International Journal of Sustainable Law, 3(1), (2026)

CIVIL AND CRIMINAL LIABILITY OF DEEPFAKE AI PLATFORM PROVIDERS IN VIOLATIONS OF THE RIGHT TO REPUTATION

Milka Kumala[1], Tuti Handayani[2], Muhammad Rizki[3], Reza Dipta Prayitna[4]

[1]Universitas Kristen Cipta Wacana
[2]Department of Law, Universitas Kristen Cipta Wacana, Malang, Indonesia

Article history
Received: 27 April 2026
Revised: 28 May 2026
Accepted: 25 June 2026

Keywords
Deepfake Artificial Intelligence
Right to Reputation
Platform Liability

Abstract
The rapid development of Artificial Intelligence (AI) has introduced deepfake technology, which is capable of manipulating audiovisual data with high realism. While this technology offers creative utility, its misuse for synthesizing non-consensual pornography, political disinformation, and character assassination severely violates an individual's right to reputation. This study examines the civil and criminal liabilities of AI deepfake platform providers under the Indonesian legal framework, specifically Law No. 1/2024 (Second Amendment to the ITE Law) and Law No. 27/2022 on Personal Data Protection (UU PDP). Using a socio-legal research method, this study analyzes how textual legal norms interact with the digital behavior of platform providers and the psychological and social vulnerabilities of victims. The results indicate a significant normative gap: current regulations primarily target the individual creators or distributors of deepfake content, while platform providers escape liability by hiding behind the "safe harbor" doctrine. Sociologically, this legal vacuum fosters a culture of impunity and exacerbates digital gender-based violence. This paper argues for a paradigm shift from passive moderation to active algorithmic accountability. Platform providers must be held civilly liable under Article 1365 of the Indonesian Civil Code for systemic negligence, and criminally liable under the corporate fault doctrine for failing to implement mandatory digital watermarking and proactive content filtration systems.

1. Introduction

Advances in artificial intelligence (AI) technology have ushered humanity into an era of digital disruption unlike anything we have ever seen before. One of the most revolutionary yet destructive AI-derived products today is deepfake a technology based on generative adversarial networks (GANs) capable of fabricating images, videos, and audio with a level of similarity that is nearly indistinguishable from the original. In the social and societal context of Indonesia, the presence of deepfake service platforms both free and subscription-based has democratized digital manipulation tools. Anyone, without needing programming skills, can now alter a person's face and voice in a video in just a matter of seconds.

However, despite its creative utility, deepfake technology has evolved into a highly dangerous cyberweapon that threatens humanity's very existence, particularly through violations of the right to reputation. Reputation is not merely an abstract concept; it is an extension of human dignity, which is constitutionally protected under Article 28G (1) of the 1945 Constitution of the Republic of Indonesia. In Indonesia's cyber landscape, the majority of deepfake abuses manifest as non-consensual deepfake pornography, digital defamation, and political disinformation. The primary victims of this phenomenon are predominantly women and public figures, who suffer social consequences ranging from social exclusion and career ruin to deep psychological trauma.

Legally speaking, Indonesia's positive law appears to be struggling to keep pace with this technological trend. Law No. 1 of 2024 on the Second Amendment to the ITE Law (New ITE Law) and Law No. 27 of 2022 on Personal Data Protection (PDP Law) have a fundamental limitation: their law enforcement focus remains heavily individual-centric. These regulations are designed to target individual actors who create or disseminate manipulative content. Conversely, the entities that provide the infrastructure, algorithms, and commercial platforms for creating deepfakes namely, AI platform providers often evade legal consequences. They hide behind the safe harbor doctrine (notice and takedown), claiming that they are merely passive technological intermediaries with no control over users' malicious intent (user-generated content).

*Corresponding author, email: milkakumala@cwcu.ac.id
doi: https://doi.org/10.71131/dj6rqv13
© 2026 The Authors
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

It is this legal gap that underscores the urgency of this research. A purely doctrinal approach is no longer sufficient to dissect this complexity, as it overlooks the economic-political dimensions of platform owners who profit from the traffic generated by manipulative content. Therefore, this study employs a socio-legal approach to explore not only what the legal texts say, but also how the legal structure and digital behavior of platform providers contribute to the persistence of victimization in the digital space. Based on this background, this study aims to answer two crucial questions: First, how are civil and criminal legal liabilities for AI deepfake platform providers constructed within the current Indonesian legal system? Second, what is the ideal platform accountability model to restore and protect citizens' right to reputation?

2. Method

This study employs a socio-legal research method, which integrates legal doctrinal analysis with an empirical-qualitative approach to examine law within its social context. Law is not viewed as a sterile, autonomous entity, but rather as a social phenomenon that interacts dialectically with the development of AI technology and the structure of digital society. Research Design: A descriptive-analytical study with a qualitative approach, exploring the operation of laws related to cyber liability and reputation protection. Research Subjects: Research subjects include regulatory texts (the ITE Law, the PDP Law, the New Criminal Code/Law No. 1 of 2023, and the Civil Code), AI platform privacy policy documents, as well as secondary data in the form of cyber victimization reports from non-governmental organizations. Research Instruments: The researcher serves as the primary instrument, aided by documentation guidelines and a conceptual analysis matrix to map platform liability. Data Collection Procedures: Data is collected through a document study (literature review) of primary, secondary, and tertiary legal materials published over the past ten years (2016–2026), as well as a digital search of content moderation mechanisms on generative AI provider platforms.

