Deepfakes at the Media–Law Boundary: A Multilingual Thematic Analysis of Platformed News Discourse

Anonymous Author(s)

Affiliation omitted for review
email@example.com

Abstract. Deepfakes are increasingly discussed as a social computing risk because they do not merely falsify images, voices, or videos. They alter the conditions under which platformed publics, news organizations, courts, and regulators decide what can be trusted. This paper examines how news discourse frames deepfakes at the boundary between media visibility and legal proof. Drawing on 132 Google News items published between 2017 and 2025 in English, Traditional Chinese, and Simplified Chinese, comprising 293,746 words, we conduct a multilingual thematic analysis supported by NVivo 14.0. The analysis identifies three connected domains: normative governance, information ecology and public cognition, and regulatory-technical response. The findings show that sampled reporting treats deepfakes primarily as a governance and liability problem, while also foregrounding public trust, viral amplification, evidentiary uncertainty, and uneven platform responsibility. We argue that deepfakes create recursive feedback loops between media salience, public cognition, legal authenticity, and regulatory intervention. The paper contributes to social computing by explaining how synthetic media risks travel across communication systems and institutional verification settings, and by proposing a coordination model linking provenance, graded uncertainty communication, platform response, and evidentiary handling.

Keywords: Deepfake · Social computing · Thematic analysis · Information ecology · Provenance · Platform governance

1 Introduction

The term deepfake emerged in 2017 and was quickly associated with the wider availability of tools for face swapping, voice imitation, and synthetic audiovisual generation [8,28,65]. Earlier manipulation techniques required specialized expertise and relatively slow production pipelines. Contemporary generative systems reduce those barriers. They can reproduce identity cues, emotional signals, institutional voices, and documentary styles. As a result, the harm of deepfakes extends beyond whether a particular file is false. The more serious question is how synthetic media reshapes the social and institutional conditions under which truth claims are accepted, doubted, contested, or ignored. This concern is

especially visible in platformed news discourse. News reports about deepfakes often move across several registers at once: celebrity victimization, non-consensual sexual imagery, electoral manipulation, corporate fraud, evidentiary disruption, platform accountability, and technical detection. These registers do not operate separately. A high-profile incident can trigger public panic, stimulate legislative proposals, pressure platforms to remove content, and create new expectations for courts or investigators. Conversely, uncertainty about evidence and attribution can feed back into public distrust, enabling the so-called liar's dividend, in which authentic material is dismissed as fake [8,56,53].

Existing research has explained deepfakes through several valuable lenses. Legal scholarship examines privacy, defamation, image-based abuse, personality rights, election law, evidentiary standards, and platform liability [15,30,34,51]. Communication and misinformation studies emphasize public trust, virality, motivated reasoning, echo chambers, and corrective information [1,41,59,64]. Technical work develops detectors, watermarking, biometric analysis, and provenance systems [6,23,38,43,49]. Yet the interaction among these domains remains under-specified. A detector may matter only when a platform preserves metadata, a newsroom communicates uncertainty, or a court treats provenance as relevant to authenticity. Similarly, a statute may fail to stabilize trust when it arrives after public belief has already been shaped by viral circulation.

This paper studies deepfakes as a boundary problem in social computing. It asks how news discourse represents the movement of synthetic media across communication, cognition, law, and governance. The analysis is based on a corpus of 132 Google News items published between 2017 and 2025 in English, Traditional Chinese, and Simplified Chinese. The corpus contains 293,746 words. It was coded through thematic qualitative text analysis using NVivo 14.0. The study does not claim to measure the real-world frequency of deepfake incidents or the actual behavior of courts, newsrooms, or platforms. Instead, it examines mediated salience: the recurring patterns through which deepfakes are made intelligible as public, institutional, and regulatory risks. Three research questions guide the analysis:

- RQ1: How are changes in newsroom capture, verification, and storytelling represented in the sampled news discourse on deepfakes?
- RQ2: How are concerns about legal tests of authenticity, admissibility, and attribution framed in the sampled reporting?
- RQ3: How are gaps in technology-neutral regulation and responsibility articulated across jurisdictions in the corpus?

The paper makes two contributions. Empirically, it maps how multilingual news discourse links verification, evidentiary uncertainty, public cognition, and governance responsibility. The coding results identify three high-level domains. Normative governance accounts for 229 of 488 coded nodes, or 46.9%. Information ecology and public cognition account for 165 nodes, or 33.8%. Regulatory and technical response accounts for 94 nodes, or 19.3%. Theoretically, the paper develops a coordination model for deepfake governance. It argues that social

computing systems need provenance that travels across newsrooms, platforms, and legal settings; uncertainty communication that does not collapse into simple true-false labels; and platform procedures that preserve evidence while limiting harmful circulation.

2 Related Work

2.1 Synthetic Media and Platformed Publics

False content often spreads faster and more widely than accurate information because novelty, anger, fear, and surprise increase sharing incentives [64]. Platform recommendation systems can amplify these incentives by rewarding engagement and clustering users into environments where existing beliefs are reinforced [59]. In such settings, deepfakes can operate as a social weapon in media manipulation, targeting groups, borrowing institutional symbols, and exploiting platform affordances to create confusion [41]. The problem is intensified by ordinary cognitive limits. Users rely on heuristics, source cues, visual fluency, and the reactions of others when deciding what to believe [3,20]. Such conditions enable the liar's dividend, allowing political actors or other speakers to dismiss authentic records as fabricated once realistic fabrication becomes publicly imaginable [8]. The result is not simply deception but reality apathy, a condition in which audiences become uncertain whether verification is possible or worth pursuing.

2.2 Verification, Evidence, and Institutional Trust

News organizations operate under time pressure and must decide whether information is newsworthy, sufficiently verified, and fairly contextualized. Courts and investigators operate under evidentiary norms of authentication, chain of custody, admissibility, probative value, and prejudice. Deepfakes make the gap between journalistic verification and legal proof more visible. A video can be treated as socially consequential before it is legally reliable, while legal uncertainty can later be reported as institutional failure [35,5]. Technical detection is useful, but detectors can be unreliable, biased, or outpaced by generative improvements [43]. Once a detectable artifact becomes known, adversarial creators can adjust their tools. This dynamic encourages a shift from after-the-fact detection toward source authentication and provenance-preserving workflows [39].

In addition, regulatory responses to deepfakes remain fragmented across jurisdictions. Most legal systems still adapt existing doctrines, including privacy, defamation, data protection, intellectual property, criminal law, and platform governance, instead of adopting a comprehensive deepfake statute [65,8]. The United States illustrates a largely decentralized model. Remedies often depend on state-level deepfake rules, tort claims, privacy doctrines, publicity rights, platform policies, and limited federal interventions in specific fields such as nonconsensual intimate imagery [7,27,60]. The European Union GDPR provides a data protection architecture for identifiable personal data, while the AI Act

places deepfake disclosure within a broader framework of risk management and transparency obligations [18,19]. The 2022 China provisions on deep synthesis services require labeling, traceability, and platform-side responsibility, while civil and criminal law provide additional routes for personality rights, personal information, obscene material, online insult, and data misuse claims [11,46,47]. However, the relationship between these obligations and courtroom evidential standards remains under-specified.

