Proceedings of the 4th International Conference on International Law and Legal Policy
DOI: 10.54254/2753-7048/2025.28977

Reimagining Safe Harbors: A Systematic Review of Liability Frameworks for Generative AI

Bingjie Xiao

School of Law, Hunan University, Changsha, China
xiaobingjie@hnu.edu.cn

Abstract. The rapid growth of generative AI (GAI) has transformed how artistic works are created and raised new challenges for copyright law. The traditional safe harbor regime is under strain. Due to widespread and instant outputs by GAI, the old notice-and-takedown approach no longer works. This paper reviews the main academic views on the liability of GAI platforms and the scope of safe harbor protection. It identifies three main positions: complete rejection of immunity, conditional immunity with new duties, and structural reform of liability rules. It explains the reasons and limits of each view. Based on this review, the paper proposes a "duty-based exemption" model: a limited immunity system that matches the platform's level of control, the predictability of risk, and the benefits received. This model aims to balance copyright protection with technological innovation.

Keywords: generative artificial intelligence, safe harbor, copyright infringement, platform liability, liability exemption

1. Introduction

Since late 2022 and the release of tools like ChatGPT, generative AI (GAI) systems have been able to create text, images, audio, and video in seconds from user prompts and large training datasets. As a result, these outputs appear to be original works. However, although GAI could boost efficiency and innovation in creating content, it also raises questions about authorship, ownership, and liability. For example, several copyright lawsuits have been filed against companies such as OpenAI and Microsoft for allegedly using protected works in training and for producing infringing content [1,2]. At the same time, legislators worldwide have begun responding. In 2023, the U.S. Copyright Office held a public inquiry on GAI. The European Union passed the Digital Services Act and the AI Act with new duties on AI platforms [3,4]. China issued interim measures treating GAI providers as content creators and imposing broad obligations on them.

The safe harbor regime, long seen as a key tool to balance innovation and rights, is now under strain. Generative AI changes platforms from passive intermediaries into active creators, challenging the basis of safe harbor protection. In response, scholars have proposed three main approaches [5-11]. One view denies immunity, treating GAI platforms as content producers and fully liable. Another suggests modifying safe harbor: granting conditional immunity if platforms meet new duties. A third calls for a new framework, reallocating liability based on the platform's role or risk level. These studies highlight the issues but often stop at debate or case analysis. Significant gaps

DOI: 10.54254/2753-7048/2025.28977

remain: there are no clear standards for key concepts such as “substantial control”, the allocation of responsibility among multiple actors is unclear, and procedural mechanisms for moving from simple removal to corrective revision are missing. These gaps lead to inconsistent court decisions, uncertain compliance for companies, and difficulties for rights holders. This paper thus provides a systematic literature review. It summarizes and evaluates the main arguments and practices, notes their shortcomings and contradictions, and aims to guide the development of a copyright liability framework fit for the GAI era.

2. Mainstream consensus

2.1. The evolution of the safe harbor framework

The “Safe harbors” refer to legal immunity granted to certain actors for specified conduct under defined conditions. In the internet context, it chiefly protects intermediary service providers such as access and hosting services. When these intermediaries satisfy statutory requirements, they are not liable for unlawful user content, including copyright infringement and defamation [12]. The design reflects a legislative attempt to balance two policies: promoting the internet economy and protecting the interests of rights holders.

The framework first appeared in the United States with Section 230 of the Communications Decency Act (1996), which granted publisher immunity for user speech but not for intellectual property, and allowed good-faith removal of harmful material without losing protection. The 1998 Digital Millennium Copyright Act (DMCA) extended safe harbor to copyright. Section 512 created four categories of immunity: transitory communications, caching, user-directed storage, and information location tools, and exempted providers lacking actual knowledge, receiving no direct financial benefit, and removing infringing material after notice [13]. The DMCA aimed to curb infringement without over-deterring platforms, enabling services like YouTube and Google. The European Union followed with the 2000 E-Commerce Directive (Directive 2000/31/EC) [14], which limited liability for “mere conduit”, “caching”, and “hosting”, and prohibited a general monitoring duty. China introduced similar rules in the 2006 Regulations on the Protection of the Right of Communication through Information Networks [15], covering automatic transmission, caching, storage, and linking, and establishing notice-and-takedown. A key difference is that the DMCA removes protection only when both control and direct financial benefit exist, while the Chinese Regulations deny immunity based on direct benefit alone.