3. Result and Discussion

3.1. Limitations of Existing Regulations and the Safe Harbor Doctrine: Legal Loopholes for Platform Providers

Research findings indicate that Indonesia's current positive law suffers from a severe technological gap in addressing the involvement of AI deepfake platform providers. Article 27A of Law No. 1/2024 (the New ITE Law), which governs attacks on honor or reputation in the digital space, explicitly targets "any person who intentionally attacks the honor...". In judicial practice, the phrase "Any Person" is consistently interpreted as referring to an individual material perpetrator (the content creator or the initial disseminator). The same applies to Law No. 27/2022 on Personal Data Protection (PDP Law). Although AI platforms process the victim's facial and voice biometric data (which is categorized as specific personal data under Article 4 of the PDP Law), the platforms consistently argue that such data processing is carried out at the user's instruction, thereby positioning themselves solely as "Personal Data Processors," not "Personal Data Controllers" who bear primary legal responsibility.

This legal helplessness is exacerbated by the incorporation of the Safe Harbor doctrine—derived from international law into Indonesian regulations, such as the Minister of Communication and Information Technology's Circular Letter No. 5 of 2016. Under this doctrine, platform providers are exempt from legal liability for illegal content created by their users, provided they offer a "notice and takedown" feature and promptly remove such content upon receiving a complaint. However, socio-legal analysis shows that this doctrine is no longer relevant and actually creates systemic injustice for victims of deepfakes. Once a pornographic or disinformation deepfake video is uploaded, it spreads exponentially (goes viral) within minutes. Although platforms remove the video after receiving a report, the damage to the victim's reputation and social dignity is already permanent (irreparable harm). This reactive content moderation model shifts the entire burden of monitoring (burden of policing) onto the victim's shoulders, while platform providers continue to reap financial gains from monetizing cyber traffic impressions.

The limitations on legal liability set forth in current national regulations reflect the legal paradigm's failure to understand the architecture of generative artificial intelligence. Safe harbor policies were originally designed for conventional web hosting providers or passive online discussion forum providers that do not intervene in any way with user-uploaded data. However, generative AI platforms actively transform raw input data through complex probabilistic computational processes to generate new outputs. This activity is clearly not merely "passive storage," but rather a process of digital content production

involving substantial modification of algorithmic instructions that are provided and fully controlled by the platform owner. From the perspective of the sociology of cyberlaw, the application of the safe harbor doctrine to the generative AI industry acts as an instrument of impunity that legitimizes the economic exploitation of victims' suffering. When manipulative videos that destroy a citizen's reputation go viral, traffic to the platform surges sharply, which in turn increases the company's valuation and advertising revenue. Platforms thus enjoy direct economic benefits from the circulation of illegal content, while the legal shield of the safe harbor exempts them from the obligation to fund the restoration of the reputation and psychological well-being of victims whose lives have been shattered by the technology they created..

Furthermore, the conventional notice-and-takedown mechanism has proven ineffective against the "hydra effect" in the digital realm, where a single deepfake video removed from one platform reappears in dozens of new copies across various mirror sites. The legal system's inability to ensure this total (permanent) erasure underscores that a reactive, downstream law enforcement approach is no longer sufficient. A radical legal reorientation is needed, shifting the focus of intervention from the post-incident, downstream dissemination phase to the algorithmic prevention phase at the upstream stage of technology production. This disparity is further exacerbated by the absence of mandatory technology audit standards for any artificial intelligence platform seeking to commercialize its services within Indonesian jurisdiction. The government appears to be allowing the domestic digital space to become an unregulated technological testing ground (an "unregulated digital Wild West") that sacrifices individual cybersecurity in the name of fostering the growth of a superficial digital economy. This failure to exercise oversight at the source is the root cause of the massive number of violations of the right to reputation facilitated by artificial intelligence-based software.

As a result, the burden of proof and the complex task of digital tracing now fall entirely on victims who lack the technical expertise. To have a deepfake video taken down, victims are often required to fill out complicated complaint forms, provide proof of identity, and even submit their own digital forensic analysis to prove that the video is a computer-generated forgery. This victim-unfriendly complaint bureaucracy reflects the institutional indifference of cyber technology corporations toward the social trauma experienced by the public. Looking deeper, the effectiveness of personal data protection guarantees under the Personal Data Protection Act is undermined by the standardization of unilateral legal disclaimer clauses (terms of service) by platforms. Through click-wrap agreements, platforms force users to agree that any technological output that infringes on third-party rights is the sole responsibility of the end user. Our civil law views this as an abuse of circumstances (misbruik van omstandigheden) in digital contracts, where consumers or users lack a balanced bargaining position, and platforms exploit this imbalance to wash their hands of the legal risks associated with biometric manipulation.

From a sociological perspective, this regulatory laxity fosters what is known as a "culture of digital impunity" a social condition in which technology developers feel they are above the law because their products are too abstract to be subject to conventional criminal laws. Legal doctrines requiring proof of personal criminal intent (mens rea) on the part of company executives become obsolete when the decision to release manipulative, uncensored videos is made automatically by autonomous AI agents. As a result, the law fails to hold accountable the legal entities that actually control this destructive ecosystem. Indonesia's digital sovereignty is also at stake, as the majority of AI deepfake platforms operate across borders without having an official legal entity within the country. Domestic law enforcement agencies often hit a wall of international jurisdiction when attempting to take action or block servers. The government's lack of firmness in enforcing rules requiring foreign Electronic System Operators (ESOs) to comply with national law renders any means of reputation restoration for Indonesian citizens who fall victim to such content nothing more than a pipe dream. Finally, the blurred line between creative parody and cyber defamation in our current law widens a gray area that benefits tech corporations. Platforms consistently use the shield of creative freedom to reject preventive censorship. Yet, from a human rights perspective, the right to reputation is a civil right that must not be violated for the economic interests of the visually manipulative entertainment industry. It is this failure to establish a clear legal demarcation between legitimate artistic expression and algorithmic crimes that perpetuates impunity for artificial intelligence platform providers.