2.3 Technology Neutrality Beyond Detection

Rules framed around effects can remain usable when model architectures, interfaces, and distribution channels change, while leaving room for lawful uses such as satire, dubbing, accessibility, artistic reconstruction, and privacy-preserving identity masking [32,50,42]. Current technical methods include reverse engineering generative models, digital fingerprints, visual heartbeat rhythms, optical flow, convolutional traces, and facial warping artifacts [2,48,6,23,38]. Commercial detection services have also expanded [61]. Early work treated abnormal blinking as a useful marker, but creators later adapted, showing how each exposed artifact may train the next generation of manipulation [37,54]. This arms race strengthens the case for upstream provenance, including cryptographic signatures or content credentials at the point of capture, although such systems remain unevenly adopted and operationally constrained [39]. Where upstream authentication is unavailable, courts may need more demanding expert evaluation of digital authenticity [12].

The distinction between misinformation and disinformation matters since false or deceptive media may circulate through attention markets before legal institutions can test provenance [58,24]. Platform monetization and selective amplification can distort public understanding, influence pre-trial publicity, and weaken confidence in legal institutions [63]. Communication research adds a further complication. The third-person effect may increase support for censorship or platform restriction, while hostile-media perceptions and motivated reasoning shape whether audiences label content as false, biased, or dangerous [13,52,10,26,20,33]. Regulatory intervention may therefore protect trust, but it may also be used to delegitimize journalism or lawful speech [55,36].

3 Data and Method

3.1 Corpus Construction

The study used Google News as a retrieval interface for news items concerning deepfakes. Google News was selected because it indexes a broad range of online news sources and shapes public visibility through ranking, language coverage, and recency signals. Searches covered 1 January 2017 to 30 September 2025. The working terms included deepfake, AI, fake, synthesis, AI evidence, and AI police, together with related legal, media, and regulatory terms. The legal domain

included civil liability, tort, criminal law, privacy, finance, policing, courts, evidence, and sexual content. The media domain included public opinion, democracy, public sphere, disinformation, and trust. The regulatory domain included technology, watermarking, prevention, monitoring, platforms, and compliance. Searches were conducted in English, Traditional Chinese, and Simplified Chinese. The multilingual design was not used to make deterministic claims about linguistic culture. It was used to improve coverage of different reporting environments and legal contexts, including transnational English-language reporting, mainland Chinese sources, Hong Kong sources, and the wider Sinophone public sphere. Items were screened using relevance, validity, and completeness. Items whose titles or abstracts did not contain deepfake-related terms, law-related terms, media-related terms, or recognized synonyms were excluded. A single article could contribute to more than one analytic domain. The final corpus contained 132 news items and 293,746 words. Only a small number of items appeared in 2017 and 2018. Coverage increased after 2022 and became heavily concentrated in 2024 and 2025, with 106 of the 132 items appearing in those two years. This temporal concentration was summarized descriptively as:

$$T_{2024\text{–}2025} = \frac{m_{2024} + m_{2025}}{M} = \frac{25 + 81}{132} = 0.803 \tag{1}$$

where m_y denotes the number of sampled items in year y and M denotes the total number of news items. This value supports treating 2024–2025 as a period of discursive acceleration rather than as one stratum in a balanced longitudinal sample. Appendix A reports the descriptive structure of the corpus. Appendix Table 1 summarizes report motivation, news type, reporting period, and geographic distribution. Figures 2 and 3 further visualize the country distribution and source geography of the sampled materials.

3.2 Coding Procedure

The study applied thematic qualitative text analysis. The goal was to identify recurring patterns in how deepfakes were framed as media, legal, technical, and governance risks. The procedure followed a staged approach: corpus familiarization, keyword-supported search, open coding, category consolidation, theme development, and interpretive synthesis [4,17,44,45]. Figure 1 summarizes this workflow from corpus construction to coding, theme aggregation, and interpretive synthesis. NVivo 14.0 was used for corpus management, coding, memo writing, word-frequency queries, and visualization. The word-frequency query in Fig. 5 provides a descriptive overview of recurring terms and supports the subsequent interpretation of thematic salience. The first coding cycle identified short text segments that referred to incidents, risks, institutions, harms, or proposed responses. These segments were organized into primary categories. Coding was not intended to count incidents. A single article could activate multiple codes. For example, a report about a fraudulent deepfake video conference could be coded under corporate fraud, platformed circulation, evidentiary uncertainty, and

institutional trust. This node-based approach captures the density of discursive associations rather than the number of unique real-world events.

To make the aggregation procedure explicit, each first-level category was treated as a coding unit with an observed node count. Let $A = \{a_1, \ldots, a_{38}\}$ denote the set of first-level categories, and let n_i denote the number of coded nodes assigned to category a_i. Each core domain $d \in D$ consists of a subset $A_d \subset A$. The domain count N_d and relative salience score S_d are defined as follows:

$$N_d = \sum_{a_i \in A_d} n_i, \qquad S_d = \frac{N_d}{\sum_{d \in D} N_d} \tag{2}$$

The resulting scores are descriptive measures of coding salience. They do not estimate the real-world frequency of deepfake incidents. This formalization is intended to improve transparency, not to turn thematic coding into incident measurement. The unit of analysis is the coded representation of a concern, not the underlying event. A single article may therefore contribute to multiple categories, and a highly elaborated report may generate several nodes. The resulting counts are best read as measures of discursive salience within the sampled corpus.

Open coding produced 488 coded nodes and 38 first-level categories. Appendix Table 2 reports the full list of first-level coding categories and their frequencies. These were consolidated into 13 second-level themes and then into three core domains: normative governance, information ecology and public cognition, and regulatory-technical response. To strengthen rigor, the analysis used iterative memoing, repeated reading of ambiguous items, negative-case checking, and explicit comparison across years and languages. Reflexive notes were recorded where reports combined legal analysis with sensational framing or where coding decisions risked importing assumptions about actual institutional practice.

4 Findings

The coding results show that deepfakes are framed primarily as a governance and liability problem, while public cognition and technical response remain closely connected. Applying Eq. 2, 229 of 488 coded nodes fell under normative governance frameworks, producing a relative salience score of 46.9%. Information ecology and public cognition accounted for 165 nodes, or 33.8%. Regulatory response and technical regulation accounted for 94 nodes, or 19.3%. Appendix Table 3 reports the aggregation from second-level categories to the three core domains. Figure 4 visualizes the relative frequency of these themes, while Figs. 6 and 7 show the same structure through hierarchy and sunburst views.

The node distribution first shows the dominance of normative governance. Criminal conviction and punishment formed the largest second-level category, with 79 nodes, followed by legislative and revision processes with 70 nodes. Civil remedies and tort determination accounted for 36 nodes, cross-border law enforcement and judicial cooperation for 25 nodes, and evidence crisis for 19 nodes. Overall, these patterns suggest that the sampled reporting often interprets deepfakes through legal boundary setting, sanction, institutional responsibility,

and proof. Figure 9 provides the corresponding descriptive view of the normative governance framework.

