Due Diligence Responsibilities of Social Media Companies for AI-Based Deepfakes in Efforts to Provide Legal Protection for Consumers

[1]Sabrina Diah Puspita, [2]Nucraha Alhuda Hasnda
[1,2] Universitas Nusa Putra, Indonesia
Email:sabrinaapuspita@gmail.com

This article aims to analyze the responsibility of social media companies based on the principle of due diligence regarding the dissemination of artificial intelligence-based deepfakes and to examine efforts to strengthen legal protection for consumers in Indonesia. The focus of this study is on the implementation of the due diligence principle by social media companies and the mechanisms of legal protection and remedies available to consumers or victims harmed by the dissemination of deepfakes. To address this issue, this study draws on theories of legal responsibility, consumer protection, and the principle of due diligence in digital platform governance. Data were collected through library research by examining legislation, scientific journals, books, international instruments, and other relevant legal sources, and were analyzed qualitatively using statutory and conceptual approaches. The results show that Indonesian regulations provide a general legal basis for holding digital platforms accountable; however, specific regulations governing deepfakes and platform due diligence standards remain fragmented and have not yet been comprehensively established, resulting in regulatory gaps in addressing deepfake-related risks. This study concludes that strengthening legal protection should focus on risk assessment, prevention, detection, reporting, and content removal mechanisms, transparency, as well as effective remedies and dispute resolution for consumers or victims.
Keywords: Artificial Intelligence; deepfake; due diligence.

This is an open access article under theCC BY-NClicense

Corresponding Author:
Sabrina Diah Puspita
Universitas Nusa Putra, Indonesia
sabrinaapuspita@gmail.com

1. Introduction

Advances in information and communication technology have brought about significant changes in various aspects of human life. One technological development that has had a major impact is *artificial intelligence* (AI), a branch of computer science that enables machines (computers) to perform tasks as well as humans do (Septian, 2024). The rise of social media has transformed how people obtain, convey, and disseminate information. Social media no longer serves merely as a means of communication and interaction; it has evolved into a platform for rapidly, effectively, and efficiently disseminating a wide range of information to a broad audience. A variety of features—such as images, videos, messages, and interactive content—allow information to be presented in a more engaging manner and made easily accessible to the public.

The distinct characteristics of social media platforms facilitate the rapid dissemination of information to a broad audience, while the development of cyberspace has also created new patterns of digital and hybrid crime that require adaptive legal responses (Runturambi & Arifin, 2025). The development of social media facilitates the rapid and widespread dissemination of information, allowing various types of content to reach the public in a relatively short time. On the one hand, this makes it easier to obtain and distribute information; on the other hand, it also poses a risk regarding the spread of false or inauthentic content (de Azevedo Cunha & Monteleone, 2021).

Technological advancements not only facilitate the creation of innovative content but also enable the production of digital content that closely resembles reality even after being manipulated. One such

technological development is the *deepfake*. "Deepfake" is a term used to describe techniques for forging digital content—particularly images, videos, and audio—using *artificial intelligence* (AI). This technology enables the creation of videos that appear highly realistic yet are entirely manipulated. According to a study by Lumenta (2026), *deepfakes* are digital content generated or manipulated using artificial intelligence and *deep learning* technology. Such manipulation can be applied to various forms of media—such as faces, voices, images, and videos—through techniques like *face swapping*, *face reenactment*, facial attribute manipulation, and audio-visual manipulation, resulting in synthetic content that closely resembles reality.

The spread of AI-based *deepfakes* can cause harm to consumers because they produce digital content that closely resembles reality, making it difficult to distinguish from authentic media. This situation can fuel fraudulent practices in the digital marketplace and influence consumers' willingness to accept and trust certain information (Ahmed et al., 2023). The study explains that *deepfakes* can have various negative impacts on consumers, including defamation, harassment, intimidation, identity theft, and other forms of harm. Furthermore, the manipulation of information through *deepfakes* can result in financial losses and ultimately erode consumer trust in organizations and the digital marketplace (Putra, 2024; Singh et al., 2025).

Social media platforms such as Facebook, Instagram, TikTok, and YouTube serve as channels for disseminating information, enabling content to reach users quickly and widely. This situation also increases the potential for the spread of *deepfakes*, as manipulated content can be easily shared with and received by a large number of users. From a consumer protection perspective, the spread of *deepfakes* via social media is problematic because this technology can be used in misleading advertisements or promotions that are difficult to distinguish from authentic content. This situation can influence consumer decisions and potentially cause harm to consumers (Mukhlis et al., 2023). Legal liability is the obligation of a party to be held accountable for its actions or negligence if they cause harm to another party.

In the context of social media, platforms—as operators of electronic systems—bear responsibility for managing user content, particularly when illegal or harmful content is present. A platform's negligence in preventing, moderating, or addressing problematic content after becoming aware of it can give rise to legal liability if harm to users or consumers can be linked to such negligence. Therefore, platforms are not only required to provide services but also to take reasonable steps to prevent and address the spread of illegal or harmful content through moderation and *notice-and-takedown* mechanisms (Dewi & Basuki, 2026).

Due diligence is the obligation to take reasonable and adequate steps to identify, prevent, and mitigate risks that could result in harm. In the context of digital platforms, *due diligence* is implemented through risk assessment, content monitoring and detection, reporting and complaint mechanisms, flagging and restricting the distribution of harmful content, takedowns, notifications to users, and evaluation of the platform's moderation system. This obligation positions platforms not only as service providers but also as parties required to undertake preventive and mitigating measures against risks arising from their activities. Consequently, the question arises as to whether a social media company's failure to exercise *due diligence* can serve as a basis for legal liability when the dissemination of *deepfakes* causes harm to consumers (Hermansyah, 2025).

Consumer protection against the spread of *deepfakes* has a legal basis in several regulations in Indonesia. The 1945 Constitution of the Republic of Indonesia provides a constitutional foundation through guarantees of personal protection and the right to obtain and convey information. Furthermore, Law No. 8 of 1999 on Consumer Protection is relevant because it regulates consumer rights, the obligations and responsibilities of business entities, and the prohibition against providing misleading information (Nurkholisah et al., 2025). In the digital realm, Law No. 11 of 2008 on Information and Electronic Transactions (ITE), as last amended by Law No. 1 of 2024, serves as the legal basis regarding electronic information and documents as well as

the governance of the digital space. Meanwhile, Law No. 27 of 2022 on Personal Data Protection provides protection for personal data, including provisions regarding the rights of data subjects, data processing, and the obligations of parties that process personal data.