3.2.The Concept of Civil Liability: A Tort Claim Based on Algorithmic Negligence

In the realm of civil law, providers of AI deepfake platforms can no longer hide behind the status of passive intermediaries. Researchers propose a legal framework based on the concept of tort (PMH) as stipulated in Article 1365 of the Civil Code (KUHPerdata). To hold platform providers liable, the element of "fault" (schuld) must be shifted from intentional fault (dolus) to systemic negligence (culpa/negligence). Platform providers are found to be negligent when they fail to exercise a reasonable standard of care in designing and operating their algorithms. Researchers have identified three layers of systemic negligence on the part of current deepfake AI platform providers, as presented in the following table:

Layers of Negligence	Specific Forms of Negligence	Legal Basis/Theoretical Framework
Phase I: Pre-Production	Does not implement a consent verification mechanism before the algorithm processes third-party facial photos uploaded by users.	The Theory of Algorithmic Accountability
Tier II: Production	We intentionally did not embed permanent digital watermarks into the metadata of the AI-synthesized video.	Article 1365 of the Civil Code (Duty of Care)
Tier III: Post-Production	Does not provide a proactive AI-based filtering system (automated content filtering) to detect and block the reappearance of previously reported deepfake content	The Doctrine of Vicarious Liability

Once this element of negligence is established, victims of defamation may seek compensation for both material and immaterial damages, the amount of which must be calculated in proportion to the scale of the content's dissemination. Furthermore, courts in Indonesia must begin to boldly apply the doctrine of Strict Liability for platforms that commercialize high-risk AI systems, under which platform providers are obligated to pay damages without the victim having to go to great lengths to prove the platform's fault. This tort claim based on algorithmic negligence opens up new horizons in Indonesia's digital civil law jurisprudence, which has historically tended to be rigid. Negligence is no longer measured by the physical actions of corporate officers, but rather by the decision-making architecture of the computer code (code-as-law) developed by the platform's development team. When a platform intentionally neglects the integration of basic security modules in order to accelerate the product's penetration into the commercial market, such an action meets the legal criteria for a breach of the standard of due care expected in a civilized cyber society.

The burden of proof in this algorithmic negligence lawsuit should apply the doctrine of limited reversal of the burden of proof (res ipsa loquitur). Given that the algorithm's source code and training data are under the exclusive control of the tech giant as the defendant, it is impossible for an individual plaintiff to prove the specific coordinates of the code's error. Therefore, the judge must order the platform provider to prove that they have implemented all optimal risk mitigation measures and demonstrate that the system failure was solely caused by unavoidable digital force majeure. Compensation for non-economic damages in cases of reputational harm caused by AI platform negligence must also be reassessed using a progressive civil law framework. Reputational damage in the digital age is permanent because digital traces cannot be completely erased from the internet's collective memory. Judges must no longer award nominal damages that amount to nothing for the billion-dollar budgets of tech corporations. Damages must be designed to be punitive in nature to compel the industry to shift its business orientation from surveillance capitalism toward a digital ecosystem that respects human dignity.

Furthermore, it must be emphasized that the right to reputation is a manifestation of a subjective right protected by civil law, the violation of which automatically triggers the victim's right to seek full restitution (restitutio in integrum). In the digital context, this remedy is not sufficient if it is merely in the form of

monetary compensation; rather, it must be accompanied by active digital rehabilitation. Platforms may be ordered to fund search engine optimization (SEO) to suppress negative links and distribute official clarification videos across their entire distribution network. The algorithmic negligence of platform providers is also evident in their inability to align technological performance with the standards of social propriety that prevail in society (maatschappelijk betamelijk). In contemporary civil law doctrine, "fault" is no longer measured solely by violations of written law, but also by disregard for the values of decency and prudence that should be inherent in a technology professional. Systemic disregard for the potential dangers of unauthorized visual processing constitutes a clear violation of the principle of good faith (te goeder trouw) in conducting commercial operations within the legal jurisdiction of Indonesia.

The technological information asymmetry between platform providers and ordinary users demands a reorientation of the concept of civil legal relationships. Platforms can no longer use the defense that users act independently, free from the platform's influence, because the architecture of digital interfaces is deliberately designed to be addictive (dark patterns) to encourage users to produce massive amounts of content without considering the ethical consequences. Thus, there is a close causal relationship between platform design incentives and the reputational harm suffered by third-party victims. From a socio-legal perspective, civil liability for algorithmic negligence also serves as an instrument for the fair distribution of risk (risk distribution theory) in the digital age. Those who create risks in society in order to reap economic benefits from the use of such technology should rightfully bear the full financial burden resulting from the failure of that technology. Shifting the burden of reputational harm and the costs of psychological recovery onto vulnerable individual victims represents a failure of the law to uphold the distributive function of justice in cyberspace. Researchers also emphasize the importance of expanding the concept of damages in cyber civil law to encompass not only actual material losses (damnum emergens) but also the loss of expected profits (lucrum cessans). Victims whose reputations are destroyed by deepfake manipulation often lose business opportunities, face termination of professional employment contracts, and even the cancellation of commercial projects. Therefore, civil courts in Indonesia must apply a comprehensive method for calculating future economic losses to ensure the full restoration of victims' financial rights.

Finally, this tort claim may be filed collectively through a class action mechanism if the deepfake platform is found to have committed systemic violations that harm numerous individuals simultaneously on a mass scale. This collective civil legal action is crucial for balancing the economic and political power disparity between civil society and global technology corporations. Through massive and structured civil litigation pressure, the artificial intelligence industry is forced to adopt ethical standards that prioritize the protection of the right to human reputation and dignity over the mere pursuit of capital accumulation.

