Proceedings of ICILLP 2026 Symposium: Comparative Politics and Legal Frameworks in Policy Development
DOI: 10.54254/2753-7048/2026.JO39103

Adapting the 'Safe Harbour' Rules for Generative AI Outputs — Centred on the Hierarchical Obligations of 'Notice and Necessary Measures'

Zirui Lu

Department of Law, Yunnan University, Kunming, China
zirui.lu.2004@outlook.com

Abstract. Copyright infringement arising from generative AI outputs can be categorised into two types: prompt-induced generation and model-memory reproduction. This typological differentiation has caused a structural shift in the prerequisites for the application of traditional notice rules, manifesting specifically in the loss of structural uniformity among the three parties involved, the dissolution of the locatability of infringing content, and the gradation of service providers' capacity to address such issues. The 'notice' rule and the 'knowledge' rule can be functionally distinguished: the former serves a procedural response function, whilst the latter serves a substantive function of determining fault. Under this regulatory framework, a tiered system of obligations should be established, grounded in the quality of the notice, bounded by technical capabilities, and operating through a mechanism of dynamic matching. Defective notices should only trigger obligations to forward the notice and rectify it, whilst valid notices should correspond to obligations to filter keywords and prompts; enhanced notices may extend to obligations to block outputs at the model level. The boundaries of obligations between application-layer developers and foundational model developers should also be distinguished on a binary basis, ensuring that disposal obligations are aligned with technical feasibility.

Keywords: generative artificial intelligence, copyright infringement, notice rules, knowledge rules

1. Introduction

With the rise and commercialisation of large generative AI models, the mode of information dissemination has undergone profound changes. Whilst AI brings unprecedented development opportunities, it has also given rise to massive and complex risks of copyright infringement. Inadequate remedies in judicial practice are no longer isolated incidents. In the Song of Youth case, the rights holder was unable to curb repeated infringements despite filing dozens of complaints [1]. Furthermore, in a technology-related case in Beijing, there was a marked divergence between the two levels of court regarding whether account suspension measures could be imposed, reflecting an insufficient regulatory framework defining the boundaries of necessary measures [2]. The ruling in the Soul Land case further noted that the necessary measures to be taken by internet service

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2026.JO39103

providers exhibit characteristics of 'dynamic development' [3]. Whether traditional enforcement measures can be effectively activated, what level of notification corresponds to what intensity of measures, and how to establish an operational correspondence between dynamically evolving necessary measures and their corresponding actions—these are all questions that existing regulations have failed to address.

The traditional notice rules set out in Article 1,195 of the Civil Code presuppose a tripartite structure comprising 'internet service providers, internet users and rights holders'; however, in the context of generative artificial intelligence, the service provider's control over generated content is reshaped by algorithmic black boxes and vast amounts of training data, causing the technical foundations for the application of traditional safe harbour rules to be structurally undermined. In the face of this challenge, clarifying the doctrinal basis for the functional separation between the notice rule and the knowledge rule, and thereby constructing a hierarchical system of obligations, is key to interpreting the application of Article 1,195 of the Civil Code in the context of generative AI.

2. A categorised analysis of copyright infringement at the output end of generative AI

Infringements arising from generative AI outputs exhibit significant internal differences in their triggering mechanisms; it is therefore necessary to distinguish between two types of infringing behaviour in order to provide a factual basis for subsequent regulatory development.

2.1. Methodological premises and technical calibration of the dual-track classification

The internal differences in infringement arising from generative AI outputs are primarily reflected in their triggering mechanisms. Some infringing outcomes depend on users constructing descriptive or suggestive prompts to bring them about, whilst others can be reliably reproduced with just a brief name. These two scenarios exhibit systematic differences in terms of the degree of user involvement, the predictability of infringing outcomes, and the scope for intervention by service providers; if the triggering mechanisms themselves are not distinguished, the subsequent allocation of obligations will lack a factual basis.

The technical core of the triggering mechanism lies in the strength of the statistical association between specific expressions in the training corpus and the output results. During the model's pre-training phase, stable associations form between high-frequency co-occurring names and expressions in the corpus, and these are solidified within the parameter structure as weight distributions. The reason a particular character name is consistently reproduced even in the absence of descriptive prompts lies in the fact that the association between that name and specific expressions occurs with extremely high frequency in the training corpus and possesses sufficient strength. The combination of stable associations and high-frequency repetition enables the model to produce highly specific infringing results based solely on a brief prompt [4]. Conversely, if the strength of the association is insufficient, the user must construct detailed descriptive or suggestive prompts to provide the model with sufficient generative cues before infringement can be triggered.