Additionally, Government Regulation No. 71 of 2019 on the Operation of Electronic Systems and Transactions mandates that Electronic System Operators ensure their systems do not facilitate the dissemination of prohibited electronic information or documents. This framework serves as a crucial basis for assessing platforms' obligations to prevent and address harmful digital content, including the dissemination of *deepfakes* (Putri & Gorda, 2025).

One real-world case of *deepfake* technology abuse in Indonesia involved the use of President Prabowo Subianto's face in content falsely offering government financial aid. This scheme caused losses to more than 100 victims spread across 20 provinces, with total losses amounting to approximately Rp65 million. This case demonstrates that *deepfake* technology is not only capable of being used to generate false information but can also be exploited as a means of fraud that harms the public. Although the perpetrators were successfully apprehended and prosecuted, the case highlights the ongoing limitations of regulations in specifically addressing the misuse of *deepfake* technology, resulting in law enforcement relying on existing provisions within the legal framework. This situation highlights the need to strengthen mechanisms for preventing and addressing the spread of *deepfakes*, including assessing the responsibility of social media companies in monitoring and mitigating content that has the potential to harm users (Heskia Putri & Astuti, 2026).

Various legal studies in Indonesia indicate that regulations such as the ITE Law, the Pornography Law, the Personal Data Protection Law (PDP Law), the Criminal Code (KUHP), and the Sexual Violence Crimes Law (TPKS Law) contain provisions that can be used to prosecute perpetrators of abuse (Kristiyenda et al., 2025). However, there is not yet a single regulation that explicitly defines and classifies *deepfakes* as a distinct criminal offense, complete with a definition, elements of the offense, digital evidence, the division of liability between creators and distributors, and specific penalties for *deepfakes*.

Existing studies have therefore provided important discussions on the harmful effects of *deepfakes*, consumer protection, the applicability of existing Indonesian legislation, and the legal liability of perpetrators. Nevertheless, these discussions have not sufficiently examined the specific position of social media companies as digital intermediaries when AI-based *deepfakes* disseminated through their services cause harm to consumers. In particular, previous studies have not comprehensively connected platform liability with *due diligence* standards covering risk assessment, prevention, detection, content moderation, reporting, *takedown*, and remedies for affected consumers within the fragmented Indonesian regulatory framework. The research gap therefore lies not merely in the absence of a specific *deepfake* regulation, but in the lack of an integrated legal analysis explaining when and to what extent a social media company's failure to exercise *due diligence* may constitute a basis for liability toward consumers. This study addresses that gap by integrating the principle of *due diligence*, digital platform liability, and consumer protection into a single analytical framework for assessing AI-based *deepfake* dissemination in Indonesia.

This situation naturally raises a critical question that forms the core of this study: How does the principle of *due diligence* serve as the basis for social media companies' liability regarding the dissemination of AI-based *deepfakes*, and what efforts can be made to strengthen legal protections for consumers against the dissemination of AI-based *deepfakes* through social media companies in Indonesia?

Based on this problem formulation, this study aims to analyze the legal liability of social media companies based on the principle of *due diligence* regarding the dissemination of AI-based *deepfakes* that have the potential to cause harm to consumers, as well as to examine the platforms' obligations in preventing,

detecting, handling, and mitigating *deepfake* content in accordance with applicable Indonesian law, thereby formulating measures to strengthen legal protection for consumers in the digital space.

The benefits of this study are both theoretical and practical. Theoretically, this study is expected to enrich the legal analysis of information technology, particularly regarding the legal liability of social media companies and the application of the *due diligence* principle in addressing the spread of AI-based *deepfakes*, as well as to broaden understanding of legal protection for consumers in the digital space. Practically, this research is expected to serve as a basis for consideration by policymakers, social media companies, academics, and law enforcement officials in strengthening mechanisms for prevention, detection, handling, and oversight regarding the spread of *deepfakes* that harm consumers.

The theoretical framework underpinning this study draws on several legal theories, including the theory of legal protection (Philipus M. Hadjon), which emphasizes the importance of protecting the rights of the public through preventive and repressive measures; the theory of legal responsibility (Hans Kelsen), which explains the relationship between violations of norms and the obligation to bear legal responsibility for the resulting consequences; and the theory of *law and technology*, which emphasizes the need for the law to adapt to developments in digital technology. This study also employs the concept of the *due diligence* principle as a foundation for assessing social media companies' obligations to identify, prevent, and mitigate the risks of *deepfake* dissemination to users and consumers.

Based on this analysis, this study considers it necessary to strengthen the accountability of social media companies through the implementation of clearer and more measurable *due diligence* principles, accompanied by effective mechanisms for prevention, detection, reporting, distribution restrictions, *takedowns*, and complaints. Such strengthening is necessary so that advancements in AI technology can be accompanied by a legal protection system capable of providing certainty, justice, and effective protection for consumers in the digital space.

2. Methodology

This study employs a normative legal research method using both the *statutory approach* and the *conceptual approach*. The statutory approach is used to analyze provisions related to consumer protection, electronic information and transactions, personal data protection, and the operation of electronic systems (Irwansyah, 2020; Yanova et al., 2023). The conceptual approach was used to examine the concepts of legal liability, legal protection, and the principle of *due diligence* in determining the obligations of social media companies regarding the dissemination of AI-based *deepfakes*. The legal materials used consisted of primary legal materials in the form of laws and regulations; secondary legal materials in the form of books, scientific journals, and relevant research findings; and tertiary legal materials in the form of legal dictionaries and other supporting sources.