Information ecology and public cognition formed the second major cluster. In this domain, political manipulation and democratic intervention accounted for 56 nodes, followed by public trust and panic with 43 nodes, victim narratives and public opinion pressure with 38 nodes, and dissemination and amplification of false information with 28 nodes. These categories show that deepfake coverage is not confined to legal doctrine. Instead, it connects synthetic media to publics, emotion, social division, credibility, and the perceived fragility of shared truth. Figure 8 illustrates this cluster by showing how public cognition, political manipulation, victim narratives, and false-information amplification are distributed within the coded material. By comparison, regulatory and technical response appeared less frequently but remained analytically important. Technology detection and prevention tools accounted for 43 nodes, public education and the ability to recognize falsehoods for 31 nodes, standards, ethics, and industry self-regulation for 13 nodes, and platform responsibility and content review for 7 nodes. Although smaller in scale, this domain indicates that technical fixes are visible in the discourse. However, they are often treated as insufficient unless linked to education, platform governance, and institutional standards.

4.1 RQ1: Newsroom Capture, Verification, and Storytelling

The sampled reporting represents deepfakes as a challenge to the ordinary pathway from capture to verification to storytelling. Capture is increasingly framed as a provenance problem, not simply a speed problem. Newsworthy material often arrives through platforms, private messages, reposted clips, screenshots, or synthetic assets designed to resemble institutional communication. The question is therefore not only whether a newsroom can obtain a file quickly, but whether it can reconstruct where the file came from, how it moved, and whether the surrounding metadata or source claims are reliable. Reports on the Hong Kong deepfake scam, for example, describe how fraudsters deceived a finance clerk through "deepfake representations of the company's CFO and other managers during a video conference," illustrating how synthetic media can now simulate institutional authority at the point of intake.

Verification also moves away from a simple true-or-false judgment toward a graduated assessment of confidence. One report noted that "false information can spread quickly and become 'true' before efforts can be made to debunk it." Other reports highlighted visible artifacts, including cases where "words and lips did not sync" or where a synthetic voice sounded odd and unnatural. Yet these cues remain unstable. A defect that exposes one generation of deepfakes may guide the next generation of improvement. The corpus therefore presents verification as an ongoing confidence assessment involving source history, corroboration, technical review, institutional context, and disclosure of uncertainty.

Storytelling becomes part of the risk. Deepfake incidents are often narrated through emotionally charged frames, including celebrity humiliation, election manipulation, corporate betrayal, school discipline, national security, and sexual

victimization. These frames can mobilize attention and policy response, but they may also narrow public understanding of the technology. The Korean Equiloom case is illustrative. An AI-generated anchor reportedly presented the scheme as a government-certified investment vehicle, while a fabricated presidential endorsement framed it as a life-changing opportunity. Such examples show how narrative packaging, emotional cues, and borrowed institutional voices can redirect attention away from verification and toward urgency, trust, and persuasion.

Three mechanisms recur across the corpus. The first is borrowed authority, where synthetic media simulates the face, voice, or format of trusted executives, officials, anchors, or institutional notices. The second is urgency, where synthetic content is framed as requiring immediate reaction before verification is complete. The third is identity anchoring, where coverage centers recognizable victims or public figures, increasing visibility while obscuring harms to ordinary individuals. These mechanisms show that deepfake risk is not reducible to deception at the file level. It is produced through the interaction between synthetic identity cues and platformed narrative circulation. Corrective reporting should therefore move beyond simple debunking and use confidence bands, traffic-light labels, short explanatory rationales, prebunking devices, and contextual boxes to reduce truth decay, repeated-exposure effects, and identity-protective resistance [57,66,40].

4.2 RQ2: Authenticity, Admissibility, and Attribution

Across the corpus, deepfakes are framed as a source of evidentiary anxiety, with authenticity, admissibility, and attribution as the three recurring concerns. Authenticity is questioned when a file may not be what it purports to be. Admissibility becomes difficult when foundation, prejudice, and probative value are contested. Attribution becomes obscure when creation, distribution, or intent is hidden across tools, users, accounts, and jurisdictions. These concerns are repeatedly connected to concrete institutional consequences. Synthetic content may trigger school discipline, employment action, police investigation, election controversy, or financial loss before its reliability is fully tested. In such cases, deepfakes operate not only as false content but also as quasi-evidence, prompting decisions in social and organizational settings that resemble legal judgment without the safeguards of legal procedure. This explains why evidence-related discourse appears even in reports that are not about formal trials. More broadly, the reporting suggests a movement from recognition-based trust toward process-based authentication. Older assumptions about audiovisual media often relied on familiar appearance, recognizable voice, and plausible setting, but deepfake risk weakens that logic. A more sustainable approach asks who captured the file, what system created it, how it was stored, whether metadata survived, whether alteration occurred, and whether independent corroboration exists. This shift aligns with broader calls for digital chain-of-custody practices and provenance-aware verification [5,21,39].

4.3 RQ3: Regulatory and Technical Gaps

The sampled reporting presents current responses as fragmented, reactive, and uneven. Technical inspection and prevention tools occupy the largest share of this domain, especially active defense and passive detection. These conditions may disappear once content is reposted, compressed, edited, or stripped of source information. Exposed weaknesses may also help adversaries refine later models [39]. Broad public release of detectors can invite reverse engineering and evasion, while restricted access raises concerns about transparency, accountability, and due process when detection results are used in moderation or evidence handling [25]. Appendix Figure 10 summarizes the regulatory and technical response domain, including detection tools, public education, platform responsibility, and standards or self-regulatory initiatives.

Furthermore, platform responsibility appears as a smaller category in the coding results, but it is analytically central. Platforms shape visibility, moderation timing, complaint channels, evidence preservation, and the persistence of harmful content. The corpus suggests that platform rules are often incident-driven and unevenly enforced. One prominent example concerned the circulation of AI-generated intimate images of Taylor Swift, where official responses emphasized that social media companies should enforce their own rules more strictly against misinformation and non-consensual intimate imagery [62]. Such cases show how high-profile celebrity harms may trigger rapid attention, while ordinary victims, financial scams, or localized harms may depend on slower complaint processes. Inconsistent enforcement weakens deterrence and allows manipulated media to persist under narrow or delayed internal rules [9].

Public education forms a second recurring response, but the corpus treats it as incomplete. Reports frequently call for media literacy, public awareness, and user caution. A further cluster concerns the absence of coherent technical-ethics standards and the weakness of ad hoc self-regulatory initiatives. Some jurisdictions emphasize disclosure, watermarking, and traceability. Others rely on privacy, personality rights, defamation, election law, criminal law, or platform terms of service. Each approach addresses part of the problem, but cross-border circulation exposes the gaps. A synthetic video can originate in one country, be hosted by a platform incorporated elsewhere, target a victim in a third jurisdiction, and influence publics across many more. The corpus therefore frames deepfake governance as a coordination problem, not simply as a problem of one missing rule.

