning online misinformation,

while the proposed European Union Artificial Intelligence Act (European Parliament and Council of the European Union (EU AI Act), 2024) classifies certain AI-driven content moderation and synthetic media systems as high-risk, imposing requirements for explainability, documentation, and human oversight.

## 5 Future directions

Future research on deepfake detection must move beyond narrow, dataset-driven evaluations toward holistic, context-aware, and ethically aligned systems. Guided by the conceptual framework, four thematic directions namely technical, methodological, ethical, and policy/governance are proposed to structure future development.

### 5.1 Technical directions

Future detection models should reflect the complexity of multimodal manipulations and evolving threat landscapes:

- i Multimodal Benchmark Development: Construct datasets that jointly assess image, video, audio, and text manipulations to capture cross-modal deepfake narratives.
- ii Cross-Domain Generalization: Mitigate overfitting through domain adaptation, self-supervised learning, and transfer learning to enhance robustness across datasets and manipulation types.
- iii Retrieval-Augmented Detection: Incorporate external evidence (e.g., verified media or fact databases) into detection reasoning for improved factual grounding and reliability.
- iv Explainability and Robustness: Advance interpretable AI methods (e.g., saliency maps, attention visualizations) that clarify decisions and resist adversarial exploitation.

### 5.2 Methodological directions

- i Adversarial-Resilient Evaluation Protocols: Establish standardized benchmarks that test model performance under realistic, adversarial, and cross-cultural conditions.
- ii Human-AI Collaboration: Design hybrid systems where explainable AI tools assist human reviewers, journalists, and policymakers in verifying authenticity and contextual accuracy.
- iii Longitudinal and Cross-Platform Studies: Examine how detection systems perform across time and social media ecosystems to measure real-world efficacy and adaptation.

### 5.3 Ethical directions

- i Data Privacy and Consent: Ensure deepfake datasets use consensual, privacy-preserving data to avoid reinforcing exploitation or harm.
- ii Bias and Fairness Auditing: Implement fairness checks and demographic audits in training datasets to prevent disproportionate impacts on marginalized groups.
- iii Transparency and Accountability: Promote open reporting of model architectures, performance metrics, and limitations to support reproducibility and ethical oversight.

## 5.4 Policy and governance directions

- i Regulatory alignment: translate policy instruments such as the *EU AI Act* and *Digital Services Act* into technical compliance standards, platform-level transparency, and auditable mechanisms.
- ii Global governance frameworks: encourage international cooperation to develop harmonized policies for detecting and labelling synthetic media.
- iii Public education and media literacy: foster interdisciplinary collaboration between AI developers, educators, and communication experts to build societal resilience against misinformation.

### 5.4.1 Practical implications

For researchers, this framework encourages the integration of technical and social dimensions, ensuring that advances in model accuracy are matched by attention to transparency and fairness. Policy makers can use it to align AI regulation with technical feasibility, creating compliance standards that are both enforceable and adaptable. Media platforms can operationalize these insights by embedding explainable detection tools within content moderation pipelines. Collectively, these directions emphasize that the sustainability of deepfake detection depends on continuous interaction between innovation, ethics, and governance.

## 6 Conclusion

This review synthesizes technical, social, and policy literature on deepfakes and detection. While there has been commendable progress in detection techniques and benchmark creation, significant challenges remain in generalization, multimodal detection, ethical safeguards, and policy enforcement. Many current models achieve strong results on specific datasets, yet their reliability weakens in real-world contexts where manipulations are more diverse and adversarial adaptive. Multimodal deepfakes, which combine video, audio, and textual fabrications, further complicate detection by exploiting gaps between unimodal approaches.

Beyond technical hurdles, unresolved ethical and legal questions persist. Non-consensual sexual deepfakes and political disinformation highlight the potential for severe personal and societal harm, underscoring the urgency of legal protections, content moderation frameworks, and victim support mechanisms. Policies such as the EU AI Act and Digital Services Act provide valuable blueprints, but enforcement depends on effective translation into technical standards and platform practices. At the same time, overzealous regulation risks constraining legitimate research and creative uses of generative technologies, demanding careful balance.

Addressing these challenges requires more than technical innovation; it demands sustained interdisciplinary collaboration. Engineers, social scientists, policymakers, and ethicists must co-design solutions that are simultaneously robust, explainable, and aligned with democratic values. Evaluation standards should evolve toward adversarial resilient benchmarks that mirror deployment conditions, while explainability tools must empower human reviewers without exposing new attack surfaces.

The promise and perils of generative AI demand coordinated responses that protect individuals and societies while enabling beneficial innovation. If pursued with transparency, inclusivity, and foresight, the next generation of deepfake detection systems

and governance structures can mitigate harms while fostering responsible use of synthetic media in education, entertainment, accessibility, and beyond. The trajectory of this field will depend on whether the global community can act not only to keep pace with technological advances, but also to shape them in the service of public trust and social good.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: FakeNewsNet for text, FaceForensics++ for video.

## Author contributions

BM: Conceptualization, Investigation, Writing – original draft, Writing – review & editing. TT: Methodology, Supervision, Writing – original draft. FM: Formal analysis, Project administration, Writing – original draft, Writing – review & editing. IO: Conceptualization, Formal analysis, Investigation, Writing – review & editing.

## Funding

The author(s) declared that financial support was not received for this work and/or its publication.