Data were obtained through a literature review and analyzed qualitatively by linking legal norms and the concept of *due diligence* to the issue of *deepfake* dissemination that harms consumers. The normative legal analysis was conducted in several stages. First, the relevant provisions in each regulation were identified and classified according to their regulatory scope, particularly provisions concerning consumer rights, obligations of electronic system operators, content governance, personal data protection, prevention of unlawful content, and mechanisms for addressing consumer harm. Second, these provisions were interpreted and systematically connected with the concept of *due diligence* to determine the forms of preventive and responsive obligations that may apply to social media companies, including risk identification, prevention, detection, content moderation, reporting, *takedown*, and mitigation measures. Third, the identified obligations were analyzed in relation to the principles of legal liability to determine the legal basis on which a platform may be held accountable when its failure to exercise reasonable preventive

or responsive measures contributes to consumer harm caused by AI-based *deepfakes*. Fourth, the regulations were compared and evaluated to identify fragmentation, normative gaps, overlaps, and the absence of specific standards governing *deepfakes* and platform *due diligence*. The results of these stages were then synthesized to formulate the forms of legal liability of social media companies and measures to strengthen legal protection for consumers.

3. Results And Discussion

The Principle of Due Diligence as the Basis for Social Media Companies' Liability Regarding the Spread of AI-Based Deepfakes

Social media companies are digital platform operators that provide systems for users to create, upload, store, and disseminate information. Under Indonesian law, companies that provide electronic services can be categorized as Electronic System Operators (ESOs). This status imposes an obligation on companies to operate electronic systems responsibly and to provide protection for users (Shabrina et al., 2026). In the spread of *deepfakes*, the role of social media companies becomes increasingly significant because platforms can expand the reach of content through distribution and recommendation systems. This situation indicates that companies not only provide a space for users to upload content but also play a role in determining how that content is disseminated to other users (Wulandari et al., 2025).

AI-based *deepfakes* create particular legal risks because manipulated images, videos, or audio can closely resemble authentic content and can be rapidly distributed through social media. For consumers, such content can facilitate fraud, identity theft, intimidation, harassment, defamation, and other forms of harm. The difficulty of identifying manipulated content also increases the importance of preventive and responsive measures by platforms (Karpinska-Krakowiak & Eisend, 2025). From a normative perspective, these characteristics are relevant not merely because *deepfakes* are technologically difficult to detect, but because they raise the question of whether the existing obligations imposed on Electronic System Operators are sufficiently specific to require platforms to identify, prevent, detect, and mitigate such risks.

The implementation of *due diligence* on digital platforms begins with risk assessment and prevention. UNESCO emphasizes that platforms should conduct regular risk assessments to identify actual and potential impacts and determine appropriate mitigation measures (UNESCO, 2023). Within the Indonesian legal framework, this preventive dimension can be linked to the general obligation of Electronic System Operators to operate electronic systems reliably, securely, and responsibly, as well as to ensure that their systems do not facilitate the dissemination of prohibited electronic information or documents. The Personal Data Protection Law also provides a more specific risk-based approach where personal data processing presents a high risk, including through the use of new technologies. These provisions provide a general normative basis for preventive obligations, particularly where *deepfake* production or dissemination involves personal or biometric data. Nevertheless, Indonesian regulations do not yet specifically require social media companies to conduct periodic *deepfake*-related systemic risk assessments or prescribe measurable risk-mitigation standards comparable to the broader *due diligence* framework. Consequently, the legal basis for prevention exists at a general level, but its application to AI-based *deepfake* risks remains fragmented and requires further specification.

Detection and monitoring constitute the next component of *due diligence*. Platforms may use automated tools and human review to identify potentially unlawful or harmful content, while labels or warnings may assist users in recognizing manipulated or synthetic content. UNESCO's transparency principles further emphasize that platforms should explain the systems and processes used in content moderation and curation (UNESCO, 2023). Indonesian regulations provide a basis for requiring Electronic System Operators to manage prohibited electronic information and maintain the reliability and security of their systems. These

obligations can support an interpretation that platforms must possess mechanisms capable of responding to unlawful content circulating through their services. Nevertheless, the existing framework does not expressly prescribe technological standards for detecting AI-generated *deepfakes*, the circumstances in which synthetic-content labels must be applied, or the level of accuracy expected from automated detection systems. Detection therefore represents an area in which a general platform obligation can be identified, while the specific standard of care required for *deepfake* detection remains legally ambiguous.

The *notice-and-action* mechanism provides a more concrete connection between *due diligence* and existing platform obligations. Such a mechanism enables users to report content suspected of violating the law and allows platforms to assess whether further action is necessary. The Digital Services Act requires an accessible reporting mechanism through which sufficiently clear notifications can be submitted (Hasibuan et al., 2025). In Indonesia, the regulatory framework for private Electronic System Operators similarly provides mechanisms concerning prohibited electronic information and access termination, thereby establishing a normative basis for platforms to respond after receiving information about unlawful content. Permenkominfo No. 5 of 2020, as amended by Permenkominfo No. 10 of 2021, strengthens this responsive dimension by regulating obligations of private Electronic System Operators in relation to prohibited electronic information and access termination. The Indonesian framework therefore provides a clearer basis for reactive action than for proactive *deepfake* risk assessment. Nevertheless, because *deepfakes* are not comprehensively classified as a specific category of prohibited content, uncertainty may arise in determining when manipulated content reaches the legal threshold requiring restriction or removal.

After a platform becomes aware of unlawful content, *takedown* or distribution restrictions may constitute appropriate mitigation measures. These actions may include content removal, restrictions on access or visibility, suspension of monetization, or restrictions on accounts disseminating problematic content (Abbas & Taeihagh, 2024). Indonesian rules on Electronic System Operators provide a legal basis for access termination in relation to prohibited electronic information, which can support *takedown* obligations where a *deepfake* falls within an existing category of unlawful content. The normative problem arises where a *deepfake* is misleading or harmful to consumers but does not clearly fall within an expressly prohibited category. Consequently, the current framework can support *takedown* obligations in particular circumstances, but it does not yet establish a comprehensive and specific standard governing when platforms must remove, restrict, label, or reduce the distribution of harmful *deepfake* content.

The implementation of *due diligence* also requires complaint and remedy mechanisms. Users affected by harmful content or moderation decisions should have access to procedures for submitting complaints, obtaining explanations, and seeking appropriate remedies. This dimension is related to Indonesian consumer protection law, which recognizes consumer rights to accurate information and legal remedies for losses, while the Personal Data Protection Law provides rights and remedial mechanisms where personal data processing causes harm. These regulations therefore provide a general legal foundation for consumer protection and redress. However, they do not comprehensively establish a platform-specific remedy mechanism for victims of AI-based *deepfakes*, including procedures for rapid content correction, restoration of impersonated identities, preservation of digital evidence, or compensation linked specifically to a platform's failure to exercise reasonable *due diligence*. The regulatory gap consequently concerns not the complete absence of remedies, but the absence of an integrated remedial framework specifically connecting *deepfake*-related consumer harm with platform responsibility.