3.3.The Concept of Criminal Liability: Corporate Liability and the Functional Theory

In addition to civil liability, progressive criminal law enforcement is essential to serve as a deterrent to exploitative AI industries. The New Criminal Code (Law No. 1/2023), which has now fully entered into force, expands the scope of criminal liability to include corporations. Platform providers, whether in the form of local limited liability companies (PT) or foreign digital corporations (Over-The-Top / OTT), can be classified as corporations that can be held criminally liable. Under the provisions of Law No. 1/2023, a criminal offense by a corporation occurs if the act is committed within the scope of the corporation's business operations, provides a benefit to the corporation, and the corporation either permits the act or fails to take adequate preventive measures. The researcher employs Identification Theory and Functional Theory to construct the criminal liability of AI platforms:

Willful Blindness: When a platform knows that its algorithmic system is consistently used to produce illegal videos but willfully turns a blind eye to it in order to maintain its monthly active user count, this action is categorized as willful conduct with indirect intent (dolus eventualis). Criminal Complicity (Deelneming): A platform may be held liable as an accomplice (medeplichtige) for intentionally providing the means, opportunity, or assistance in the form of a super-fast AI rendering tool to facilitate the commission of the criminal offense of defamation. The criminal sanctions that may be imposed on the platform-providing corporation are not limited to monetary fines, but also include additional sanctions of a structural-educational nature. These sanctions range from the suspension of the application's operational license within Indonesian jurisdiction, orders for forced modification of the platform's algorithms under the supervision of an independent curator, to the total shutdown of the platform (digital death penalty).

The application of the "willful blindness" doctrine within the realm of national cyber criminal law marks a significant conceptual leap in dismantling the classic defenses of technology corporations. Platform administrators can no longer claim "we didn't know" regarding cybercrimes occurring on their servers. In the era of big data, platform providers have real-time analytics dashboards capable of detecting anomalies in data traffic, the most frequently searched keywords, and the types of files most often processed by their AI systems. Turning a blind eye to the digital knowledge right before their computer screens constitutes a form of corporate criminal intent (mens rea) that warrants criminal prosecution.

Furthermore, the imposition of additional criminal sanctions in the form of the suspension of operating licenses or forced algorithmic restructuring is a criminal justice instrument specifically designed to restore cyber justice. If an AI corporation is proven to consistently disregard public safety in pursuit of capital growth, the courts, acting on behalf of the state, possess digital sovereignty authority to perform surgical interventions on their technological architecture. This decisive step is vital to demonstrate that the state's legal sovereignty must not be subject to or dictated by the power of transnational technology corporate oligarchies. Furthermore, the concept of functional criminal liability affirms that the status of a subject of criminal law is inherent to the corporation because, in sociological reality, the platform acts as an autonomous organic entity. A platform is not merely a stack of hardware and dead code, but a functional legal subject capable of directing the behavior of millions of its users through the engineering of digital incentives. When this functionality is exercised without accounting for the risks of cybercrime, criminal blame automatically shifts from the individual managers to the corporate entity itself.

In the context of proving fault, the element of corporate intent (corporate mens rea) can be inferred from internal policies, the compliance culture, and lax corporate governance structures. If the corporate articles of incorporation or Standard Operating Procedures (SOPs) contain no provisions for preventive safeguards against visual identity manipulation, this demonstrates the existence of systematic institutional negligence. The absence of this culture of legal compliance indicates that the corporation has, from the outset, tolerated the use of its technology for illegal activities in order to boost market valuation. Researchers also propose the application of the Theory of Vicarious Liability in cyber criminal law to hold foreign platforms accountable. This theory allows the state to hold parent corporations liable for criminal acts facilitated by their agents or automated software operating within Indonesian jurisdiction. Thus, there is no longer any room for transnational digital corporations to evade national legal accountability by claiming that their technological operations are run automatically by artificial intelligence without human intervention.

Another important aspect that needs to be analyzed is the criminal doctrine of Corporate Strict Liability for types of digital crimes that cause massive destructive impacts. Given that the dissemination of deepfake-based non-consensual pornography can destroy a victim's social life in an instant, criminal law must abandon the traditional requirement to prove criminal intent. Once a platform is proven to have processed biometric data without authorization and produced visual output that violates dignity, the corporation is immediately deemed to have committed a criminal offense without the need to prove the psychological motivation of its directors. In terms of law enforcement, expanding the concept of criminal complicity under Article 40 of Law No. 1/2023 becomes highly relevant to hold platforms accountable as hidden intellectual authors. Platforms actively design recommendation algorithms that promote provocative deepfake content to go viral in order to boost corporate advertising revenue. This act of amplifying the circulation of illegal content for economic gain shifts the platform's status from merely a passive provider of a medium to an active participant (co-perpetrator) who contributes to expanding the scale of the impact of such cybercrimes.

In addition to high cumulative fines, the type of penalty most feared by technology corporations is the imposition of digital incapacitation. This penalty is enforced through permanent domain name system (DNS) blocking and the removal of applications from official digital app stores operating within Indonesian jurisdiction. The enforcement of this penalty is equivalent to a corporate death penalty because it instantly cuts off their capital flow and business interactions in the domestic market, thereby sending a very strong law enforcement message to other players in the AI industry. Through a socio-legal approach, the criminalization of these platform-providing corporations is also a form of human rights protection with a social dimension. Criminal law must not be harsh only on small individual users of the applications, while being lenient toward the major capital owners who produce these cybercrime machines. Balancing the scale of punishment between end-users and platform providers is an absolute prerequisite for achieving legal order and substantive justice in the era of artificial intelligence.

Finally, future updates to cybercrime laws must incorporate the principle of aggressive extraterritorial jurisdiction into specific legislation governing AI. Any global technology corporation whose services are accessed and cause harm to the reputation of Indonesian citizens must be subject to prosecution in national criminal courts. This legal firmness is essential to ensure that Indonesia's digital sovereignty is not undermined by foreign cyber technology empires that often hide behind the geographical borders of tax haven nations.