Accordingly, the substantive criterion for classification lies in the dominant mechanism triggering the infringement. Where the design of the user's prompt is the dominant factor—and the infringing content is not necessarily generated by the model but may only arise when guided by user input—it is classified as the 'prompt-induced generation type'; where the dominant factor is the fixed and highly frequent, stable parametric memory in the training corpus—with user input serving merely as an activation signal and the infringing result exhibiting a high degree of certainty and repeatability—it is classified as the 'model memory reproduction type' [5]. The differences between the two types

DOI: 10.54254/2753-7048/2026.JO39103

of infringement in terms of their generative principles directly determine the respective degrees of foreseeability and capacity for intervention that service providers possess, thereby providing a classificatory basis for the construction of a tiered system of obligations.

2.2. Case law support

With regard to model memory reproduction-type infringement, the Ultraman case [6], heard by the Guangzhou Internet Court, provides a test case. In this case, the plaintiff merely entered the character name 'Ultraman', and the defendant's service consistently produced images substantially similar to the plaintiff's work; users were able to trigger infringing results without resorting to any descriptive constructs. On this basis, the court established the effectiveness-based criterion that 'normal use of prompts by users cannot generate infringing images'. At the same time, it dismissed the plaintiff's claim to have relevant content removed from the training dataset on the grounds that the defendant had not actually carried out model training. Application-layer developers can intervene at the input and output stages through prompt filtering and content moderation, yet they have no means of accessing the memory mechanisms within the model's parameter layer; the misalignment between the capacity to address the issue and the party bearing the obligation thus becomes clearly discernible.

In contrast, the nation's first 'AI text-to-image' copyright case heard by the Beijing Internet Court directly corroborated the logic of prompt-induced generation. The judgement explicitly stated that users made an intellectual contribution to the generated content by inputting prompts and setting parameters, thereby reflecting their personalised expression [7]. In the context of regulating infringement, this stance serves as indirect evidence of the user's control and dominant position at the input stage in prompt-induced generation scenarios; consequently, the primary focus of the platform's duty of care should shift towards the input stage.

2.3. The structural impact of type differentiation on the prerequisites for the application of notification rules

The differentiation of content types has a threefold impact on the prerequisites for the application of traditional 'notice-and-takedown' rules.

Firstly, the undermining of the structural uniformity of the three-party framework. The traditional 'notice-and-takedown' mechanism presupposes a relatively clear three-party structure: the user, as the content provider, uploads or inputs content; the platform, as the service provider, fulfils its intermediary obligations; and the rights holder issues a notice to assert their rights. The 'developer-operator-user' tripartite structure broadly corresponds to this presupposition [8]. However, under the dual-track classification, in the 'prompt-induced generation' type, users substantively participate in the creation of infringing content by constructing prompts, and their status is closer to that of a traditional content provider; in the 'model memory reproduction' type, users merely input a brief name, and the determining factor for the infringing result is the memory associations embedded in the model's parameters, making users closer to 'triggers' rather than 'contributors'. The conduct of which party should serve as the basis for liability assessment in a notice varies depending on the type of infringement.

Secondly, the dissolution of the locatability of infringing content. In the prompt-induced generation type, the infringing result—as the output of a specific interaction—can be identified and fixed. However, in model memory reproduction-type infringement, the infringing content is not stored in a fixed form at a specific address, but is embedded within the model's parameter structure

DOI: 10.54254/2753-7048/2026.JO39103

in the form of a probability distribution; it does not exist independently in any form that can be directly identified until the next interaction occurs. There is as yet no definitive answer as to whether the subject of a rights holder's notice should be a specific output or a particular memory association embedded within the model itself.