Transparency and accountability are necessary to enable users and relevant authorities to understand how platforms conduct content moderation. Transparency can be achieved through accessible moderation policies, reports on moderation activities, and information regarding the use of automated systems in decision-making (Andreas et al., 2021). Existing Indonesian regulations impose general responsibilities on

Electronic System Operators and provide regulatory mechanisms for oversight and compliance. These provisions support transparency as part of responsible electronic-system governance, but they do not establish detailed *deepfake*-specific transparency obligations, such as disclosure of detection rates, labeling practices, algorithmic amplification of synthetic content, or the effectiveness of moderation measures. The existing framework therefore provides a general accountability structure without fully defining the information that platforms should disclose regarding their management of AI-generated manipulative content.

Evaluation constitutes the final component of the *due diligence* process and requires platforms to periodically assess whether prevention and mitigation mechanisms effectively reduce identified risks. UNESCO emphasizes that platform *due diligence* systems, risk assessments, and mitigation measures should be periodically reviewed and that the results should be made publicly available (Septian, 2024). Indonesian regulations provide general obligations concerning the reliability, security, and accountability of electronic systems, which may support continuing evaluation as part of responsible system governance. Nevertheless, they do not expressly require periodic audits of *deepfake*-related risks, publication of the results of such assessments, or measurement of the effectiveness of detection and mitigation mechanisms. This distinction demonstrates that international *due diligence* standards provide a more structured risk-management model, while Indonesian law currently supplies general and sectoral legal bases that have not yet been integrated into a comprehensive *deepfake*-specific framework.

The normative analysis therefore demonstrates that Indonesian law does not entirely lack a legal basis for imposing responsibilities on social media companies. Obligations concerning reliable and secure electronic systems, prohibited content, access termination, consumer protection, personal data protection, and remedies can collectively provide a general basis for platform responsibility. The principal regulatory weakness lies in the absence of specific and measurable standards that translate these general obligations into a complete *due diligence* framework for AI-based *deepfakes*. Risk assessment, prevention, detection, *notice-and-action*, *takedown*, complaint, remedy, transparency, and evaluation are therefore not wholly outside the existing legal framework, but their scope, threshold, and implementation remain fragmented or ambiguous. This fragmentation makes it difficult to determine consistently when a platform's failure to prevent or respond to harmful *deepfake* content constitutes a breach of legal obligation and consequently gives rise to legal liability.

Efforts to Strengthen Legal Protection for Consumers Against the Spread of AI-Based Deepfakes by Social Media Companies

Not every victim of a *deepfake* can be considered a consumer. Consumer status arises when a *deepfake* is related to commercial activities, promotions, or transactions involving goods and/or services, thereby influencing consumer decisions (Hasibuan et al., 2025). *Deepfakes* in advertisements may portray a person as if they are providing a testimonial or recommending a product, even though the information is false. This situation can mislead consumers and influence purchasing decisions (Runturambi & Arifin, 2025). Consumer harm resulting from *deepfakes* can take the form of economic loss from purchasing products based on false information, as well as health and safety risks if *deepfakes* are used to promote health products or medications with misleading claims (Singh et al., 2025). Thus, *deepfake* victims must be distinguished between victims in general and those who are consumers. A victim can be categorized as a consumer if there is a connection between the *deepfake* information, the decision to use or purchase goods and/or services, and the losses they incur (Putra, 2024).

Law No. 8 of 1999 on Consumer Protection (UUPK) serves as the primary basis for protecting consumers against misleading information in commercial activities. In the context of *deepfakes*, the use of manipulated content in advertisements or promotions directly relates to consumers' right to obtain accurate, clear, and

honest information because purchasing decisions may be based on false representations. The UUPK therefore provides a normative basis for protecting consumers from commercial *deepfakes*, although its provisions were not specifically designed to regulate AI-generated content. Law No. 1 of 2024 on Information and Electronic Transactions (ITE Law) provides another layer of protection when *deepfakes* are disseminated through electronic systems and contain information or content that violates the law. Its application, however, still depends on the nature and legal consequences of the disseminated content because the ITE Law does not specifically classify *deepfakes* as a distinct category of unlawful content. Law No. 27 of 2022 on Personal Data Protection (PDP Law) becomes relevant when the creation or dissemination of a *deepfake* involves a face, voice, biometric data, or other personal data capable of identifying an individual without a lawful basis for processing. The PDP Law therefore provides a normative basis for addressing the misuse of identity or personal data in *deepfake* production and, through its risk-based provisions, also requires impact assessment where personal data processing presents a high risk, including through the use of new technologies.

The obligations of social media companies can be further linked to Government Regulation No. 71 of 2019 on the Operation of Electronic Systems and Transactions. This regulation requires Electronic System Operators to operate their systems reliably, securely, and responsibly and to ensure that their systems do not facilitate the dissemination of prohibited electronic information or documents. In the context of *deepfakes*, this obligation provides a general normative basis for requiring platforms to adopt preventive and responsive measures when manipulated content falls within an existing category of prohibited information. This framework is reinforced for private Electronic System Operators by Permenkominfo No. 5 of 2020, as amended by Permenkominfo No. 10 of 2021, particularly in relation to prohibited electronic information and access termination. Accordingly, Indonesian law already provides a legal basis for requiring platforms to respond to unlawful content. The limitation is that these provisions do not specifically determine how platforms must identify *deepfake* risks, what technological detection standards must be applied, when synthetic content must be labeled, or when manipulated content that is harmful but not clearly prohibited must be restricted or removed.