3.4. Comparative Analysis of AI Safety Regulations: A Comparative Study of the EU AI Act and Indonesian Cybersecurity Law

Comparative law serves as a crucial methodological tool for uncovering the extent to which Indonesia's positive law lags behind in responding to the landscape of generative artificial intelligence when compared to global jurisdictions. At the international level, the European Union has established the gold standard for artificial intelligence governance through the enactment of the European Union Artificial Intelligence Act (EU AI Act). This regulation adopts a risk-based approach that explicitly classifies AI systems into four risk levels: unacceptable risk, high risk, limited risk, and minimal risk. Based on this classification, deepfake technology is categorized as a system with limited risk that carries an absolute transparency obligation upstream before the product is released to the market.

The transparency requirements under the EU AI Act specifically require platform providers to ensure that any audio or visual output generated by an artificial intelligence system is clearly and legibly labeled as manipulative content or computer-generated content. If platform providers fail to comply with these standards, they will face massive global administrative fines, which could amount to a certain percentage of the corporation's global annual revenue. This approach demonstrates that European law unequivocally positions technology providers as the primary parties responsible for the potential social harms caused by their products a paradigm shift that moves away from legal reliance on end-user fault alone.

Conversely, if we examine Indonesia's cyber legal framework whether through the New ITE Law (Law No. 1/2024) or the PDP Law (Law No. 27/2022) the concept of technology risk classification is entirely absent. Indonesian positive law remains agnostic regarding the type of technological architecture used to commit legal violations. As a result, the domestic regulatory system treats generative artificial intelligence platforms which possess massive social destructive potential on par with conventional blog platforms or providers of static websites. This lack of legal differentiation creates acute legal uncertainty and allows high-risk platforms to operate without adequate preventive oversight from the state.

From the perspective of the sociology of cyber law, the strength of the EU AI Act lies in its success in shifting the compliance burden from the downstream societal sector to the upstream industrial sector. In Indonesia, because our cyber regulations are reactive, the burden of proof and of mitigating the harmful effects of digital manipulation falls squarely on the victims. Victims must prove the video is false, while platforms continue to profit from data circulation. This comparison reveals the sociological reality that Indonesia's current cyber laws still operate under 20th-century logic that idolizes digital market freedom, while the technological landscape has shifted toward aggressive algorithmic capitalism.

Furthermore, the EU AI Act incorporates the concept of periodic independent audits of the source code and training data used by artificial intelligence development corporations. This aims to detect algorithmic bias or potential unauthorized data harvesting as early as the laboratory phase. In Indonesia, let alone conducting source code audits, regulatory authorities such as the Ministry of Communication and Digital Affairs do not even possess formal legal instruments to compel foreign platforms to disclose the transparency of how their recommendation algorithms operate. This structural weakness underscores that Indonesia has not yet achieved technological sovereignty under the law.

Another fundamental discrepancy lies in the European Union's doctrine of ex-ante compliance versus Indonesia's approach of ex-post enforcement. Under the legal framework of the EU AI Act, artificial intelligence systems classified as high-risk must undergo a rigorous conformity assessment before receiving the CE marking to operate legally. This approach closes the loophole that would allow corporations to use the general public as guinea pigs for unproven technology. In contrast, Indonesia's cyber laws leave the digital gateway wide open, allowing web- and mobile-based deepfake applications to be freely downloaded and used by millions of domestic users without undergoing cognitive and ethical security screening by the state from the outset.

Furthermore, the EU AI Act takes a bolder step by strictly prohibiting the use of AI systems that manipulate human cognitive behavior in ways that could cause physical or psychological harm. This regulation provides a strong legal basis for holding platforms criminally liable when their algorithms intentionally steer users toward manipulative content to boost user engagement. In Indonesia, the lack of regulations regarding algorithmic cognitive manipulation forces law enforcement to rely on conventional

fraud or defamation provisions in the ITE Law, which are extremely difficult to meet in terms of criminal elements when the perpetrator is an autonomous AI-based agent.

From an institutional perspective, the European Union has established the European AI Office, which functions as a centralized oversight body with high technical authority to monitor transnational compliance. This agency is staffed by digital forensic experts, data scientists, and cyber law specialists capable of keeping pace with the speed of innovation in the cyber industry. In contrast to this situation, AI governance in Indonesia remains overlapping and fragmented across various ministries and agencies, lacking a single dedicated independent authority with adequate technical expertise. As a result, the policies produced are often hesitant, piecemeal, and fail to provide an integrated solution to the structural problems of digital visual exploitation.

Furthermore, the philosophies underlying the establishment of these two legal frameworks reveal a stark disparity in their orientations. The EU AI Act is built on a human-centric approach that prioritizes the protection of citizens' fundamental rights over market expansion interests. On the other hand, Indonesia's information technology governance policies, including regulations under Government Regulation No. 71/2019 on PSTE, remain heavily influenced by a one-sided focus on digital economic growth and investment facilitation (market-centric approach). It is this compromise in the name of economic growth that often sacrifices the protection of individuals' right to reputation under the shadow of the innovation freedom of giant cyber industries.

Therefore, the urgency of implementing legal transplantation of the EU AI Act's core principles into the national legal framework is undeniable. Indonesia does not need to blindly copy the entire European codification, but the adoption of a risk-based approach and mandatory labeling obligations at the upstream level is non-negotiable. This legal update is urgently needed to align the instruments protecting the right to reputation of Indonesian citizens with the universally recognized standards of digital human rights in the developed world.

3.5. Standardization of Technical Duty of Care: Mandates for Digital Watermarking and Biometric Authentication Systems

The legal framework for civil torts based on algorithmic negligence requires objective technical parameters to determine whether a platform provider has breached its duty of care. Researchers argue that the most fundamental duty of care that the generative AI industry must currently meet technically is the implementation of cryptographic digital watermarking that is invisible and indelible. This watermark is not merely a visual logo in the corner of a video, but rather binary data encryption embedded into the visual file's metadata, enabling search engines and social media platforms to automatically detect and label the video as AI-generated.