Thirdly, the gradation in the capacity for intervention. In the case of prompt-induced generation, relatively effective preventive measures can be achieved through prompt filtering and content moderation at the application layer; application-layer developers possess the practical ability to anticipate and address such infringements. In contrast, for model memory reproduction, the triggering mechanism is embedded within the training and parameter structure of the base model; application-layer developers are unable to predict in advance whether specific terms will consistently trigger infringements, nor can they intervene in the model's parameters. In existing theory, the element of 'knowledge' is typically assessed as a holistic mental state without distinguishing specific technical capabilities. Whether this holistic logic of presumption of fault can continue to be uniformly applied to the two categories of infringement, where the capacity for handling has already undergone structural differentiation, has become an unavoidable issue.

3. The asymmetrical application of the notice rule and the knowledge rule in AI contexts

The threefold structural instability illustrates the technical obstacles to the application of the notice rule, rather than a loss of legitimacy in its application. The latter requires independent normative justification.

3.1. Confirmation of the basis for the analogous application of the notice rule

The deep-seated root of the academic divergence regarding whether the notice rule can be applied to copyright infringement arising from generative AI outputs lies in differing definitions of the scope of a service provider's control. By subdividing control into macro-level control over the algorithmic system itself and micro-level control over specific generated outputs, it becomes apparent that a service provider's control over the latter relies to a large extent on external information inputs; the procedural function of the notice mechanism lies precisely in bridging this information gap.

Existing theories have formed four representative positions regarding the applicability of the notice-and-takedown rules to copyright infringement arising from generative AI outputs, each presenting arguments from different normative dimensions. Xu Wei characterises the notice-and-takedown rules as an imputation clause, arguing that their applicability depends on whether the service provider can take effective measures against infringing content [9]; Wang Liming and Bao Ding Yurui propose a dynamic duty of care framework based on technological neutrality [10]; Liang Zhiwen advocates referring to the notice-and-takedown rules for search services [4]; Cheng Xiao, by contrast, argues that the rule should be excluded from application, on the grounds of the active nature of the behavioural structure and the technical uncertainty surrounding the normative purpose [11]. Taken as a whole, the approach of analogical application holds a relative advantage in terms of the completeness of its argumentation. It is grounded in the factual premise of the service provider's limited control over generated content, supported by the procedural information-filling function of the notification mechanism as a normative basis, and underpinned by the interpretative extensibility of Article 1195 of the Civil Code as a doctrinal foundation. These three dimensions of argumentation collectively delineate the basic conditions under which the notification rules are triggered in AI contexts, whilst also providing a normative foundation for the subsequent definition of 'what intensity of notification triggers what level of measures'. This framework of reasoning is likewise

DOI: 10.54254/2753-7048/2026.JO39103

corroborated at the comparative law level. In the case of Elsevier v. Cyando, the Court of Justice of the European Union confirmed three key principles: firstly, platform operators themselves do not generally constitute 'communication to the public'; secondly, the application of the safe harbour is conditional upon the platform operator's lack of knowledge of 'specific illegal content'; general knowledge is insufficient to deprive it of its safe harbour status; and thirdly, Member States are permitted to stipulate that rights holders must issue an explicit notice prior to obtaining an injunction [12]. These three key points correspond to the position of analogical application from the three dimensions of the characterisation of the subject, the specification of the knowledge requirement, and the clarity of the notice, respectively.

3.2. Functional separation of the notice rule and the knowledge rule

Building upon the application by analogy, a further question arises as to whether the notice rule and the knowledge rule can be applied separately. Articles 1195 and 1196 of the Civil Code together constitute a procedural, interactive mechanism. Triggered by the submission of evidence of infringement by the rights holder, this mechanism involves a series of procedural steps—including notification by the platform, a counter-notice by the internet user, and the platform's decision to maintain or lift the necessary measures—thereby forming a procedural workflow designed to resolve disputes swiftly. Procedural obligations themselves and the substantive liability arising therefrom belong to different levels of assessment; the proper fulfilment of procedural obligations does not give rise to tortious liability, and a breach thereof merely provides a possibility—rather than a certainty—for liability to arise; the determination of substantive liability must still be judged by reference to the constituent elements of tortious liability themselves [13]. This structure, in which procedure and substance are mutually independent, also applies to understanding the functional positioning of the notice rule, the normative content of which focuses on the ex post facto disposal process itself and does not establish substantive standards for assessing the subjective fault of service providers.