With respect to covered entities, most jurisdictions grant near absolute safe harbor protection to basic technical services (such as network access and automatic transmission), but add more conditions for content hosting platforms (such as video sharing sites and social media) and for information location tools (such as search engines), and keep adjusting the scope as technology evolves. In the United States, courts refined the rules through cases such as Viacom v. YouTube, distinguishing general awareness from specific knowledge of infringement, and narrowly interpreting “control” and “direct financial benefit”, so that only active participation removes protection. In the European Union, the Court of Justice developed the “active role” theory, as in L’Oréal v. eBay, holding that selecting or optimizing content may forfeit immunity. The 2019 Copyright Directive (Article 17) further imposed preventive obligations on large platforms, marking a “post–safe harbor” era where claims of neutrality are no longer enough [1]. In China, courts follow the 2006 Regulations on the Protection of the Right of Communication through Information Networks but refine boundaries through cases [15]. In IFPI v. Yahoo (Alibaba), failure to remove all infringing links after notice was deemed a breach [15]. Courts also developed the “red flag”

DOI: 10.54254/2753-7048/2025.28977

standard, presuming knowledge where infringement is obvious, even without notice. The "no direct profit" rule was later clarified by judicial interpretation. In recent years, Chinese courts have broadened the safe harbor standard, requiring platforms to act reasonably after valid notice, extend removal to clearly infringing content, and apply the "red flag" standard to presume knowledge in obvious cases, aiming to balance copyright protection with industrial development.

2.2. Core debates in current scholarship

The rapid growth of generative artificial intelligence (GAI) has reopened the question of whether its providers should receive safe harbor protection like traditional intermediaries. Unlike early services that only stored or transmitted user material, GAI produces text, images, and other outputs from prompts. Scholars widely agree that the notice-and-takedown model shaped by Web 1.0 and 2.0 does not fit the scale and unpredictability of such outputs [5,6].

From this premise, three positions have formed. The first rejects safe harbor altogether: GAI is viewed as an active creator that should bear direct liability, as argued by Pérez [5]. Luk [6] also criticizes the safe harbor provisions, claiming that the law should not simply adjust to technological change but should instead impose clear liability on GAI providers. The second seeks conditional reform: safe harbor may remain but only with added duties. Zou and Zhang [7] accept that GAI providers might be treated as "content producers", yet they would keep immunity if they filter and cooperate with regulators. Zhang [8] proposes a "notice-and-action" system that requires complaint channels and post-notice model revision, while Revolidis [3] analogizes some GAI functions to hosting or search, bringing transparency, risk checks, and notice duties under the EU Digital Services Act into play. The third calls for structural redesign. Lin and Guan [9] propose "AI Harbors" that allocate duties across the supply chain and grant immunity only after those duties are met. Choi [10] argues that vicarious liability should replace safe harbor: when a platform controls outputs and profits from them, it should be held liable. This, in turn, would push platforms to adopt preventive measures like filtering in order to avoid liability. Lim [11] advances a risk-based mix in which high-risk uses require licensing, while low-risk uses follow a lighter notice system. Taken together, these approaches differ on whether to deny immunity, to condition it on new obligations, or to rebuild the regime by role and risk. The core issue is whether GAI's features have moved platforms beyond passive intermediation and whether the law should adapt old rules or design a new governance framework.