The failure of platform providers to integrate digital watermarking technology into their rendering systems must be classified as a form of product defect under consumer protection and civil law theory. Platforms that release super-realistic video creation software to the public without accompanying digital safeguards are akin to launching commercial motor vehicles without a braking system. Such actions constitute a clear case of design negligence that directly opens the door wide to violations of the reputation rights of innocent third parties. In addition to digital watermarks, the second layer of technical safeguards that must be mandated by law is the implementation of a Biometric Consent Verification System. Before a platform's algorithm is permitted to process facial or voice modifications, the system must require the user to upload written proof of consent or real-time identity verification from the individual whose face will be manipulated. If a user uploads a photo of another person's face without a valid consent authentication token, the platform's algorithm must automatically block that prompt and lock the user's account.

Through socio-legal analysis, platforms' reluctance to implement these technical restrictions stems from their fear of losing user convenience (user friction). The tech industry has always championed the principles of decentralization and barrier-free access in pursuit of exponential user growth targets. However, in a vulnerable cyber society, convenience without accountability is a recipe for digital anarchy. Civil law must step in to compel platforms to balance their profitability interests with the obligation to keep the digital space safe for human dignity and respect.

This technical standardization also serves as a tool to shift the burden of proof in civil court. If a victim files a tort claim, the judge need only examine whether the harmful deepfake video contains digital watermark metadata from the defendant's platform. If evidence is found that the platform released the video without adequate labeling encryption, then the platform's negligence is deemed legally proven (ipsa loquitur). This provides very strong and efficient legal protection for victims in asserting their civil rights.

Furthermore, the determination of the operational boundaries of the standard of care must not be left to the self-regulatory mechanisms of the technology industry, which tend to be permissive. The state, through its telecommunications regulatory authority, is obligated to codify this "security by design"

architecture into legally binding technical regulations. When artificial intelligence system developers distribute their products within Indonesian jurisdiction without complying with cryptographic protocols for file origin marking, such corporations have automatically committed statutory negligence (negligence resulting from a violation of the law). This legal approach instantly nullifies any form of unilateral disclaimer of liability typically included in the clauses of their electronic contracts.

From the perspective of cyber-tracking efficiency, the adoption of digital watermarking technology based on blockchain architecture or open consortia such as the Coalition for Content Provenance and Authenticity (C2PA) must be viewed as a professional obligation of digital integrity. This upstream tagging creates a transparent and accountable provenance chain from the moment digital content is first created until it is distributed on social media. Any negligence or intentional act by developers to sever this data provenance chain constitutes material evidence of a breach of civil good faith, thereby giving rise to the public's right to seek compensation for the risk of immaterial losses suffered due to exposure to visual disinformation. Furthermore, the mandatory implementation of a biometric authentication system at the account registration (onboarding) stage for generative AI software acts as a digital moral filter. The concept of civil prudence requires developers to anticipate potential misuse of the technology (foreseeable misuse) by consumers. Given the inherent nature of deepfakes, which are highly susceptible to exploitation for the creation of non-consensual pornography and political character assassination, the absence of a verification gateway for the user's true identity and biometric facial mapping of the target constitutes a form of reckless disregard for the safety of the public's subjective rights.

From a socio-legal perspective, the standardization of these technical safeguards will necessitate a restructuring of economic incentives within the domestic artificial intelligence market. Platforms will no longer be able to engage in unfair competition by offering unlimited visual manipulation features to inflate their venture capital valuations, while shifting the negative externalities—such as the social trauma suffered by victims—onto the state's judicial system. The requirement to include biometric consent authentication features ensures that the cost of cybercrime is internalized into corporate production costs, thereby fostering ethical industry competition that prioritizes the protection of human dignity.

Finally, this technical standardization strengthens the bargaining position of national civil cyber law enforcement in testing the validity of electronic evidence in court. With standardized digital watermarks, the digital forensic process no longer takes months and incurs high costs that often dampen victims' resolve to seek justice. Synchronization between formal civil law and these encrypted computer security standards is an absolute prerequisite for transforming Article 1365 of the Civil Code into a tactical, progressive, and responsive digital litigation instrument against the threat of algorithmic repression in the modern era..

3.6. Mechanisms for Restoring Victims' Rights (Restitutio in Integrum) Through a Cyber Restorative Justice Approach

Law enforcement regarding violations of the right to reputation caused by deepfake technology must not stop merely at punishing the perpetrators or imposing administrative fines on the state treasury. The primary focus of just law must be placed on the full restoration of the victim's psychological, social, and digital condition to its original state prior to the crime (restitutio in integrum). Researchers propose the integration of the Cyber Restorative Justice Approach as a progressive legal resolution model to bridge the interests of victims and the responsibilities of platform providers.

Cyber restorative justice shifts the focus from traditional retributive justice—which emphasizes physical retribution—to functional restorative justice. In cases of massive-scale digital defamation, the form of restoration most needed by victims is the permanent destruction of the digital traces of manipulative content from the internet ecosystem. This is where platform providers play a central role as the executors of restoration. Platforms should no longer remain passive and wait for complaints; rather, they must be legally mandated to use their search algorithms to track down, block, and delete all copies of such deepfake videos across their entire network of servers.

This restorative model also requires platforms to fund digital reputation rehabilitation projects for victims. These rehabilitation costs include funding for digital public clarification campaigns using promotional optimization algorithms, the provision of professional psychological counseling services for victims, and direct financial compensation for the loss of victims' future economic opportunities. Through

the platform's active involvement in funding the recovery, the state successfully shifts the burden of social harm from vulnerable individuals onto technology corporations with massive economic capacity.