The 'knowledge rule' established in Article 1197, on the other hand, serves a purely substantive legal function, taking 'knowledge' or 'ought to have known' as the core elements for establishing fault, and retrospectively inferring the existence of fault based on objective facts such as the obviousness of the infringing content and the service provider's technical capabilities. The notification rule and the knowledge rule do not stand in a progressive or inclusive relationship in their application, but rather in a parallel relationship. A rights holder may either directly assert liability under the knowledge rule where they can prove that the service provider was aware of the infringement, or, where it is difficult to prove subjective fault, may instead issue a notice of infringement [14]. This further demonstrates that the operation of the notification rule does not depend on the establishment of fault under the knowledge rule.

3.3. The technical limitations on the 'knowledge' rule and the normative basis for their separate application

The key to whether the 'knowledge' rule applies lies in whether the service provider can make a prior assessment of the specific re-enactment of the infringement before the content is output. Research by Biderman et al. indicates that whether a specific training string has been memorised by the model cannot be reliably predicted by studying smaller models or partial training checkpoints, unless the testing computational power reaches a significant proportion of the full training computational power of the target model [15]. This implies that, given the computational resources currently

DOI: 10.54254/2753-7048/2026.JO39103

available, it is difficult for service providers to make a reliable prior assessment of whether a specific output will result in the reproduction of memorised content, thereby severing the causal link in knowledge required for the presumption of knowledge. If the standard of duty of care under the 'knowledge rule' were to be applied indiscriminately to the output stage, it would be tantamount to requiring service providers to bear a general duty of scrutiny that is difficult to fulfil in practice under current technological conditions.

The scope of application of the 'knowledge rule' should therefore be restricted to situations where service providers are technically capable of identifying and intercepting infringing content but fail to do so; it should no longer serve as a general requirement triggering a duty to take action. Such restriction must satisfy two conditions. The issue of apparent infringement addresses the evidential question of 'what facts are sufficient to presume that a service provider ought to have known of the infringement': the appearance of infringement must be objectively recognisable to a reasonable person. Section 512(c)(1)(A) of the US Digital Millennium Copyright Act (DMCA) establishes the dual knowledge requirements for the application of safe harbour provisions—actual knowledge and 'red flag awareness', the latter referring to facts or circumstances sufficient to make the infringing activity apparent [16]. In Viacom v. YouTube, the Second Circuit Court of Appeals interpreted the 'red flag' standard strictly, requiring that knowledge must relate to 'specific and identifiable instances of infringement'; general knowledge is insufficient to override safe harbour protection [17]; Capitol v. Vimeo further clarified that 'red flag awareness' must be assessed against an objective standard [18]. Technical feasibility holds independent normative significance in the context of generative AI: even if an infringing appearance is objectively recognisable to a reasonable person, the 'knowledge' rule should not be triggered if the service provider lacks the practical ability to identify and block that specific infringement under current technical conditions. The obviousness of the appearance provides the epistemological basis for establishing negligence, whilst technical feasibility defines the limits of the ability to fulfil obligations; together, these constitute a dual filtering mechanism for the 'knowledge' rule.

The position that the rules on limitation of liability should be applied separately from the rules on notification has precedents in Chinese judicial practice. In the Joy of Life case, heard by the Chongqing High People's Court, the court used whether the link was removed within a reasonable timeframe as the criterion for determining procedural timeliness with regard to notifications of specific infringing links; at the same time, it independently made a substantive determination of 'should have known' based on objective facts such as the scale of dissemination on the platform and viewing popularity [19]. The handling of individual notification cases and subjective knowledge of the overall state of infringement fall under two distinct evaluation frameworks—procedural compliance and substantive fault—each with its own independent criteria for assessment, and neither need be a prerequisite for the other.

4. The regulatory framework for a tiered system of necessary notification measures

Since notification rules can trigger an obligation to take action, the standard for the intensity of that obligation must be clarified. There are objective differences in the clarity and traceability of information provided in notices by rights holders, whilst service providers' capacity to control the output, input and parameter layers varies across a spectrum; these two factors jointly determine that the intensity of necessary measures should be differentiated. The principle of proportionality requires that measures be commensurate with technical feasibility; it is inappropriate to impose obligations on service providers that exceed their actual capacity for control. Therefore, a tiered

DOI: 10.54254/2753-7048/2026.JO39103

system of assessment criteria must be established, centred on the quality of notices, technical capabilities and their dynamic interplay.