5 Discussion: Toward a Coordination Model

The discussion moves from corpus-level patterns to a coordination model for deepfake governance. This model is not presented as a direct empirical finding. It is a synthesis built from recurring concerns in the sampled reporting: verification delay, evidential uncertainty, platform amplification, fragmented responsibility, and public trust. Appendix Figure 11 presents the integrated media–law–regulation framework that links these concerns across institutional domains.

5.1 From Content Moderation to Provenance Infrastructure

The findings suggest that deepfake governance should move beyond content-centered response. Labeling, removal, downranking, and evidential exclusion remain necessary, but they often occur after circulation has already shaped belief, reputation, or institutional reaction. A provenance-centered approach asks a different question: whether a record has a traceable history that can travel across capture, publication, platform circulation, complaint handling, investigation, and adjudication. For news organizations, this means documenting how synthetic-risk material was obtained, what checks were performed, what confidence level was assigned, and what uncertainty remains. For platforms, it means preserving content credentials where available, retaining audit logs for high-risk incidents, and creating review pathways that do not destroy evidence needed by victims, journalists, investigators, or courts. For courts and investigators, it means requiring clearer disclosure of source, handling history, alteration risk, and machine involvement before contested synthetic media is treated as probative. These duties do not require one vendor or one technical architecture. They require interoperable documentation practices. This approach also clarifies the role of technology neutrality. A technology-neutral rule should define duties by function, not by the name of a model or tool. Relevant functions include generation, distribution, amplification, monetization, reporting, preservation, and adjudication. Such an approach remains flexible while avoiding the false comfort that all actors are merely neutral intermediaries. The practical question is what role an actor played in making synthetic content credible, visible, profitable, persistent, or legally consequential. Watermarking and content credentials may support this infrastructure, but they cannot carry the whole burden because adversarial removal, compression, and cross-platform transfer can weaken technical signals [29,43].

5.2 Communicating Uncertainty Without Collapsing Trust

A binary true–false label is often too crude. A file may be unverified, partially altered, contextually misleading, synthetically generated, authentically captured but deceptively edited, or genuine but strategically accused of being fake. Users therefore need signals that communicate confidence without producing panic, complacency, or generalized distrust. Graded labels offer one practical route. Newsrooms could distinguish verified, provenance incomplete, manipulation suspected, synthetic confirmed, and context disputed. Platforms could pair warnings with short explanations, reporting tools, and appeal routes. Courts could treat provenance as part of authenticity and probative value, not only as an admissibility switch. These measures should be designed for moments of rapid circulation. Long warnings are unlikely to help when users share quickly. Short rationales, traffic-light systems, source-history indicators, and prebunking prompts are more likely to support calibrated judgment when they are backed by consistent institutional practice [57,66].

Uncertainty communication also has limits. Warnings may be rejected when audiences view them as partisan, censorial, or platform-serving. Repeated exposure can increase familiarity even after correction, and overuse of warnings

may teach users that all audiovisual evidence is equally unreliable [22]. The goal should therefore be calibrated trust, not blanket skepticism. Public literacy should teach users when to pause, what cues to inspect, where to report, and how to interpret probabilistic claims. Literacy is not only an educational add-on. It is part of the governance system, especially when paired with visible provenance cues, transparent moderation outcomes, and institutional disclosure practices [16,34].

5.3 Closing Feedback Loops Across Institutions

The corpus indicates that deepfake risk moves through feedback loops. Information ecology generates pressure for legal action. Legal uncertainty shapes media reporting and public trust. Platform responses affect what evidence remains available. Regulatory measures reshape incentives for creators, distributors, and news organizations. These loops can stabilize trust when they are coordinated, but they can intensify distrust when each institution acts alone.

A coordination model should include four layers. The first is capture and creation, where content credentials, watermarking, and source records should be preserved when feasible. The second is circulation, where platforms should provide labeling, rapid review for high-risk synthetic content, evidence preservation, and due-process safeguards [31]. The third is institutional verification, where newsrooms, investigators, and courts should use provenance-aware procedures and communicate uncertainty in compatible terms. The fourth is public literacy, where users learn not only that deepfakes exist, but how verification, provenance, and confidence signals work. Appendix Figure 12 translates these layers into a technology-neutral coordination model for deepfake risk control. In particular, this model does not assume that all deepfakes are harmful. Synthetic media can support entertainment, accessibility, dubbing, parody, artistic reconstruction, privacy protection, and legitimate experimentation. The governance task is to preserve lawful expression while constraining deceptive impersonation, non-consensual sexual imagery, electoral manipulation, financial fraud, and fabricated evidence [8,14]. A function-based approach is better suited to that balance than a tool-specific ban.

5.4 Implications for Social Computing Research

These findings suggest three implications for social computing research. First, deepfake governance should be studied as a socio-technical process. A model that synthesizes a face or voice becomes consequential only when it enters systems of sharing, ranking, verification, moderation, and adjudication. Future work should examine whether repost chains preserve source cues, whether warning labels appear before sharing, whether complaint channels collect usable evidence, and whether provenance marks survive compression or cross-platform transfer.

Second, evaluation should move beyond file-level detection accuracy. Detection benchmarks remain useful, but the broader question is whether an intervention reduces harmful belief formation, preserves evidence, and avoids chilling legitimate

expression. A detector may work in the lab but fail in practice if users distrust labels, platforms strip metadata, or courts cannot inspect the system. Evaluation should therefore combine technical performance with behavioral and institutional measures, including sharing intent, trust calibration, time to correction, complaint resolution, and evidentiary preservation.

Third, multilingual and cross-jurisdictional corpora should be treated as core evidence. Deepfake incidents rarely remain within one language or legal system. A political fabrication, intimate-image abuse case, or financial scam may appear across transnational media, local reporting, platform posts, and official notices. Each setting may foreground different victims, authorities, remedies, and uncertainties. Comparative social computing research can identify which risks are globally shared and which arise from local media systems, regulatory structures, and platform ecosystems.

These implications also refine technology-neutral governance. Neutrality should not mean indifference to design. Duties should be specified by function and effect, while remaining attentive to the technical and social architectures that make harm more likely. For social computing, the contribution is not a complete legal solution to deepfakes. It is an account of how synthetic media risk travels through coupled socio-technical systems, and how provenance, verification, platform governance, and public literacy can be aligned.

6 Limitations and Conclusion

This study has several limitations. Google News is a visibility interface rather than a complete archive of deepfake incidents. The corpus likely over-represents scandal, celebrity harm, political controversy, and legally dramatic cases, while under-representing benign, creative, routine, or lower-conflict uses of synthetic media. In addition, node counts measure coding density rather than incident frequency, since one news item may activate several codes. Despite these limits, the study shows how deepfakes are constructed as a social computing risk at the media–law boundary. Sampled reporting does not treat deepfakes only as a technical detection problem. It connects synthetic media to governance, public trust, institutional proof, platform responsibility, and regulatory lag. Normative governance was the largest coded domain, followed by information ecology and public cognition, and then regulatory-technical response. This pattern suggests that deepfakes are publicly understood through institutional and social consequences as much as through technical features.