3. The decline of the traditional safe harbor framework

The traditional safe harbor regime, exemplified by DMCA §512, rests on three pillars: notice and takedown, protection for passive intermediaries, and technological neutrality [12,13]. It was meant to shield online services to promote industry growth while protecting copyright. With the rise of generative AI (GAI), which shifts content production from distribution to creation, these foundations face serious strain.

From a theoretical perspective, safe harbor assumes platforms are passive and neutral. Pérez [5] and Mohan and Jansi [16] argue that GAI, relying on complex models and large training sets, acts as a "co-creator" rather than an intermediary. Pérez [5] add that control through model design, data selection, and fine-tuning further undermines neutrality. Statutorily, Henderson et al. [17] note that the DMCA covers content "stored at the direction of the user", while GAI outputs stem from prompts and model parameters, placing them outside safe harbor. U.S. courts echoed this in Williams-Sonoma v. Amazon, where liability arose from algorithmic selection. By contrast,

DOI: 10.54254/2753-7048/2025.28977

Revolidis [3] cites YouTube and Cyando to argue that new functions do not erase intermediary status. Yet this overlooks platforms' active role in data collection and parameter design, making the "passive intermediary" assumption hard to sustain for GAI. Technically, the notice-and-takedown model faces new limits. Outputs are generated in real time, often not stored, which makes infringement difficult to detect or remove. In Doe v. GitHub, the court held that using copyrighted data for training is not itself infringement unless outputs are substantially similar to protected works. Choi [10] observes that GAI creates infringement on a massive and hidden scale, as models can reproduce or "remember" training data, increasing monitoring costs. Lim [11] further emphasizes that removing copyrighted content from a trained model is nearly impossible, and current tools to detect infringement are limited.

On responsibility allocation, Lee et al. [18] highlight that the GAI supply chain involves multiple actors—data collectors, model trainers, deployers, and users—while safe harbor was built for a single service provider. Choi [10] argues that platforms profit without assuming risk, while Luk [6] stresses that notice and takedown shifts the burden onto copyright holders, undermining fairness and proportionality. Thus, the two-party model of "user uploads, platform stores" collapses when liability spreads across multiple actors. Legislative practice reflects this shift. The EU's Digital Services Act maintains exemptions but adds risk assessment and content governance duties, while the AI Act requires the disclosure of copyrighted materials used in training. Both impose proactive governance [3,4]. In China, the 2023 Interim Measures for Generative AI Services classify providers as "content producers" and impose full liability, discarding the intermediary premise [7]. Overall, the trend in multiple jurisdictions shows that the traditional safe harbor is giving way to preventive, transparent, and active duties, reflecting a move from passive immunity to proactive governance in the GAI era.

4. Reconstruction of safe harbor rules for generative AI

Confronting the impact of generative artificial intelligence (GAI) on safe harbor rules, current scholarship follows three paths that differ on whether GAI remains an intermediary, whether outputs can be controlled by traditional tools, and how to regulate training and generation.

The rejection path argues that general-purpose generative AI should not benefit from safe harbor. Pérez [5] regard such systems as "originators" of content and call for systemic risk assessments and trusted reporter mechanisms under Article 34 of the DSA. Mohan and Jansi [16] also contend that GAI is not an "intermediary" and propose a "Digital India Act" to define liability rather than grant exemption. This view leaves no space for immunity, but it treats all GAI as homogeneous and overlooks the varied levels of initiative and control across different systems and functions in the supply chain. Demanding that platforms be responsible for all outputs "spontaneously generated" by models amounts to strict liability, which is technically unrealistic and would impose prohibitive compliance and litigation costs, stifling startups and open-source projects.

The modification path retains safe harbor in principle but ties it to duties. Rosati [1], drawing on EU case law such as YouTube and Ziggo, argues that GAI providers may enjoy protection unless they assume an "active role", such as knowingly allowing infringing outputs without preventive measures like filters. The challenge lies in extending the "hosting provider" analogy to AI platforms and in defining "active role" when model operations are opaque and intertwined with micro-controls like prompt engineering or preset parameters. Lemoine and Vermeulen [14] oppose excluding GAI from the DSA altogether and propose treating it contextually, sometimes like hosting or search services.