4.1. Tiered assessment criteria for the quality of notices

In scenarios of copyright infringement at the output end of generative artificial intelligence, the analogous application of notification rules depends not only on the service provider's control over the generated content, but also on the subjective clarity and objective locatability of the information provided by the rights holder. The single notification pathway of 'a link constitutes identification' found in traditional online services encounters structural obstacles when applied to generative AI. As generated content is not stored as static files at specific web addresses but is dynamically generated based on probabilistic algorithms, it is necessary to establish a tiered standard for 'notification quality' that takes into account both technical characteristics and information density, to serve as the procedural starting point for triggering necessary measures of varying intensity by service providers.

Depending on the completeness of the information, the quality of a notice can be categorised into three progressive levels.

First, an incomplete notice, where the notice lacks the basic elements required to locate the infringing act and cannot be converted into an enforceable course of action. In the Baidu Netdisk Gilded Chopsticks case, the rights holder provided only a hash value and the title of the work, without supplying a specific sharing link or supplementary attribute information. The Guangdong Provincial High People's Court ruled that such notices, lacking relevant supplementary information, cannot be converted into an actionable identification pathway and therefore do not give rise to a substantive obligation to take remedial action [20]. The legal consequence of a defective notice is that it does not trigger a substantive obligation to take remedial action.

Secondly, a basic notice must include proof of the rights holder's identity, preliminary evidence of rights ownership, and search criteria that meet the minimum technical localisation requirements. In cases where the content is generated through keyword prompts, the notice should include images of the infringing material, screenshots of interaction records, and the time of generation; in cases where the content is reproduced from a model's memory, the notice should specify the particular high-frequency keywords or characteristic elements that trigger the reproduction of the infringement. The function of the basic notice is to enable a reasonable service provider, without being subject to a duty of proactive review, to identify suspected infringement through routine technical means and pinpoint it to a specific user account.

Thirdly, an enhanced notice should, building upon the basic notice, proactively disclose more in-depth technical evidence, including precise prompt combinations, algorithm-induced pathways, test reports demonstrating reproducibility, or analyses of copyright feature watermarks. In the ByteDance v. Tencent (Soul of War) case, the court did not uphold the rights holder's claim that the platform should filter all potentially infringing videos, on the grounds that the rights holder should actively cooperate with the service provider to propose screening scopes, standards or methods [21]. If no specific preventive guidelines are provided, the platform lacks the technical basis to implement filtering and blocking measures [22]. By analogy with the field of generative AI, enhanced notices—by providing screening scope, criteria or reproducible pathways—overcome the service provider's knowledge barrier, thereby upgrading the duty of care from passive 'ex post disconnection' to active 'source filtering'.

DOI: 10.54254/2753-7048/2026.JO39103

4.2. Defining the boundaries of measure intensity based on technical capability

The delineation of technical capabilities must be based on the actual control that application-layer developers exercise over the infringement process. The scope of control is distributed along a gradient ranging from 'prompt filtering at the input end' to 'content interception at the output end' and 'adjustment of underlying parameters': the first two fall within the control of the application layer, whilst the latter constitutes an uncrossable red line. The Ultraman case established a fundamental boundary: the defendant accessed a third-party large language model via an API interface; the court held that its duty of reasonable care was limited to implementing technical measures within its controllable scope (such as prompt filtering) to ensure that users could not generate infringing images when using relevant prompts normally; the claim seeking the removal of materials from the training dataset was dismissed on the grounds that the defendant was not the actual trainer of the model and could not influence the base model's dataset [6]. This boundary was exceptionally extended in the Ultraman LoRA Model case. Although the defendant utilised a third-party open-source model as a foundation, it exercised actual control over the LoRA overlay models that users had trained themselves and shared on the platform's marketplace. On this basis, the court imposed a duty to remove individually generated images, whilst requiring the complete removal of the LoRA model itself, which possessed the capability to repeatedly generate infringing content [23] —the extent of control marks the limit of the obligation to take necessary measures; beyond this boundary lies the underlying realm beyond the reach of technical capabilities. The legal characterisation of the attributes of model weights in the Getty Images v. Stability AI case further addresses the question of whether the scope of such measures can extend to the model parameter layer. The High Court of Justice in the United Kingdom held that, once training is complete, model weights are merely probability distributions crystallised through statistical learning; they do not themselves store any content from copyrighted works, nor do they constitute or carry infringing copies [24]. This ruling rejects the universal legal proposition that model weights are invariably infringing copies; mandating the deletion, deconstruction or modification of large models at the output stage not only lacks a legal basis but also exceeds the reasonable limits of technical feasibility.