The broader implication is that deepfake governance must be coordinated across systems. Detection tools are useful but incomplete, legal rules are necessary but reactive, media literacy is valuable but insufficient without institutional support, and platform moderation remains uneven unless tied to due process and evidence preservation. A social computing approach should therefore focus on the connective tissue of governance: portable provenance, calibrated uncertainty labels, evidence-preserving platform pathways, and institutional practices that align media truth with legal proof.

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of Economic Perspectives **31**(2), 211–236 (2017)
2. Asnani, V., Yin, X., Hassner, T., Liu, X.: Reverse engineering of generative models: Inferring model hyperparameters from generated images. arXiv preprint arXiv:2106.07873 (2021)
3. Bikhchandani, S., Hirshleifer, D., Welch, I.: A theory of fads, fashion, custom, and cultural change as informational cascades. Journal of Political Economy **100**(5), 992–1026 (1992)
4. Braun, V., Clarke, V.: Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health **11**(4), 589–597 (2019)
5. Breen, D.C.: Silent no more: How deepfakes will force courts to reconsider video admission standards. Journal of High Technology Law **21**, 122–168 (2021)
6. Caldelli, R., Galteri, L., Amerini, I., Del Bimbo, A.: Optical flow based cnn for detection of unlearnt deepfake manipulations. Pattern Recognition Letters **146**, 31–37 (2021)
7. Chabinsky, S., Pittman, F.P.: Usa: Data protection 2019. In: Gabel, D., Hickman, T. (eds.) The International Comparative Legal Guide to Data Protection 2019. Global Legal Group (2019)
8. Chesney, R., Citron, D.K.: Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review **107**, 1753–1820 (2019)
9. Chiu, A.: Facebook wouldn't delete an altered video of nancy pelosi. what about one of mark zuckerberg? The Washington Post (2019)
10. Corbu, N., Oprea, D.A., Negrea-Busuioc, E., Radu, L.: They can't fool me, but they can fool the others! third-person effect and fake news detection. European Journal of Communication **35**(2), 165–180 (2020)
11. Cyberspace Administration of China and others: Provisions on the administration of deep synthesis of internet information services (2022)
12. Daubert v. Merrell Dow Pharmaceuticals, Inc.: 509 u.s. 579 (1993), united States Supreme Court
13. Davison, W.P.: The third-person effect in communication. Public Opinion Quarterly **47**(1), 1–15 (1983)
14. De Ruiter, A.: The distinct wrong of deepfakes. Philosophy & Technology **34**(4), 1311–1332 (2021)
15. Delfino, R.A.: Pornographic deepfakes: The case for federal criminalization of revenge porn's next tragic act. Actual Problems of Economics and Law **14**(1), 105–123 (2020)
16. Durach, F., Bârgăoanu, A., Nastasiu, C.: Tackling disinformation: Eu regulation of the digital space. Romanian Journal of European Affairs **20**(1), 5–20 (2020)
17. Elliott, V.: Thinking about the coding process in qualitative data analysis. The Qualitative Report **23**(11), 2850–2861 (2018)
18. European Union: Regulation (eu) 2016/679 of the european parliament and of the council: General data protection regulation (2016), arts. 4(1), 5(1)(d), and 17
19. European Union: Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (2024), artificial Intelligence Act, Art. 50
20. Flanagin, A.J., Metzger, M.J.: Perceptions of internet information credibility. Journalism and Mass Communication Quarterly **77**(3), 515–540 (2000)
21. Freeman, L.: Weapons of war, tools of justice: Using artificial intelligence to investigate international crimes. Journal of International Criminal Justice **19**(1), 35–53

(2021)
22. Gilbert, D.T.: How mental systems believe. American Psychologist **46**(2), 107–119 (1991)
23. Guarnera, L., Giudice, O., Battiato, S.: Fighting deepfake by exposing the convolutional traces on images. IEEE Access **8**, 165085–165098 (2020)
24. Habgood-Coote, J.: The term fake news is doing great harm. The Conversation (2018)
25. Harwell, D.: Top ai researchers race to detect 'deepfake' videos: 'we are outgunned'. The Washington Post (2019)
26. Jang, S.M., Kim, J.K.: Third person effects of fake news: Fake news regulation and media literacy interventions. Computers in Human Behavior **80**, 295–302 (2018)
27. Kerr, O.S.: Computer Crime Law. Thomson/West (2006)
28. Kietzmann, J., Lee, L.W., McCarthy, I.P., Kietzmann, T.C.: Deepfakes: Trick or treat? Business Horizons **63**(2), 135–146 (2020)
29. Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T.: A watermark for large language models. In: Proceedings of the 40th International Conference on Machine Learning. pp. 17061–17084. PMLR (2023)
30. Kirchengast, T.: Deepfakes and image manipulation: Criminalisation and control. Information and Communications Technology Law **29**(3), 308–323 (2020)
31. Klonick, K.: The new governors: The people, rules, and processes governing online speech. Harvard Law Review **131**, 1598–1670 (2017)
32. Koops, B.J.: Should ict regulation be technology-neutral? In: Starting Points for ICT Regulation: Deconstructing Prevalent Policy One-Liners, pp. 77–108. T.M.C. Asser Press (2006)
33. Kruger, J., Dunning, D.: Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology **77**(6), 1121–1134 (1999)
34. Kugler, M.B., Pace, C.: Deepfake privacy: Attitudes and regulation. Northwestern University Law Review **116**, 611–680 (2021)
35. Langbein, J.H.: Historical foundations of the law of evidence: A view from the ryder sources. Columbia Law Review **96**, 1168–1202 (1996)
36. Lees, C.: Fake news: The global silencer. Index on Censorship **47**(1), 88–91 (2018)
37. Li, Y., Chang, M.C., Lyu, S.: In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. arXiv preprint arXiv:1806.02877 (2018)
38. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv:1811.00656 (2018)
39. Lin, H., Holland, H.: The danger of deepfakes: Responding to bobby chesney and danielle citron. Lawfare (2018)
40. Majerczak, P., Strzelecki, A.: Trust, media credibility, social ties, and the intention to share towards information verification in an age of fake news. Behavioral Sciences **12**(2), 51 (2022). https://doi.org/10.3390/bs12020051
41. Marwick, A., Lewis, R.: Media manipulation and disinformation online. Data and Society Research Institute (2017)
42. Maxwell, W., Bourreau, M.: Technology neutrality in internet, telecoms and data protection regulation. Computer and Telecommunications Law Review (2014)
43. Mirsky, Y., Lee, W.: The creation and detection of deepfakes: A survey. ACM Computing Surveys **54**(1), 1–41 (2021). https://doi.org/10.1145/3425780
44. Naeem, M., Ozuem, W.: Understanding misinformation and rumors that generated panic buying as a social practice during covid-19 pandemic. Information Technology and People **35**(7), 2140–2166 (2022)