DOI: 10.54254/2753-7048/2025.28977

Zou and Zhang [7], noting that China's Interim Measures already define AI providers as "content information producers", argue that the law should also allow them to obtain an exemption if they fulfill specific duties. Rosati [1], by contrast, starts from immunity in principle but removes it when providers take an active role. Though their directions differ, both approaches resist blanket rules. Zhang [8] advances an "input out, output in" model: non-expressive training use is beyond copyright control, while outputs should be subject to conditional safe harbor duties similar to the DMCA. Providers must maintain effective notice systems and fine-tune models after notice to prevent recurrence, with "camouflaged copying" as a ground for losing immunity.

Yet problems remain. U.S. precedent accepts large-scale scanning as transformative use, but Samuelson [2] warns that this justification weakens for high-value licensed content such as news or photo libraries. The EU Copyright Directive also allows rights holders to retain consent rights for text and data mining. Current lawsuits by The New York Times and Getty Images against OpenAI and Stability AI hinge on whether models "remembered" copyrighted material and reproduced outputs too similar to originals, undermining markets. Technical issues persist as well: reliable "machine unlearning" is still difficult, and Heng and Soh [19] show that forgetting cannot easily be achieved without harming performance. Overall, the modification path seeks to adapt the safe harbor by combining notice-and-revision duties with due diligence.

The reconstruction path seeks to rebuild liability and immunity around roles and risk levels. Lin and Guan [9] propose an "AI harbor" model with differentiated duties: data providers ensure transparency and legality, developers remove memorized content and use watermarks, and deployers filter outputs and address repeat infringement. They also suggest third-party audits and certification, though this requires strong regulatory capacity and risks raising compliance barriers that favor large platforms. Lim [11] proposes a "notice plus license" hybrid model: high-risk uses require prior licensing, while low-risk uses follow a notice-and-revision system. He criticizes the EU's licensing regime as too burdensome for SMEs. His approach complements Zhang's [8] output-end governance but faces questions of how to judge whether revisions are adequate and who has the authority to decide compliance. Choi [10] instead calls for a full reconstruction using vicarious liability: when platforms both control outputs and profit, they should be liable. He also recommends integrating the EU DSA's risk-based obligations into compliance, creating a hybrid "U.S. standards plus EU governance" framework.

In sum, the modification path, represented by Rosati [1], links immunity to an active role and due diligence, moving from notice-and-takedown to notice-and-revision. The reconstruction path redistributes duties by function or risk or substitutes vicarious liability for safe harbor. The rejection path denies exemption entirely. While these approaches differ in reasoning, they share one conclusion: unconditional safe harbor is no longer viable, and future liability must be allocated in a graded system aligned with control, foreseeability, and benefit.

5. Discussion

A review of current scholarship shows that the central issue is whether, once generative output has broken the premise of user uploads and passive storage, the law should still grant an exemption or instead reassign responsibility. Three approaches emerge. The first views platforms as "content producers", arguing that their role in generation has changed their identity and exemption should be denied or tightly limited [5,6,16]. This makes liability clear but risks stifling innovation. The second accepts greater involvement but still sees platforms as intermediaries, supporting limited exemption with reforms such as adapting notice-and-takedown to the features of generative AI [1,4,7,8]. The third proposes reconstructing responsibility by role and risk, rejecting blanket exemption or single-

DOI: 10.54254/2753-7048/2025.28977

actor liability and suggesting differentiated duties along the supply chain with conditional exemptions or hybrid mechanisms like "notice plus license" [9-11]. While all three acknowledge changing platform roles, they differ in logic: one stresses "generation equals responsibility", the second emphasizes "processes can be controlled", and the third highlights "responsibility should match risk".