Accordingly, technical capabilities in generative AI scenarios can be categorised into three progressive levels, each corresponding to different boundaries for measures. Firstly, output-level handling refers to the interception, blocking or deletion of specific generated results at the interactive terminal; this constitutes the baseline capability that all service providers must possess. Secondly, intervention at the input level and during the generation process, which includes prompt filtering (maintenance of a keyword database and blocking of sensitive instructions) and the use of Retrieval-Augmented Generation (Note: Retrieval-Augmented Generation (RAG) refers to a class of technical architectures in which external knowledge base retrieval is incorporated during the generation phase, using non-parametric memory to augment the output of parametric models [25].) to introduce comparisons with external vector indices, as well as the use of a safety classifier (Note: A 'safety classifier' refers to an independent classification model that performs a secondary risk assessment of a model's output; it is typically deployed at the output stage, acting as a filtering layer separate from the generation process [26].) to review the risk of infringement at the output stage. This level does not involve the parameter structure of the foundation model and falls within the standard capabilities of application-layer developers with independent operational backends. Thirdly, adjustments to model parameters and training data, including fine-tuning the weight distributions of neural networks, reverse erasure, or cleaning and reorganising pre-training datasets.

DOI: 10.54254/2753-7048/2026.JO39103

Of these, full retraining and large-scale data cleansing are the exclusive preserve of foundational model developers; whilst mild fine-tuning methods such as RLHF [27] and Constitutional AI [28] have manageable computational requirements, they still constitute interventions involving the modification of weights and should therefore be the responsibility of foundational model developers; they must not be imposed on application-layer developers who merely call underlying services via API interfaces.

4.3. Rules for the dynamic matching of the intensity of measures

The quality of the notice determines the information foundation, whilst technical capabilities define the boundaries of implementation; the dynamic alignment of these two factors determines the actual intensity that necessary measures should possess. This alignment is reflected in two progressive rules and an overarching proportionality principle.

Firstly, the obligation to match the baseline for defects or foundational notices. Where the rights holder merely provides general proof of ownership and compares it with the results of the infringement, such a notice can only trigger the minimum obligation on the part of application-layer developers: to implement keyword filtering at the output stage or to block content retrospectively, thereby addressing the most superficial issues in cases where prompt-induced generation occurs. The standard for meeting this minimum obligation should adopt the effectiveness-based test established in the Ultraman case: service providers must, within the limits of their controllable technical capabilities, ensure that users cannot generate substantially similar images when using the relevant prompt words in the course of normal use [6]. The term 'normal use' here constitutes a reasonable limitation: if a user forcibly induces the model to output infringing content through multiple rounds of dialogue, the deliberate construction of concealed descriptions, or the circumvention of safety words, this exceeds the everyday interaction habits of ordinary users and the service provider's reasonable expectations for prevention under their baseline obligations; the platform should not bear joint and several liability for such actions.

Secondly, the duty of active prevention through notice matching is strengthened. Where a notice meets the statutory criteria and can accurately pinpoint the infringing characteristics, the primary duty-bearer remains the application-layer developer; however, the duty to take action extends to the input level: the establishment of a prompt-blocking database targeting specific subject-matter of rights. This rule can prevent the vast majority of infringement cases induced by prompts. For example, in the Ultraman LoRA Model case, this rule manifested as an obligation to delete specific LoRA models within the scope of control in their entirety, as the LoRA models themselves possess the capacity to be repeatedly invoked by other users and trigger infringement on a large scale [23]. The ByteDance v. Tencent (Soul of War) case also confirms that, upon receiving notices of ongoing infringement and becoming aware that algorithmic recommendations are fuelling the spread of infringement, service providers should bear a stricter duty of care and proactively develop or deploy content-recognition-based pre-filtering and blocking mechanisms in accordance with the screening scope and criteria provided by the rights holder [21].