Consumer protection against *deepfakes* in Indonesia is therefore multi-regulatory in nature. The Consumer Protection Law provides the basis for protecting consumers from misleading commercial information; the ITE Law governs relevant aspects of electronic information and its dissemination; the PDP Law protects personal data used in creating or processing *deepfakes*; and Government Regulation No. 71 of 2019, together with the regulatory framework for private Electronic System Operators, establishes obligations related to electronic system governance and prohibited content. These provisions demonstrate that Indonesia does not face a complete legal vacuum regarding *deepfake*-related harm. The normative problem instead lies in regulatory fragmentation: existing provisions can be applied according to the nature of the conduct and harm, but they have not been integrated into a specific framework defining *deepfakes*, determining the responsibilities of platforms, and establishing measurable *due diligence* standards. This fragmentation can create legal uncertainty because the distinctive characteristics of AI-based synthetic content have not yet been comprehensively accommodated within the existing framework (Ahmed et al., 2023).

Another weakness concerns the division of legal responsibility among creators, uploaders or distributors, technology developers, and platform companies. Existing regulations can more readily be applied to individual perpetrators whose conduct falls within an established legal prohibition, whereas the threshold for platform liability based on failure to prevent, detect, or respond to harmful *deepfakes* remains less clearly defined. The same ambiguity applies to *due diligence* obligations concerning risk assessment, detection, monitoring, synthetic-content labeling, responses to user reports, and restriction or removal of *deepfake*

content. Although the general obligations imposed on Electronic System Operators can serve as a normative basis for prevention and response, Indonesian regulations have not established uniform and measurable standards specifically for these mechanisms. Consequently, the regulatory issue is not the complete absence of platform obligations, but uncertainty regarding the scope and standard of care that platforms must satisfy when managing *deepfake*-related risks (Mukhlis et al., 2023).

Strengthening legal protection therefore requires clarification of both substantive norms and platform responsibilities. Indonesia does not yet have a specific regulatory definition and classification of *deepfakes*, meaning that enforcement continues to rely on the ITE Law, the PDP Law, the Consumer Protection Law, the Criminal Code, and other relevant provisions according to the nature of the act and the resulting harm (Dewi & Basuki, 2026). This condition should be understood as a specific regulatory gap rather than a complete absence of applicable law. Clearer regulation is particularly necessary to establish the legal position of AI-generated synthetic content, distinguish the responsibilities of creators, distributors, and platforms, and define when a platform's failure to exercise reasonable *due diligence* may give rise to legal liability. Such clarification would address the current lack of a clear liability model and measurable standards for platform obligations (Lumenta, 2026).

The *due diligence* framework also needs to be translated into more concrete platform obligations. Indonesian regulations have not yet specified in detail uniform requirements for *deepfake*-related risk assessment, detection, labeling, monitoring, *notice-and-takedown procedures*, transparency, and periodic evaluation (Wulandari et al., n.d.). Existing obligations concerning reliable and secure electronic systems, prohibited content, personal data protection, and access termination provide a general legal foundation, but they do not determine specific standards for detecting or labeling AI-generated content. The absence of uniform labeling and transparency requirements may make it difficult for users to distinguish authentic content from synthetic content, while the lack of standardized response and redress mechanisms may hinder victims in obtaining timely content removal, identity restoration, protection, or compensation. Strengthening these mechanisms would therefore complement rather than replace the existing legal framework.

The strengthening of consumer protection should ultimately integrate preventive, responsive, and remedial obligations within a coherent platform-accountability framework. Risk assessment and monitoring should enable platforms to identify foreseeable *deepfake* risks; detection and labeling should reduce consumers' exposure to misleading synthetic content; *notice-and-takedown* mechanisms should provide a clear procedure for responding to unlawful content; and complaint and remedy mechanisms should enable affected consumers to obtain effective redress. Transparency regarding moderation practices and periodic evaluation would further enable regulators and users to assess whether these measures function effectively. The principal weakness of Indonesian law therefore lies not only in the absence of a specific definition of *deepfake*, but also in the lack of an integrated system connecting platform *due diligence*, legal liability, and consumer remedies (Hermansyah, 2025). Strengthening this connection would provide greater legal certainty regarding what platforms are required to do, when failure to act may constitute a breach of legal obligation, and what forms of protection are available to consumers harmed by AI-based *deepfakes*.

The Prabowo Subianto *Deepfake* Case – Indonesia

A notable case in Indonesia involved the use of *deepfake* videos depicting President Prabowo Subianto and other state officials as if they were offering government financial assistance to the public. According to the Indonesian National Police, the perpetrators disseminated manipulated videos accompanied by captions and telephone numbers to attract individuals seeking financial assistance. Victims who contacted the listed numbers were subsequently asked to transfer money under the pretext of administrative fees for the disbursement of the promised aid. In one case, a perpetrator identified as AMA defrauded 11 victims, while

another perpetrator, identified as JS, was reported to have defrauded approximately 100 victims across 20 provinces and obtained approximately Rp65 million from the scheme. These facts were documented by the Indonesian National Police through its cybercrime investigation. The perpetrators were charged under Article 51 paragraph (1) in conjunction with Article 35 of Law No. 1 of 2024 concerning the Second Amendment to the ITE Law, as well as Article 378 of the Criminal Code concerning fraud.

The case demonstrates that existing Indonesian law provides a basis for prosecuting perpetrators when *deepfake* manipulation constitutes electronic information manipulation and is used as an instrument of fraud. However, the application of these provisions remains primarily perpetrator-oriented and does not itself establish a comprehensive standard governing the preventive responsibility of social media platforms through which the fraudulent content is disseminated. The Indonesian National Police also reported coordination with the Ministry of Communication and Digital Affairs to block and remove accounts disseminating such content, indicating that *takedown* can function as a responsive measure after unlawful *deepfake* content has been identified. The limitation is that such intervention largely operates after dissemination has occurred, whereas the *due diligence* principle requires attention to preventive measures before consumer harm materializes.

From the perspective of platform *due diligence*, the case therefore illustrates the relationship between detection, reporting, distribution restriction, and *takedown* obligations. Existing obligations imposed on Electronic System Operators provide a general legal basis for responding to prohibited electronic information, but they do not specifically determine the technological standards that social media companies must employ to detect fraudulent *deepfakes* before they reach consumers. Nor do they establish uniform requirements for synthetic-content labeling, early warnings, or risk-based restrictions on the distribution of suspected *deepfake* content. Consequently, strengthening consumer protection requires not only prosecution of individual perpetrators but also clearer standards requiring platforms to implement reasonable detection mechanisms, provide rapid reporting channels, restrict the distribution of identified fraudulent content, and promptly remove unlawful *deepfakes* (Nurkholisah et al., 2025). The Prabowo case thus demonstrates that the principal regulatory gap concerns the transition from reactive content removal to measurable preventive *due diligence* obligations for digital platforms.