These divisions rest on two key questions. First, legal classification: should generative AI platforms be treated as intermediaries or as content creators? If the latter, liability rises sharply and exemption may vanish; if the former, exemption can remain but only under stricter duties. Second, enforcement: should governance rely mainly on ex post remedies or shift toward ex ante prevention? The former stresses improved notice, timelines, and appeals; the latter requires technical tools such as filtering, labeling, data review, and output correction. The issue is whether platform involvement justifies removing the exemption entirely or attaching conditions to keep it. This paper argues that the more useful question is under what conditions "limited exemption" should be granted. Liability should depend on technical control, foreseeability of risks, and benefits gained. Actors with greater control, clearer foreseeability, and higher benefit should bear heavier duties. Only if obligations matching their role are fulfilled and compliance shown should exemption be allowed. Safe harbor should shift from identity-based immunity to duty-based exemption.

Responsibility can be reassessed through three perspectives. First, role differentiation: data providers, developers, deployers, and distributors have different levels of control, so duties and exemptions should differ accordingly. Second, governance sequence: notice-and-takedown cannot address dynamic output. Ex ante, platforms should plan compliance and keep technical records; during operation, they should establish a full "notice–response–justification–appeal" process open to external review, as in the DSA; ex post, they should include correction and restoration beyond removal. Copyright law provides useful references, such as counter-notice procedures and Articles 17 and 20 of the DSM Directive. Third, risk levels: high-risk uses should face stricter assessment, audit, and appeal duties, while medium- and low-risk uses may follow lighter requirements like labeling and transparency. This layered system manages risks while preserving innovation.

Still, gaps remain. First, no clear standard defines "active role" or "substantial control", leading to inconsistent rulings and uncertain compliance. Second, multi-actor responsibility lacks coherent rules on who acts first, who bears proof, and who carries ultimate liability. Third, shifting from content removal to model correction lacks standards for notice validity, response timelines, adequacy of revision, and recurrence. Without these, duty-based exemption cannot be verified, leaving rights holders without remedies and platforms unable to prove compliance. Future research should focus on three areas. First, designing a conditional safe harbor, where the exemption depends on duties proportionate to control and risk, is reduced if unmet. Second, clarifying supply chain roles: data providers must ensure legality, developers must govern training data and constrain models, deployers must filter outputs and handle complaints. Allocation should follow principles such as the lowest-cost risk preventer acting first, information holders bearing proof, and main beneficiaries bearing liability. Third, creating minimum procedural standards across systems, covering notice elements, timelines, revision criteria, and appeals, ensuring external verification and judicial applicability.

In sum, the debate should shift from whether safe harbor applies to how responsibility should be allocated. By clarifying concepts, defining roles, and establishing procedures, rejection, modification, and reconstruction can operate within a unified framework rather than in opposition. This approach better reflects the complexity of generative AI.

DOI: 10.54254/2753-7048/2025.28977

6. Conclusion

This paper has reviewed three main scholarly positions on whether generative AI (GAI) platforms may benefit from safe harbor protection: outright rejection of exemption, retention of conditional exemption with additional duties, and reconstruction of an entirely new liability framework. The consensus across the literature is that GAI's active content generation challenges the assumptions of "technological neutrality" and the "notice-and-takedown" mechanism, making the logic of passive exemption increasingly unworkable. Nevertheless, significant gaps remain at the practical level. There is still no clear and workable standard for what constitutes an "active role" or "control" by AI platforms. Rules for allocating responsibility among multiple actors are not yet coordinated. Likewise, consensus has not formed on how to move from simply removing infringing content toward processes that prevent infringement through model correction. Future research should therefore, explore a model of "conditional safe harbor". Under such a model, exemption would depend on benchmarks such as the platform's degree of technical control, the foreseeability of risks, and the extent of benefits gained. Clear duties should be assigned to data providers, model developers, and deployers. At the same time, a comprehensive governance process should be built, covering ex ante prevention, ongoing monitoring, and ex post correction. Only through such a structure can the law maintain a dynamic balance between encouraging innovation and safeguarding rights.