Sociologically, the application of restorative justice in cyberspace faces the challenge of resistance from the public, who have already consumed visual hoaxes. Therefore, restorative legal rulings must require platforms to display automated correction notifications to every internet user detected as having previously viewed or shared such illegal deepfake videos. This radical step is highly effective in breaking the chain of social stigma and massively educating the public about the falsity of such content.

More specifically, the dimension of restorative cyber justice calls for a reconfiguration of the "right to be forgotten" as stipulated in Article 26 of the ITE Law so that it can be applied proactively. In the context of deepfake victims, this right is often ineffective because the process of removing inaccurate electronic information requires a court order, which is a time-consuming process. Through a cyber-restorative framework, AI platforms are mandated to develop automated filtering systems that prevent the re-uploading of victims' biometric digital assets that have been declared legally invalid. Shifting the burden from bureaucratic judicial mitigation to autonomous technological obligations is a cornerstone of realizing the essence of restitutio in integrum in the digital age.

From a legal psychology perspective, the trauma experienced by victims of visual manipulation is no less destructive than that experienced by victims of physical violence in the real world. The manipulation of a victim's face or voice into obscene or disinformative narratives creates an instant character assassination that triggers social isolation and acute anxiety disorders. Therefore, traditional civil remedies that assess damages solely based on tangible material losses are considered highly reductive. A restorative cyber approach breaks this rigidity by incorporating mandatory mental health restitution costs—which must be fully borne by platform providers—as an integral part of their liability for the negative externalities of their AI products.

Furthermore, the active participation of victims (a victim-centered approach) is an absolute prerequisite in determining the parameters of digital restitution. In cyber restorative mediation forums facilitated by independent authorities, victims are granted full voting rights to determine the most effective method of reputation restoration for themselves, such as demanding the display of an official apology on the platform's homepage for a specified period. This contrasts with conventional retributive criminal justice, where victims are often relegated to the role of mere crown witnesses, alienated from the process of determining the restoration of their own honor.

In the context of international civil law enforcement, this cyber restorative mechanism offers a solution to the cross-border jurisdictional deadlock. When AI-providing platforms operate from abroad without a physical representative office in Indonesia, the enforcement of conventional compensation rulings is nearly impossible. However, by threatening a nationwide boycott of commercial access if the platform refuses to facilitate a restorative recovery program for Indonesian citizens, the state can compel such transnational corporations to comply with a cyber-restorative contract. This approach underscores that a nation's digital sovereignty is manifested through the extraterritorial protection of its citizens' human rights.

From a sociological perspective of cyber law, the institutionalization of cyber restorative justice also serves as a means of collective re-education for netizens who often act as perpetrators of stigma reinforcement (cyberbullying). When platforms disseminate algorithmic correction notifications, this simultaneously dismantles the cognitive biases of the public that have already come to trust false content. Law, therefore, no longer appears as a fearsome stick at the end of the chain, but rather as a social architect that restructures the public's ethical awareness and digital literacy so that they do not easily exploit the suffering of fellow human beings in the virtual space.

Finally, national cyber law reform must establish these restorative cyber mechanisms not as a voluntary option for platforms, but as a mandatory legal obligation. A platform's failure to facilitate the process of restoring a victim's right to reputation must result in the imposition of harsher corporate criminal sanctions. Only through this combination of punitive civil sanctions and humane restorative mechanisms can the law restore the human dignity that has been temporarily stripped away by the unchecked development of artificial intelligence in Indonesia.

Conclusion

This study concludes that Indonesia's current legal landscape is experiencing an acute regulatory lag in the face of visual exploitation based on artificial intelligence (generative AI deepfakes). The Safe Harbor doctrine, which has long served as a shield of impunity for cyber platform providers, has proven to be no longer relevant and, in fact, perpetuates systemic injustice against victims of violations of the right to reputation. Through the lens of contemporary civil law, platforms can no longer be positioned as passive intermediaries but rather as functional legal entities that can be held accountable through the legal framework of Tort (PMH) based on algorithmic negligence. This negligence manifests in the platform's failure to apply technical standards of care (duty of care) at the upstream level, such as the absence of permanent cryptographic digital watermarking mechanisms and mandatory biometric consent verification systems for users.

In the realm of criminal law, the enactment of the New Criminal Code (Law No. 1/2023) opens the door to expanding corporate criminal liability through the doctrine of willful blindness and the functional theory. Platforms that intentionally allow the circulation of harmful content to secure economic gains from user engagement may be deemed to possess corporate mens rea and warrant sanctions ranging from operational suspension to digital incapacitation. Through a comparative study with the EU AI Act, this research underscores the importance for Indonesia to immediately shift its technology governance paradigm from one that is market-centric and reactive (ex-post) to a human-centric and preventive (ex-ante) approach based on risk stratification.

As a final recommendation, the restoration of victims' digital human rights must be grounded in the principle of restitutio in integrum through the institutionalization of a cyber restorative justice mechanism. The responsibility for digital footprint restoration, automatic rehabilitation of reputation, and punitive non-pecuniary damages must be shifted from the shoulders of vulnerable victims to technology corporations possessing colossal economic capacity. An integrative, responsive, and extraterritorially sovereign national cyber law reform is an absolute prerequisite for protecting the human dignity of Indonesian citizens from the threat of algorithmic colonialism in the modern era.

Author Contributions

Milka Kumala: Conceptualization, Methodology, Project administration, Writing - original draft preparation.

Tuti Handayani: Data curation, Investigation, Validation.

Muhammad Rizki: Formal analysis, Resources, Software.

Reza Dipta Prayitna: Supervision, Visualization, Writing - review & editing.

All authors have equal contributionsto the paper. All the authors have read and approved the final manuscript.