45. Naeem, M., Ozuem, W., Howell, K., Ranfagni, S.: A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. International Journal of Qualitative Methods **22** (2023)
46. National People's Congress: Civil code of the people's republic of china (2020), arts. 990 and 1019
47. National People's Congress: Criminal law of the people's republic of china (2023), 2023 Amendment, Arts. 246, 253(1), 293(1), 364, and 365
48. Peckham, O.: Deepfake detection and authentication technologies (2021), online commentary on synthetic media detection and attribution
49. Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., Liu, Y., Zhao, J.: Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4318–4327 (2020)
50. Reed, C.: Taking sides on technology neutrality. SCRIPTed **4**, 263–284 (2007)
51. Reid, S.R.: The deepfake dilemma: Reconciling privacy and first amendment protections. University of Pennsylvania Journal of Constitutional Law **23**, 209–258 (2021)
52. Rojas, H., Shah, D.V., Faber, R.J.: For the good of others: Censorship and the third-person effect. International Journal of Public Opinion Research **8**(2), 163–186 (1996)
53. Schick, N.: Deepfakes: The Coming Infocalypse. Twelve (2020)
54. Scoles, S.: Deepfakes are getting better. but they're still easy to spot. Wired (2018)
55. Siebert, F.S., Peterson, T., Schramm, W.: Four Theories of the Press. University of Illinois Press, Urbana (1956)
56. Silbey, J., Hartzog, W.: The upside of deep fakes. Maryland Law Review **78**, 960–966 (2019)
57. Swire-Thompson, B., DeGutis, J., Lazer, D.: Searching for the backfire effect: Measurement and design considerations. Journal of Applied Research in Memory and Cognition **9**(3), 286–299 (2020). https://doi.org/10.1016/j.jarmac.2020.06.006
58. Tan, D.: The law and media: A continuing tension. Journal of Media Law (2022)
59. Tufekci, Z.: Twitter and Tear Gas: The Power and Fragility of Networked Protest. Yale University Press (2017)
60. United States Congress: Tools to address known exploitation by immobilizing technological deepfakes on websites and networks act (2025), tAKE IT DOWN Act
61. Usmani, T.N., Hussain, M.Z., Hasan, M.Z.: Sentinel ai: Revolutionizing cybersecurity with intelligent intrusion detection (2025), manuscript
62. Vituscka, A., Smith, D.: Taylor swift ai images expose the weakness of platform rules on deepfake abuse. News report (2024)
63. Vojak, B.A.: Fake news: The commoditization of internet speech. California Western Law Review **54**(1) (2017)
64. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
65. Westerlund, M.: The emergence of deepfake technology: A review. Technology Innovation Management Review **9**(11), 39–52 (2019)
66. Wood, T., Porter, E.: The elusive backfire effect: Mass attitudes' steadfast factual adherence. Political Behavior **41**, 135–163 (2019)

Appendix

A Corpus Statistics and Coding Tables

Table 1. Corpus descriptive statistics.

Statistic	Details	Number
Motivation of report	Report the cases; discuss cases	85; 47
Types of news	News article; news analysis; feature article	84; 46; 2
Periods of report	2017; 2018; 2019; 2020; 2021; 2022; 2023; 2024; 2025	1; 2; 5; 3; 2; 2; 6; 25; 81
Locations of report	United States; international; mainland China; Hong Kong; United Kingdom; Australia; South Korea; India; Singapore; Denmark; and other locations	26; 22; 17; 11; 10; 10; 9; 7; 4; 4; remaining locations: 1–2 each

Table 2: First-level coding categories and frequencies.

Code	Category	Frequency
a1	Legislative response to public opinion events	31
a2	Law revision incorporates new technology	18
a3	Legislative lag and legal gap	21
a4	Personality rights and portrait rights infringement	27
a5	Difficulty in proof and accountability	9
a6	Old charges and penalties	9
a7	Establish a new charge	19
a8	Sentencing and deterrence	32
a9	Criminal liability pursuit on platform	19
a10	Jurisdictional conflict	5
a11	Lack of international collaboration mechanism	2
a12	Public opinion trial interferes with judicial independence	5
a13	Regulatory intervention and requirements	13
a14	Traditional verification methods are difficult to identify	4
a15	Current law on deepfake evidence is vague	1
a16	Reduces the credibility of the witness	2
a17	Increases the cost of evidence verification	7
a18	Increases the complexity of pre-trial evidence disclosure	4
a19	Causes jury bias	1
a20	Social media viral spread	9
a21	Pretends to be news	12
a22	Regulatory lag leads to the spread of false information	7
a23	Erosion of trust	26

Code	Category	Frequency
a24	Targeted social panic	13
a25	Debunking rumors proves futile and the backfire effect	4
a26	Celebrity victim effect	27
a27	Public sympathy drives the agenda	11
a28	Electoral interference and disinformation	28
a29	Geopolitical attack tool	9
a30	Creates social division and unrest	19
a31	Passive detection technology	17
a32	Active defense technology	26
a33	Platform policies and rules	4
a34	Disclaimer and liability avoidance	3
a35	Missing technical ethics framework	7
a36	Industry self-regulation initiative	6
a37	Media literacy education	9
a38	Lack of ability to recognize fraud	22

Table 3. Core theme aggregation.

Core domain	Corresponding second-level category	Frequency
Normative governance framework	Legislative and revision process	70
	Civil relief and tort identification	36
	Criminal conviction and punishment	79
	Cross-border law enforcement and judicial cooperation	25
	Evidence crisis	19
Information ecology and public cognition	Dissemination and amplification of false information	28
	Public trust and panic	43
	Victim narratives and public opinion pressure	38
	Political manipulation and democratic intervention	56
Regulatory response and technical regulation	Technology detection and prevention tool	43
	Platform responsibility and content review	7
	Standards, ethics, and industry self-regulation	13
	Public education and the ability to recognize falsehoods	31

B Figures

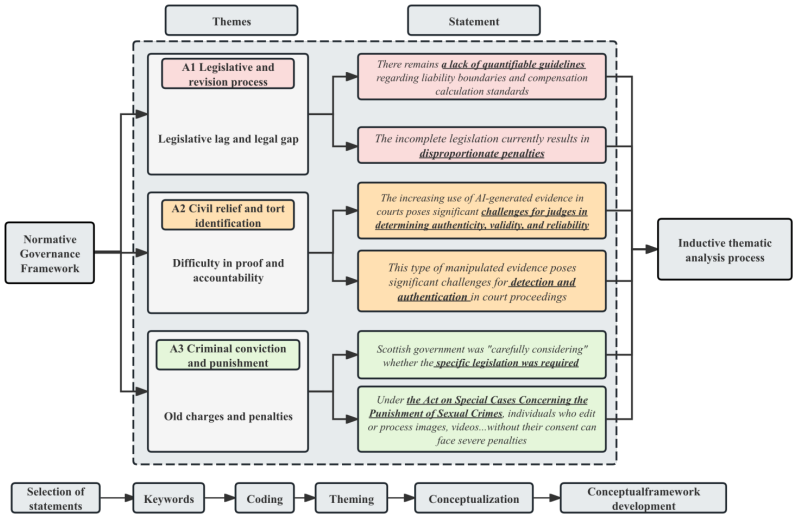

Fig. 1. Thematic analysis workflow for deepfake governance.