References

[1] Rosati, E., 2025. Infringing AI: Liability for AI-generated outputs under international, EU, and UK copyright law. European Journal of Risk Regulation, 16(2), pp.603-627.
[2] Samuelson, P., 2024. Fair use defenses in disruptive technology cases. UCLA L. Rev., 71, p.1484.
[3] Revolidis, I., 2024. Generative AI content misuse and the DSA. Available at SSRN 5124005.
[4] Lemoine, L. and Vermeulen, M., 2023. Assessing the Extent to Which Generative Artificial Intelligence (AI) Falls Within the Scope of the EU's Digital Services Act: an Initial Analysis. Available at SSRN 4702422.
[5] Pérez, G.E.M., 2025. From Curators to Creators: Navigating Regulatory Challenges for General-Purpose Generative AI in Europe. JIPITEC–Journal of Intellectual Property, Information Technology and E-Commerce Law, 16(2).
[6] Luk, A., 2024. The relationship between law and technology: comparing legal responses to creators' rights under copyright law through safe harbour for online intermediaries and generative AI technology. Law, Innovation and Technology, 16(1), pp.148-169.
[7] Zou, M. and Zhang, L., 2025, January. Navigating China's regulatory approach to generative artificial intelligence and large language models. In Cambridge Forum on AI: Law and Governance (Vol. 1, p. e8). Cambridge University Press.
[8] Zhang, J., 2025. Input out, output in: towards positive-sum solutions to AI-copyright tensions. Journal of Intellectual Property Law & Practice, p.jpaf037.
[9] Lin, Y. and Guan, T., 2025. From safe harbours to AI harbours: reimagining DMCA immunity for the generative AI era. Journal of Intellectual Property Law & Practice, p.jpaf043.
[10] Choi, S., 2025. Assessing the Efficacy of Third-Party Liability Copyright Doctrines Against Platforms That Host AI-Generated Content. BCL Rev., 66, p.1087.
[11] Lim, D., 2024. NOTIFICATION AND PERMISSION-BASED APPROACHES FOR GENERATIVE AI PLATFORMS. Available at SSRN.
[12] Sun, Y., 2014. The Road of Cooperation: Assessing the Evolving Interaction between Copyright Owners and ISPs —From ISP Legislation to the Graduated Response. Available at SSRN 2430213.
[13] Beard, T.R., Ford, G.S. and Stern, M.L., 2017. Fixing Safe Harbor: An Economic Analysis. Phoenix Center Policy Paper, (52).
[14] de STREEL, A. and Husovec, M., 2020. The e-commerce Directive as the cornerstone of the Internal Market. Available at SSRN 3637961.

DOI: 10.54254/2753-7048/2025.28977

[15] Ma, X., 2014. Establishing an Indirect Liability System for Digital Copyright Infringement in China: Experience from the United States' Approach. NYU J. Intell. Prop. & Ent. L., 4, p.253.
[16] Mohan, K., & Jansi, S. (2024). Determining due diligence principles to enable safe harbour protection for Generative Artificial Intelligence. International Journal of Research Publication and Reviews, 5(11), 5390–5396. Retrieved from https: //ijrpr.com/uploads/V5ISSUE11/IJRPR35382.pdf (accessed 28 Sept 2025).
[17] Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M.A. and Liang, P., 2023. Foundation models and fair use. Journal of Machine Learning Research, 24(400), pp.1-79.
[18] Lee, K., Cooper, A.F. and Grimmelmann, J., 2024, March. Talkin"Bout AI Generation: Copyright and the Generative-AI Supply Chain (The Short Version). In Proceedings of the 2024 Symposium on Computer Science and Law (pp. 48-63).
[19] Heng, A. and Soh, H., 2023. Selective amnesia: A continual learning approach to forgetting in deep generative models. Advances in Neural Information Processing Systems, 36, pp.17170-17194.