## References

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. Neural Inf. Proc. Syst.* 33, 1877–1901.
- Brown, T., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., et al. (2025). Deepfake-Eval-2024: a multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv*.
- Chen, Y., Zhang, L., and Niu, Y. (2024a). Forgelen: data-efficient forgery focus for generalizable forgery image detection. *arXiv*.
- Chen, Y., Zhang, L., and Niu, Y. (2024b). Guided and fused: efficient frozen CLIP-ViT with feature guidance and multi-stage feature fusion for generalizable deepfake detection. *arXiv*.
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *CVPR*, 1800–1807.
- Chuk-Ke, C., and Dong, Y. (2024). Misinformation and literacies in the era of generative artificial intelligence: a brief overview and a call for future research. *Emerg. Media* 2, 70–85. doi: 10.1177/27523543241240285
- Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. *NAACL*. 417, 4171–4186. doi: 10.18653/v1/N19-1423
- Dolhansky, B., Howes, R., Pflaum, B., Baram, N., and Ferrer, C. C. (2019). The DeepFake detection challenge dataset. *arXiv*.
- Donahue, C., McAuley, J., and Puckette, M. (2019). WaveGAN: spectral audio synthesis with generative adversarial networks. *arXiv*.
- European Parliament and Council of the European Union (EU DSA) (2022). Regulation (EU) 2022/2065 on a single market for digital services: Official Journal of the European Union.
- European Parliament and Council of the European Union (EU AI Act) (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence: Official Journal of the European Union.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets" in *Advances in neural information processing systems*. Eds. J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, vol. 27.
- Gowrisankar, B., and Thing, V. L. L. (2024). An adversarial attack approach for explainable AI evaluation on deepfake detection models. *Comput. Secur.* 139:103684. doi: 10.1016/j.cose.2023.103684
- Green, D. M., and Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Idiongo, P. (2024). The impact of fake news on public trust in traditional media outlets. *J. Commun.* 5. doi: 10.47941/jcomm.1984
- Karras, T., Laine, S., and Aila, T. (2021). Analyzing and improving the image quality of StyleGAN3: CVPR, 8110–8119.
- Kozik, R., Ficco, M., Pawlicka, A., Pawlicki, M., Palmieri, F., and Choraś, M. (2024). When explainability turns into a threat—using XAI to fool a fake news detection method. *Comput. Secur.* 137:103599. doi: 10.1016/j.cose.2023.103599
- Kumar, A., Nguyen, T., and Li, J. (2022). WaveFake: a dataset for synthetic voice detection: Interspeech, 2110–2114.
- Li, Y., Chang, M. C., and Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics: CVPR, 3207–3216.
- Loth, A., Kappes, M., and Pahl, M. O. (2024). Blessing or curse? A survey on the impact of generative AI on fake news. *arXiv*.
- Mirsky, Y., and Lee, W. (2021). The creation and detection of deepfakes: a survey. *ACM Comput. Surv.* 54, 1–41. doi: 10.1145/3425780
- National Sexual Violence Resource Center. 2024. Taylor Swift and the dangers of deepfake pornography. Available online at: <https://www.nsvrc.org/blogs/feminism/taylor-swift-and-dangers-deepfake-pornography>
- Nguyen, H., Fang, F., Yamagishi, J., and Echizen, I. (2020). Deep learning for deepfake detection: analysis and challenges. *IEEE Trans. Inf. Forensics Secur.* 15, 1879–1893.
- Nguyen, T., Nguyen, C. M., Nguyen, D., and Nahavandi, S. (2022). Fakevoices: dataset for synthetic voice detection. *IEEE Access* 10, 108046–108066. doi: 10.1109/ACCESS.2022.3211069
- Pearson, J., and Zinets, N. 2022. Deepfake footage purports to show Ukrainian president capitulating. Available online at: <https://www.reuters.com/world/europe/deepfake-footage-purports-show-ukrainian-president-capitulating-2022-03-16/>
- Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images: ICCV, 1–11.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization: ICCV, 618–626.

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2020). FakeNewsNet: a data repository with news content, social context and spatiotemporal information for fake news research. *Big Data* 8, 171–188. doi: 10.1089/big.2020.0062
- Siarohin, A., Sangineto, E., Lathuiliere, S., and Sebe, N. (2019). “First order motion model for image animation” in *Advances in neural information processing systems*, vol. 32, 7137–7147.
- Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cogn. Sci.* 12, 257–285. doi: 10.1207/s15516709cog1202\_4
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). “Attention is all you need” in *Advances in neural information processing systems*, vol. 30.
- Verdoliva, L., et al. (2019). Media forensics and deepfake detection: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14, 910–932.
- Wang, Z., Cheng, Z., Xiong, J., Xu, X., Li, T., Veeravalli, B., et al. (2024). A timely survey on vision transformer for deepfake detection. *arXiv*.
- Yermakov, A., Cech, J., and Matas, J. (2025). Unlocking the hidden potential of CLIP in generalizable deepfake detection. *arXiv*.
- Zhou, X., Zafarani, R., Shu, K., and Liu, H. (2019). Fake news: Fundamental theories, detection strategies, and challenges *ACM Transactions on Information Systems*, 39, 1–40.
- Zhang, L., Li, H., and Wang, J. (2024). Vision transformers in deepfake detection: accuracy, generalization, and practical benchmarks. *IEEE Trans. Inf. Forensics Secur.* 19, 4321–4338.
- Zhou, X., Zafarani, R., Shu, K., and Liu, H. (2021). Fake news: fundamental theories, detection strategies, and challenges. *ACM Trans. Inf. Syst.* 53, 1–40. doi: 10.1145/3395046