The intensity of both of the aforementioned obligations is subject to the principle of proportionality: measures must be commensurate with the severity of the infringement and must not cause disproportionate harm. The prohibition on general monitoring obligations constitutes a specific application of the principle of proportionality in the context of the 'notice-and-takedown' framework. The essence of this rule does not lie in the prohibition of filtering technology itself, but rather in whether the scope of the filtering is targeted—filtering directed at specific works clearly identified by the rights holder constitutes lawful, specific scrutiny; whereas blanket pre-screening of

DOI: 10.54254/2753-7048/2026.JO39103

all uploaded content or all generated results exceeds the necessary limits [29, 30]. Domestic judicial rulings also adhere to this position [31]. The Joy of Life case clarified that measures should be limited to what is sufficient to prevent the continuation of infringement without causing disproportionate harm, and that the cessation of infringement is restricted to content that has been found to be infringing in the specific case [19]. The Baidu Netdisk Gilded Chopsticks case applied this upper limit at the operational level: blocking the sharing function constitutes a reasonable and necessary measure, but deleting files stored on the server does not in itself constitute a necessary measure, as the files themselves may not necessarily be infringing, and deletion would affect other users who use them solely for personal study and research [20]. Necessary measures must be limited to specific acts of infringement that have already been established; technical feasibility must not be used to expand the scope of action beyond the content or user groups specified in the rights holder's claim. This constraint applies at every level of the two tiers of obligations (Note: This upper-bound constraint imposed by the principle of proportionality on the intensity of necessary measures is likewise reflected in overseas judicial practice. In Roadshow Films Pty Ltd v. iiNet Ltd, the High Court of Australia held that even where a service provider possesses an ultimate technical means of intervention, it should not be required to adopt such a measure where the cost of doing so is disproportionately high and the measure is not targeted at a specific infringing act [32].)

On this basis, 'model memory reproduction' infringements impose distinct requirements regarding the allocation of remedial obligations. The mechanism triggering such infringements is embedded within the model's parameter structure, rendering traditional application-layer interception ineffective. If the rights holder submits advanced evidence, such as a precise algorithmic induction path (constituting an enhanced notice), the obligation to take remedial action must be extended upwards to the developers of the foundational model, who must then employ low-level measures—such as model parameter adjustments, targeted fine-tuning or reinforcement learning using negative examples—to eradicate the risk of infringement. It should be noted that this extension must be targeted at 'specific objects identified by the rights holder', thereby forming a point-to-point defence, rather than involving a blanket pre-screening of all generated content.

In summary, there is a structural difference in the level of control over the algorithmic system between application-layer developers and foundational model developers, (Note: Overseas courts have also begun to distinguish between the different roles and liabilities of developers of underlying models and application-layer platforms. In the case of Andersen v. Stability AI, the United States District Court for the Northern District of California treated Stability AI (the model developer) and DeviantArt and Midjourney (application-layer platforms) differently as early as the pre-trial stage [33].), which determines the differing levels of responsibility they bear under the three tiers of remedial obligations. Application-layer developers possess more direct configuration authority over model outputs, and their obligations are concentrated on output-level handling and prompt filtering; the obligations of foundational model developers are limited to the model parameter level within the scope of their technical capabilities, and they should not be held liable for obligations beyond this scope of control. This distinction aligns with the actual structure of the industry chain, thereby avoiding the imposition of obligations that cannot be practically fulfilled upon entities lacking the requisite capabilities (Note: Some comparative law scholars refer to interventions at the model parameter level as 'model disgorgement', a category encompassing techniques such as model retention, machine unlearning, and isolation. As an emerging remedial direction, its technical feasibility and its conformity with the principle of proportionality both remain to be tested, and it has not yet been generally recognised in judicial practice [29].)

DOI: 10.54254/2753-7048/2026.JO39103

5. Conclusion

The impact of copyright infringement arising from generative AI outputs on traditional safe harbour rules is manifested in the erosion of the technical foundations underpinning their application. The logical starting point for addressing this lies in typological differentiation: distinguishing between two types of infringement—prompt-induced generation and model-memory reproduction. These two types exhibit systemic differences in the degree of user participation, the foreseeability of infringing outcomes, and the scope for service provider intervention; the subsequent allocation of obligations must be premised on this factual basis.