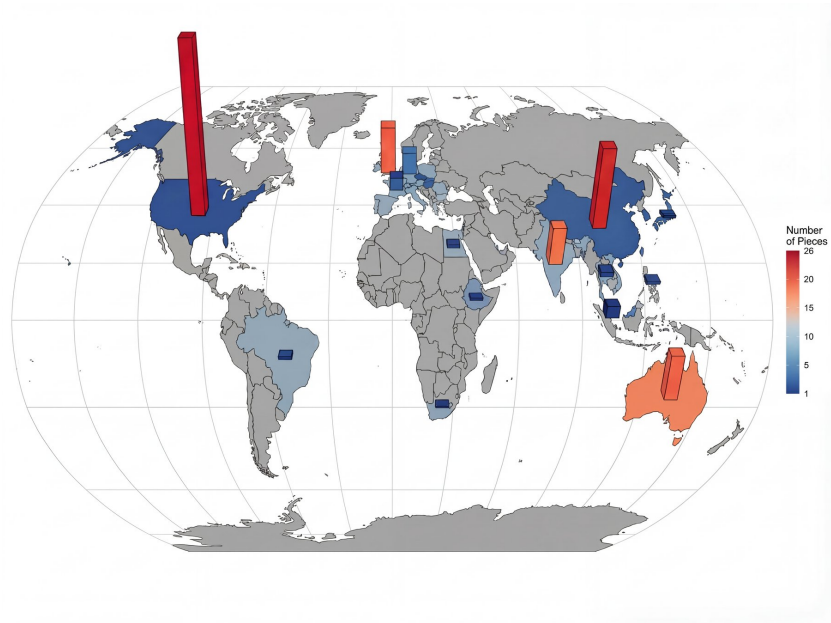

Fig. 2. Country distribution heatmap.

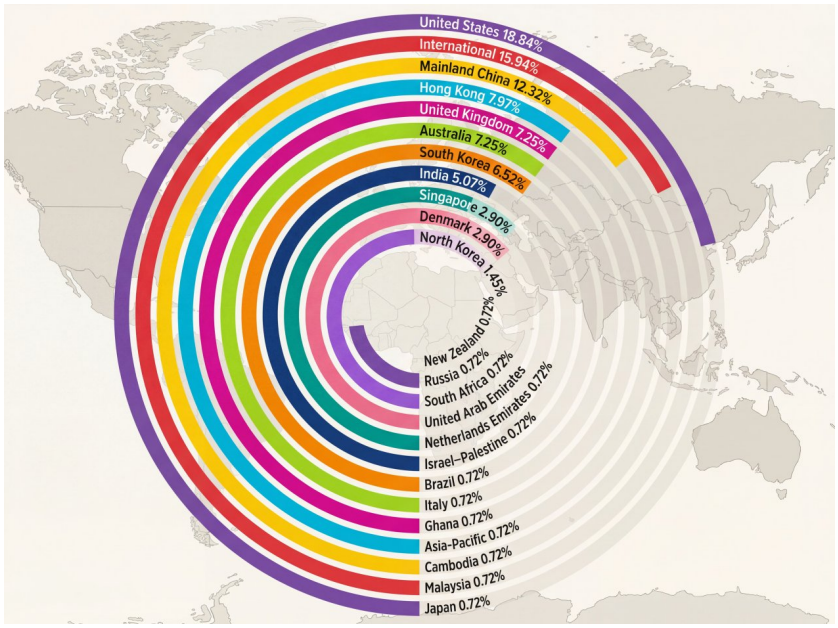

Fig. 3. Source geography of the corpus.

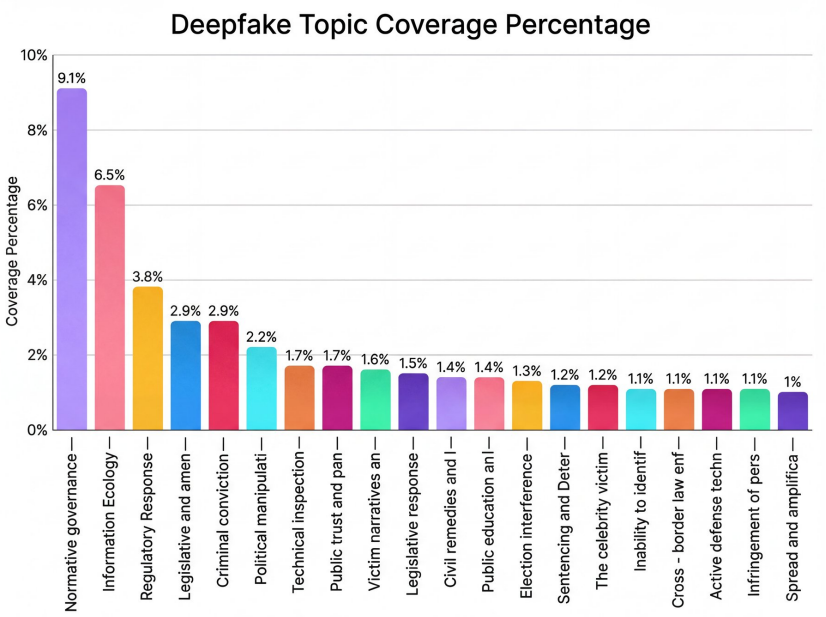

Fig. 4. Theme frequency in the deepfake corpus.

Fig. 5. Word-frequency query for theme analysis.

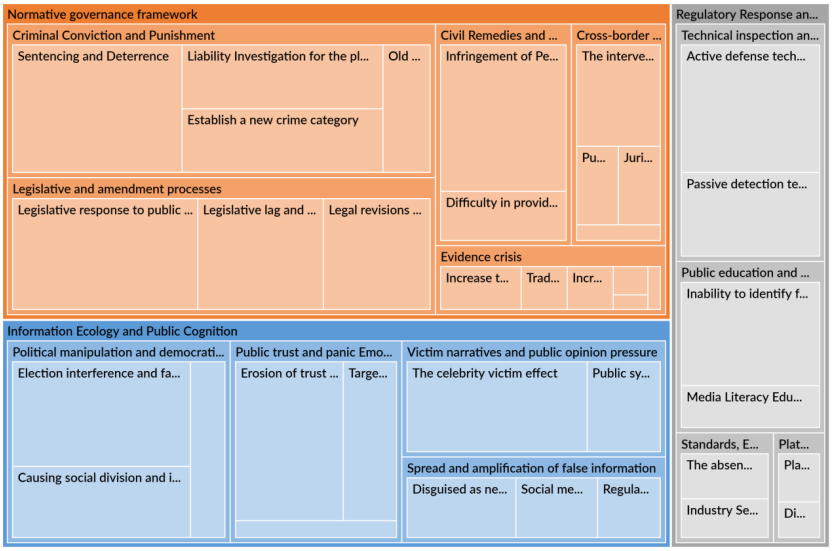

Fig. 6. Hierarchy chart for identified themes.