Deepfake Fraud Case Involving a Finance Director in Singapore

A comparable case occurred in Singapore in March 2025 and involved a finance director of a multinational corporation who was targeted through a sophisticated business impersonation scheme using *deepfake* technology. According to the Singapore Police Force (SPF), the victim was initially contacted through WhatsApp by a scammer impersonating the company's Chief Financial Officer and was subsequently invited to participate in a Zoom video conference concerning a purported restructuring of the company's regional business. During the video conference, the company's Chief Executive Officer and other participants were impersonated using *deepfake* technology. Following the conference and subsequent communication with an individual impersonating a lawyer, the victim was instructed to transfer more than US$499,000 from the company's HSBC account to a local corporate account controlled by the scammers. More than US$494,000 was subsequently transferred from that account to bank accounts in Hong Kong.

The victim became aware of the fraud on March 27, 2025, when the scammers requested an additional transfer of approximately US$1.4 million. The victim immediately contacted HSBC, which notified the Singapore Police Force's Anti-Scam Centre (ASC). The ASC traced the proceeds and coordinated with the Hong Kong Police Force's Anti-Deception Coordination Centre (ADCC). Through this coordination, the authorities successfully withheld the full amount transferred to the Hong Kong accounts, while the remaining amount in the Singapore mule account was also seized. The official account of the case therefore

demonstrates that rapid reporting, financial tracing, institutional coordination, and cross-border cooperation can function as effective remedial mechanisms after a *deepfake*-enabled fraud has occurred.

The Singapore case also illustrates the preventive dimension of *due diligence*. In a joint advisory, the Singapore Police Force, the Monetary Authority of Singapore, and the Cyber Security Agency of Singapore warned businesses about scams involving digital manipulation and recommended verification of unusual requests through established communication channels. The authorities specifically emphasized verification of unsolicited communications and urgent fund-transfer instructions. These preventive measures demonstrate that *deepfake*-related protection does not depend solely on detecting manipulated audiovisual content; it can also involve procedural safeguards capable of interrupting fraudulent transactions even when the synthetic content itself is difficult to identify.

From the perspective of this study, the Singapore case provides a comparative illustration rather than a direct legal basis for imposing obligations on Indonesian platforms. Its relevance lies in demonstrating how preventive verification, rapid reporting, financial intervention, and coordinated recovery mechanisms can complement content detection and removal. In the Indonesian context, these measures may inform the development of more responsive consumer-protection mechanisms, but their adoption must remain grounded in Indonesian legal obligations applicable to Electronic System Operators, consumer protection, electronic information, and personal data. The comparison therefore reinforces the need for an integrated *due diligence* framework in which prevention, detection, reporting, *takedown*, and victim remedies operate as interconnected obligations rather than isolated responses after consumer losses have already occurred.

Consumer/Victim Redress Mechanisms

Consumer and victim redress mechanisms for AI-based *deepfakes* need to be distinguished between remedies that already have a legal basis under Indonesian positive law (*de lege lata*) and mechanisms that constitute proposals for strengthening future regulation (*de lege ferenda*). This distinction is important because not every form of redress discussed in the context of digital-platform governance has been expressly established as a specific legal obligation for social media companies in Indonesia.

First, halting the dissemination and removing unlawful *deepfake* content can be linked to existing Indonesian rules concerning prohibited electronic information and access termination. Government Regulation No. 71 of 2019 on the Operation of Electronic Systems and Transactions and the regulatory framework governing private Electronic System Operators provide a legal basis for restricting or terminating access to electronic information or documents that fall within legally prohibited categories. Accordingly, *takedown* is not entirely a proposed mechanism because access termination for unlawful electronic content already has a basis in Indonesian positive law. However, a specific *Fast-Track Removal* procedure for *deepfake* content has not been expressly established as a uniform legal obligation applicable to platforms. Fadhila's proposal to strengthen *takedown* mechanisms through a *Fast-Track Removal* process should therefore be understood as a *de lege ferenda* recommendation intended to accelerate the removal of harmful *deepfake* content, particularly where continued circulation prolongs the harm suffered by victims.

Second, victim redress may involve removing or restricting access to content that unlawfully uses a victim's identity, particularly where a *deepfake* exploits a person's face, voice, or other personal data. This form of redress has a positive legal basis insofar as Law No. 27 of 2022 on Personal Data Protection recognizes data-subject rights concerning the processing of personal data, including the termination of processing, deletion, and/or destruction of personal data under the conditions prescribed by law. Thus, the substance of the *right to erasure* is already recognized within Indonesia's personal data protection framework. Nevertheless, Indonesian law does not establish a separate *deepfake*-specific *right to erasure* that automatically requires social media platforms to remove every form of synthetic content upon a victim's

request. Fadhila's identification of the *Right to Erasure* as a technical remedy should therefore be understood as relying on existing personal-data rights while proposing their more explicit and effective application to *deepfake*-related harm.

Third, in the context of *deepfakes* used for advertising or promotion, consumer redress through the correction of misleading information is connected to rights and obligations already recognized under Law No. 8 of 1999 on Consumer Protection. The UUPK recognizes consumers' right to obtain accurate, clear, and honest information concerning goods and/or services and prohibits business actors from providing misleading information in commercial activities. Nababan points out that the use of *deepfakes* in social media advertising can mislead consumers and potentially cause harm; therefore, legal action against misleading information or advertisements is consistent with existing consumer-protection principles. However, a specific obligation requiring social media platforms to issue corrections for every misleading *deepfake* advertisement has not been comprehensively regulated. Accordingly, the consumer's entitlement to truthful commercial information is part of *de lege lata*, whereas a standardized platform-based correction mechanism for *deepfake* advertising constitutes an area for *de lege ferenda* regulatory strengthening.