Funding

This research and the writing of this scientific article were conducted independently and funded entirely by the authors. No funding was received from any public, commercial, or nonprofit sponsors for this research.

Declaration of Conflicting Interests

The authors declare that there are no potential conflicts of interest—whether financial, personal, or professional—with any party or organization that could affect the objectivity, conduct, writing, and/or publication of this scientific article. Declarations of interest: none.

Acknowledgment (Optional)

The author would like to express his deepest gratitude to the academic community of the Faculty of Law at Cipta Wacana Christian University for the academic support provided throughout the research

process. The author also extends his appreciation to colleagues in the constitutional and socio-legal studies community in Malang, whose discussions and valuable critical feedback have helped refine the ideas presented in this article.

References

A. Mustofa (2022) "Pertanggungjawaban Pidana Korporasi atas Kebocoran Data Pribadi Berdasarkan UU PDP," *Jurnal Hukum & Pembangunan*, vol. 52, no. 3, pp. 677-694

A. Sofyan (2024) "Algorithmic Negligence: Menggugat Kelalaian Penyedia Platform AI Lewat Pasal 1365 KUHPerdata," *Jurnal Hukum Perdata*, vol. 16, no. 2, pp. 142-158

A. Wicaksono (2023) "Pertanggungjawaban Perantara Digital (Intermediary Liability) dalam Hukum Siber Indonesia," *Jurnal Hukum Teknologi*, vol. 7, no. 1, pp. 34-51

B. Harris (2021) "Deepfakes, Fake Porn, and the Cyber-Violence Against Women: A Socio-Legal Analysis," *International Journal of Cyber Criminology*, vol. 15, no. 1, pp. 88-104

B. Santoso (2025) "Socio-Legal Methodology in Conceptualizing Digital Harms in Indonesia," *Indonesian Journal of Socio-Legal Studies*, vol. 4, no. 1, pp. 75-92

C. E. A. Karnow (2017) "The Application of Strict Liability to High-Risk Artificial Intelligence Systems," *Ohio State Technology Law Journal*, vol. 13, no. 2, pp. 301-329

D. K. Citron and R. Chesney, (2019) "Deepfakes and the New Information Warfare," *California Law Review*, vol. 107, no. 6, pp. 1753-1822

H. Siregar (2023). *Hukum Pidana Baru Indonesia: Komparasi KUHP Lama dan UU No. 1/2023*. Jakarta: Sinar Grafika

J. Fransiska and T. Prasetyo (2023) "Penerapan Keadilan Restoratif Berdasarkan Undang-Undang Perlindungan Data Pribadi," *Jurnal Hukum Pidana dan Penegakan Hukum*, vol. 5, no. 1, pp. 45-59

J. Lipton (2019) "The Future of Section 230 and Platform Liability for Deepfakes," *Harvard Journal of Law & Technology*, vol. 33, no. 1, pp. 143-168

K. Barker (2018) "Online Defamation and the Right to Reputation: Choosing the Mirror over the Window," *Journal of Media Law*, vol. 10, no. 2, pp. 115-138, 2018.

K. Wijaya (2025) "Green Arbitration and Digital Disruption: Emerging Trends in ADR," *Journal of Alternative Dispute Resolution*, vol. 12, no. 2, pp. 201-218

M. Zimmer (2028) "The Ethics of Using Synthesized Media Data in Socio-Legal Research," *Journal of Empirical Legal Studies*, vol. 15, no. 4, pp. 812-835

M. Irwan (2024) "Dinamika Pembatasan Kekuasaan Negara dalam Konteks Transformasi Digital di Indonesia," *Jurnal Konstitusi*, vol. 21, no. 2, pp. 210-232

M. Wardaya (2024) "Penyerangan Kehormatan di Ruang Siber Menurut Hukum Pidana Nasional," *Mimbar Hukum*, vol. 36, no. 1, pp. 112-130

N. Diakopoulos (2016) "Accountability in Algorithmic Decision Making," *Communications of the ACM*, vol. 59, no. 2, pp. 56-62

P. M. Marzuki (2021) *Penelitian Hukum: Edisi Revisi*. Jakarta: Prenadamedia Group

P. Yudha (2024) "Model Pemidanaan Korporasi dalam UU No. 1 Tahun 2023 bagi Kejahatan Berbasis Kecerdasan Buatan," *Jurnal Hukum Pidana Indonesia*, vol. 19, no. 2, pp. 155-171

R. Wacks (2020). *Privacy and Media Freedom*. Oxford: Oxford University Press

R. Ramadhan (2024) "Analisis Perubahan Kedua UU ITE: Menjamin Kebebasan Berpendapat atau Memperkuat Represi Digital?" *Jurnal Komunikasi Hukum*, vol. 10, no. 1, pp. 12-29

R. Tambunan (2025) "Deepfakes and the Erosion of Democratic Trust: A Constitutional Perspective," *Indonesian Constitutional Review*, vol. 8, no. 1, pp. 45-68

S. A. Nugroho (2023) "Doktrin Safe Harbor dan Relevansinya dalam Menjerat Kejahatan Siber Kontemporer," *Jurnal IUS Kajian Hukum*, vol. 11, no. 2, pp. 189-204

S. M. Kircengaard (2020) "Smart Contracts and the Traditional Law of Obligations: A Comparative Study," *Computer Law & Security Review*, vol. 36, p. 105-11

T. Sukesti (2022) "Pelanggaran Hak Biometrik dan Tantangan Hukum Acara Pidana Indonesia," *Jurnal Hukum Progresif*, vol. 10, no. 3, pp. 310-325

W. Utami (2023) "Digital Gender-Based Violence: Ketiadaan Hukum Proaktif dalam Menangani Non-Consensual Deepfake Pornography," *Jurnal Wanita dan Hukum*, vol. 4, no. 2, pp. 89-107