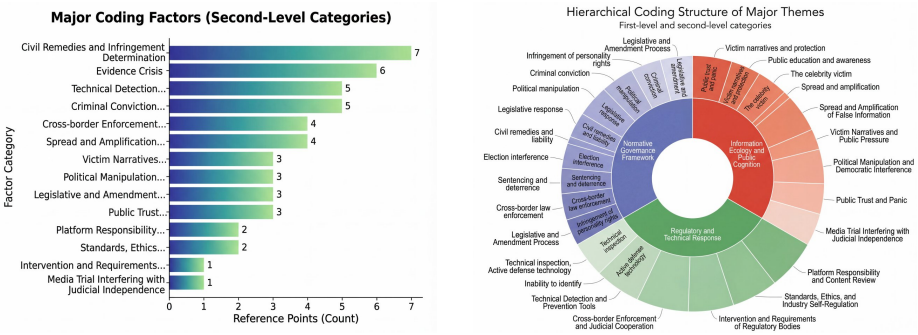

Fig. 7. Sunburst charts showing major factors identified for coding.

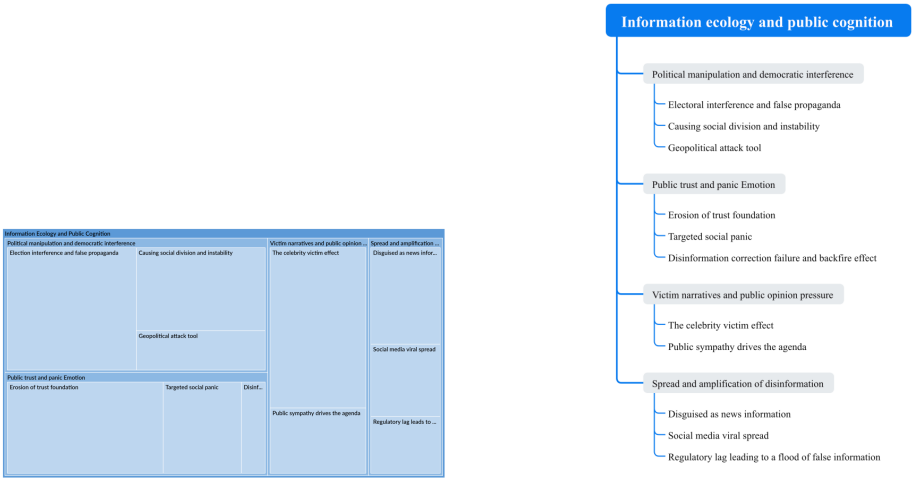

Fig. 8. Descriptive view of information ecology and public cognition.

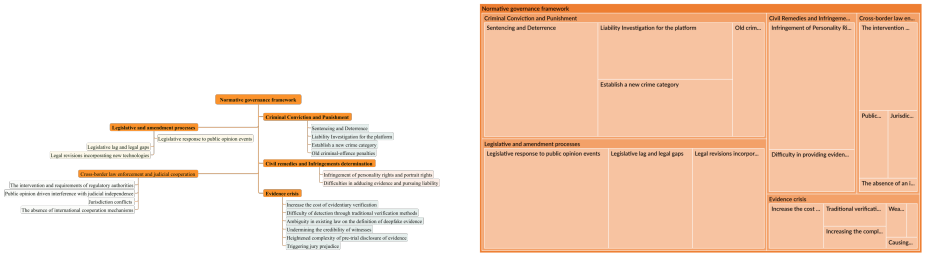

Fig. 9. Descriptive view of normative governance framework.

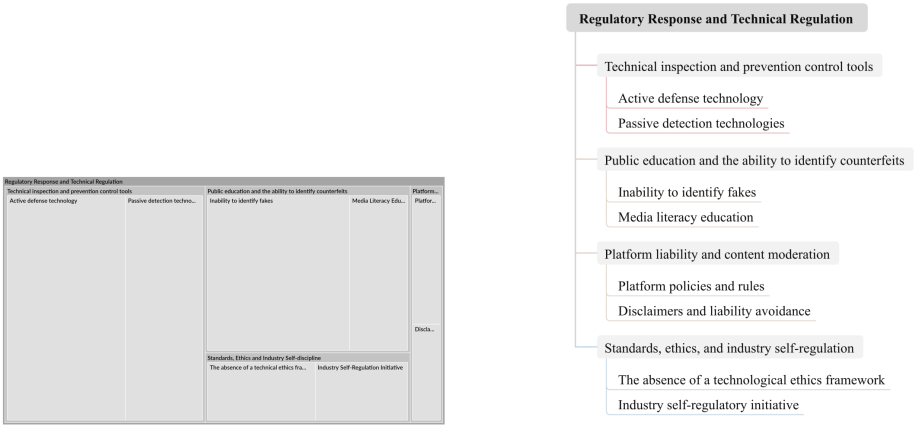

Fig. 10. Descriptive view of regulatory response and technical regulation.

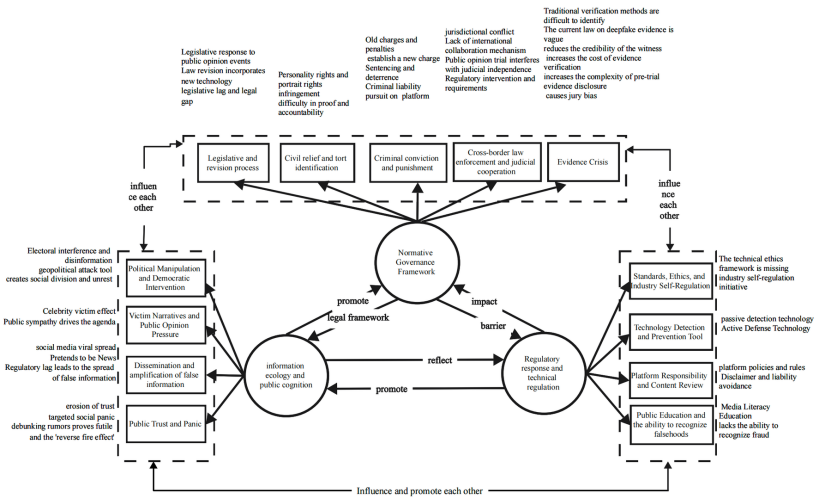

Fig. 11. Integrated media–law–regulation framework for deepfake governance.

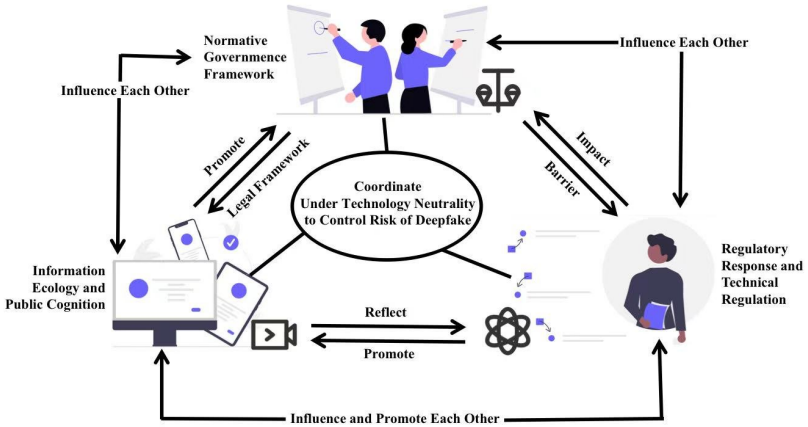

Fig. 12. Technology-neutral coordination model for deepfake risk control.