Fourth, compensation has a positive legal basis where consumers suffer actual losses within a consumer relationship. Law No. 8 of 1999 on Consumer Protection provides a basis for business actors' responsibility to compensate consumers for losses arising from goods and/or services, subject to the conditions and allocation of responsibility established by law. Accordingly, economic losses resulting from consumers purchasing products based on false or misleading *deepfake* promotions may fall within existing consumer-redress mechanisms where the legal requirements for liability are satisfied. Research on platform liability for consumer harm further indicates that digital platforms cannot always be viewed as passive parties when system vulnerabilities or failures in verification mechanisms contribute to consumer harm. Nevertheless, Indonesian positive law does not yet establish an automatic right to compensation specifically against social media platforms merely because harmful *deepfake* content has circulated through their services. Platform compensation liability therefore requires a demonstrable legal basis connecting the platform's obligation, breach or failure to act, the resulting harm, and the relevant liability regime. A specific compensation regime based on a platform's failure to exercise *deepfake due diligence* would consequently require further regulatory clarification.

Fifth, victims need access to an accessible complaint mechanism through which they can report unlawful *deepfake* content and request action from the platform. Existing regulations governing private Electronic System Operators provide a normative basis for handling prohibited electronic information and responding to notifications concerning unlawful content. Putra's research indicates that the platform liability system in Indonesia has led to an obligation to remove illegal content upon notification; however, greater transparency and oversight mechanisms are still needed to ensure that this obligation is effectively enforced. Therefore, complaint and notification mechanisms relating to unlawful electronic content have a basis in the existing regulatory framework. What remains insufficiently regulated is a uniform *deepfake*-specific complaint procedure establishing clear response periods, notification requirements, transparency standards, victim follow-up procedures, and accountability where platforms fail to respond. These additional procedural standards should therefore be understood as *de lege ferenda* measures for strengthening the effectiveness of existing complaint mechanisms.

Sixth, where complaints submitted to platforms do not resolve consumer harm, consumers may use dispute-resolution mechanisms recognized under Indonesian law, including litigation and available alternative dispute-resolution mechanisms such as mediation or arbitration in accordance with the applicable legal framework. These mechanisms form part of the existing consumer-protection and dispute-resolution framework and therefore constitute *de lege lata*. *Online Dispute Resolution* (ODR), however,

should be distinguished from these established mechanisms insofar as a comprehensive and uniform ODR framework specifically designed for disputes involving *deepfakes* and social media platform liability has not yet been established. Hidayati and Suartini found that the growth of digital transactions has increased the need for ODR because online dispute-resolution mechanisms can provide consumers with easier access to dispute settlement, although issues concerning transparency, security, and fairness remain. Accordingly, the use of ODR as a specialized and integrated mechanism for *deepfake*-related consumer disputes should be regarded primarily as a *de lege ferenda* proposal for strengthening access to redress rather than as an already applicable platform obligation.

The analysis demonstrates that Indonesian positive law already provides several foundations for consumer and victim redress, particularly through access termination for unlawful electronic content, personal-data rights, protection against misleading commercial information, compensation under consumer-protection law, complaint mechanisms, and existing dispute-resolution procedures. The principal regulatory gap lies in the absence of an integrated *deepfake*-specific redress framework that connects these existing rights and obligations with the responsibilities of social media platforms. Therefore, *Fast-Track Removal*, standardized *deepfake*-specific correction and complaint procedures, specific platform compensation standards based on failures of *due diligence*, and an integrated ODR mechanism should be understood as *de lege ferenda* measures. This distinction prevents proposed regulatory strengthening from being interpreted as an obligation already imposed by Indonesian positive law.

4. Conclusion

The principle of *due diligence* can serve as a basis for assessing the legal responsibility of social media companies for the dissemination of AI-based *deepfakes* because Indonesian law already provides a general legal foundation for platform obligations through regulations governing Electronic System Operators, electronic information and transactions, consumer protection, and personal data protection. These regulations establish obligations related to reliable and secure electronic systems, the handling of prohibited content, access termination, consumer rights, and personal data protection. However, the existing legal framework has not yet established specific and measurable *due diligence* standards for AI-based *deepfakes*, particularly regarding risk assessment, prevention, detection, monitoring, synthetic-content labeling, *notice-and-action*, *takedown*, transparency, and periodic evaluation. Therefore, platform responsibility is not based on a complete absence of applicable law, but on general and sectoral obligations whose application to *deepfake*-related risks remains fragmented and, in certain aspects, ambiguous. This condition makes it necessary to clarify the scope and standard of care required from social media companies in determining when a failure to prevent or respond to harmful *deepfake* content may give rise to legal liability.

Strengthening consumer protection against AI-based *deepfakes* therefore requires the integration of preventive, responsive, and remedial measures within a clearer platform-accountability framework. The mechanism for redressing consumers or victims cannot be limited to simply removing the content but must be carried out in a sustained and comprehensive manner. Remedies can range from halting dissemination and *taking down* or restricting access to unlawful content, correcting misleading information, restoring the victim's identity and reputation, providing accessible and rapid complaint mechanisms, to awarding compensation where legally provable losses and the requirements for liability are established. If resolution through the platform does not provide adequate redress, consumers or victims need access to applicable dispute-resolution mechanisms, including mediation and arbitration, while an integrated *Online Dispute Resolution* (ODR) mechanism can be developed as part of future regulatory strengthening. Accordingly, strengthening social media companies' *due diligence* should connect prevention and content governance

with effective consumer remedies, while clearly distinguishing obligations already recognized under Indonesian positive law from *de lege ferenda* measures that require further regulatory development.

5. References

Abbas, F., & Taeihagh, A. (2024). Unmasking Deepfakes: A Systematic Review of Deepfake Detection and Generation Techniques Using Artificial Intelligence. *Expert Systems with Applications*, *252*, 124260. https://doi.org/10.1016/j.eswa.2024.124260

Ahmed, S., Ng, S. W. T., & Bee, A. W. T. (2023). Understanding The Role of Fear of Missing Out and Deficient Self-Regulation in Sharing of Deepfakes on Social Media: Evidence from Eight Countries. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1127507

Andreas, C., Priandi, S., Simamora, A. N. M. B., & Mardianto, M. F. F. (2021). Analisis Hubungan Media Sosial dan Media Massa dalam Penyebaran Berita Hoaks berdasarkan Structural Equation Modeling-Partial Least Square. *MUST: Journal of Mathematics Education, Science and Technology*, *6*(1), 81. https://doi.org/10.30651/must.v6i1.8816

de Azevedo Cunha, M. V., & Monteleone, S. (2021). Data Protection, Freedom of Expression, Competition and Media Pluralism: Challenges in Balancing and Safeguarding Rights in The Age of Big Data. In *Research Handbook on EU Media Law and Policy*. Edward Elgar Publishing. https://doi.org/10.4337/9781786439338.00018

Dewi, R. S., & Basuki, A. (2026). Relevansi Kepentingan Hukum Pelapor dengan Dakwaan dalam Perkara Deepfake. *Al-Zayn: Jurnal Ilmu Sosial & Hukum*, *4*(2), 1738–1747. https://doi.org/10.61104/ALZ.V4I2.4554

Hasibuan, A. H. H., Setyawan, D. N., & Rasyid, Y. A. (2025). Legal Protection For Consumers Against The Potential Harm Caused By Promotions Using Deepfakes. *JIHK*, *6*(2), 116–126. https://doi.org/10.46924/jihk.v6i2.232

Hermansyah, M. G. R. N. (2025). Penggunaan Artificial Intelligence dan Deep Fake Analisis Kasus Pidana Video Deepfake yang Mencatut Presiden Prabowo Subianto Tahun 2025. *JIC: Jurnal Hukum Dan Konstitusi*, *1*(3), 132–146. https://doi.org/10.64272/yqyf4m65

Heskia Putri, Y., & Astuti, P. (2026). Penegakan Hukum Deepfake di Polda Jawa Timur Sebagai Bentuk Kejahatan Siber. *Indonesian Journal of Contemporary Law*, *3*(01), 1–21. https://journal.unesa.ac.id/index.php/ijcl/article/view/50868

Irwansyah. (2020). *Penelitian Hukum; Pilihan Metode dan Praktik Penulisan Arikel*. Mitra Buana Media.

Karpinska-Krakowiak, M. (Mags), & Eisend, M. (2025). Realistic Portrayals of Untrue Information: The Effects of Deepfaked Ads and Different Types of Disclosures. *Journal of Advertising*, *54*(3), 432–442. https://doi.org/10.1080/00913367.2024.2306415

Kristiyenda, Y. S., Faradila, J., & Basanova, C. (2025). Pencegahan Kejahatan Deepfake: Studi Kasus terhadap Modus Penipuan Deepfake Prabowo Subianto dalam Tawaran Bantuan Uang. *Aladalah: Jurnal Politik, Sosial, Hukum Dan Humaniora*, *3*(2), 149–164. https://doi.org/10.59246/aladalah.v3i2.1263

Lumenta, H. N. (2026). Tanggung Jawab Hukum Platform Media Sosial Atas Penyebaran Konten Hoaks di Indonesia. *Lex Technologia: Jurnal Hukum Dan Masyarakat Digital*, *1*(1), 1–10. https://ejournalfish.unima.ac.id/index.php/lextechnologia/id/article/view/9

Mukhlis, Soekorini, N., Taufik, Moh., & Aranggraeni, R. (2023). Ilusi Realitas: Tinjauan Hukum Pidana Indonesia terhadap Penipuan Daring Berbasis Deepfake AI di Media Sosial. *Lex Journal: Kajian Hukum Dan Keadilan*, *7*(2), 585–618. https://doi.org/10.25139/lex.v7i2.12605

Nurkholisah, S., Rismana, D., Nugroho, A. E., Munjiyah, A., & Ayunisa, Q. (2025). Deepfake Sebagai Bentuk Kejahatan Siber Baru: Tantangan Kriminalisasi dalam Hukum Pidana Indonesia. *Jurnal USM Law Review*, *8*(3), 2421–2445. https://doi.org/10.26623/julr.v8i3.13060

Putra, G. R. A. (2024). Tanggung Jawab Hukum Platform Media Sosial terhadap Konten Pengguna di Indonesia: Potret Conditional Safe Harbor, Sistem SAMAN, dan Tarikan Menuju Akuntabilitas Proaktif. *Buletin Adalah*, *8*(7), 63–84. https://doi.org/10.15408/ADALAH.V8I7.51883

Putri, P. N. T., & Gorda, A. A. A. N. T. R. (2025). Pengaturan Hukum dalam Penanggulangan Deepfake Artificial Intelligence (AI) terhadap Anak sebagai Korban Kejahatan Siber di Indonesia. *Jurnal Magister Hukum Udayana (Udayana Master Law Journal)*, *14*(3), 713–728. https://doi.org/10.24843/JMHU.2025.v14.i03.p09

Runturambi, A. J. S., & Arifin, R. (2025). New Patterns and Trends of Migration: Hybrid-Crimes among Indonesian Migrant Workers in Southeast Asia. *Regional Science Policy & Practice*, *17*(10), 100215. https://doi.org/10.1016/j.rspp.2025.100215

Septian, G. D. (2024). Systematic Literature-Review: Analisis Komparasi Media Sosial dalam Penyebaran Informasi. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *8*(3), 771–780. https://doi.org/10.35870/jtik.v8i3.2353

Shabrina, A. N., Naila, G. S., Nuryansyah, G. P., & Amanda, R. (2026). Bahaya Deepfake dalam Penyalahgunaan Grok AI di Platform X Sebagai Bentuk Disinformasi. *Integrative Perspectives of Social and Science Journal*, *3*(01 Januari), 195–207. https://ipssj.com/index.php/ojs/article/view/1132

Singh, A., Kaur, I., Upadhyay, A. K., & Trivedi, M. (2025). *Safeguarding Free Speech While Tackling Misinformation on Social Media Platforms* (pp. 4–21). https://doi.org/10.2991/978-2-38476-426-6_2

Wulandari, S. B., Setiyawati, A., Yulanda, M., Rizky, M., Saputra, P., & Pambudi, S. B. (2025). Menakar Intermediary Liability: Tantangan Platform Digital dalam Mengawasi Deepfake Sexual Content. *Konstitusi: Jurnal Studi Hukum*, *2*(1), 1–12. https://glonus.org/index.php/konstitusi/article/view/294

Yanova, M. H., Komarudin, P., & Hadi, H. (2023). Metode Penelitian Hukum: Analisis Problematika Hukum dengan Metode Penelitian Normatif dan Empiris. *Badamai Law Journal*, *8*(2), 394–408. https://doi.org/10.32801/DAMAI.V8I2.17423