The notice rule and the knowledge rule may be functionally separated. The notice rule serves a procedural response function, focusing on ex post facto handling procedures, and does not establish substantive criteria for assessing subjective fault. The knowledge rule serves the function of determining substantive fault, with 'knowledge' or 'ought to have known' as the core elements for establishing fault. In the context of generative AI, the 'knowledge' rule should be subject to the combined constraints of 'obviousness of the infringing appearance' and 'technical feasibility'. The former requires that the infringing appearance be objectively recognisable to a reasonable person, whilst the latter requires that the service provider possess the practical ability to identify and block specific infringements under existing technical conditions. The combination of these two conditions restricts the 'knowledge' rule to an objective assessment of the duty of care, whilst reserving regulatory space for the 'notification' rule to independently fulfil its procedural response function.

The tiered system of obligations centres on the quality of notifications and is bounded by technical capabilities. Notification quality is categorised into three levels: defective notifications, basic notifications and enhanced notifications; technical capabilities are categorised into three tiers: output-level handling, input-level and intervention in the generation process, and adjustment of model parameters and training data. The dynamic alignment between these two aspects manifests as follows: a defective or basic notification triggers baseline obligations at the output level, whilst an enhanced notification triggers proactive prevention obligations at the input level; for models of the 'memory-reproduction' type, under the conditions of an enhanced notification, these obligations may extend to the adjustment of parameters at the foundational model layer. Application-layer developers and foundational model developers thereby assume a two-tiered hierarchy of obligations, ensuring that disposal obligations are aligned with technical feasibility and avoiding the imposition of a general duty of scrutiny upon service providers.

Looking ahead, how specific trigger rules for foundational model developers will gradually accumulate based on judicial precedents, and under what conditions remedies at the parameter level —such as model deactivation—will be technically feasible, will require ongoing scrutiny during the process of adapting the rules governing generative AI infringement.

References

[1] Appeal in the case of So-and-so Baba (China) Network Technology Co., Ltd. Qingdao Branch et al. v. Shanghai So-and-so Digital Technology Co., Ltd. concerning infringement of the right to disseminate works via information networks (2022) Shandong Provincial High People's Court, Civil Judgement No. (2022) Lu Min Zhong 2743.

[2] Case concerning the infringement of the right to disseminate works via information networks between Beijing X Technology Co., Ltd. 1 and Beijing X Technology Co., Ltd. 2 (2023) Beijing Intellectual Property Court, Civil Judgement No. (2023) Jing 73 Min Zhong 386.

[3] Ruling on the review of non-litigious provisional measures in the case of Beijing Weibo Shijie Technology Co., Ltd., Shenzhen Tencent Computer Systems Co., Ltd. et al. (2021) Chongqing No. 1 Intermediate People's Court, (2021) Yu 01 Xing Bao No. 1-1.

DOI: 10.54254/2753-7048/2026.JO39103

[4] Liang, Z. W. (2025). The legal characterisation of generative AI outputs under copyright law and the rules governing liability. *Legal Science, 6*, 129-146. DOI: 10.3969/j.issn.1000-4238.2025.6.fax202506009.
[5] Zheng, Z. F., & Luo, L. C. (2024). On the 'safe harbour' rule for generative artificial intelligence. *Artificial Intelligence, 1*, 91-99. DOI: 10.16453/j.2096-5036.202402.
[6] Shanghai Xinchuanghua Cultural Development Co., Ltd. v. Guangzhou Nianguang Co., Ltd. (Internet infringement liability dispute) (2024) Guangzhou Internet Court, Civil Judgement No. (2024) Yue 0192 Min Chu 113.
[7] Li v. Liu: Copyright infringement dispute (2023) Beijing Internet Court, Civil Judgement No. (2023) Jing 0491 Min Chu 11279.
[8] Xu, W. (2023). On the legal status and liability of generative AI service providers: A case study of ChatGPT. *Legal Science (Journal of Northwest University of Political Science and Law), 41*(4), 69-80. ISSN: 1674-5205.
[9] Xu, W. (2024). On the application of the notice rule in generative artificial intelligence infringement. *Modern Law Science, 46*(3), 144-157. DOI: 10.3969/j.issn.1001-2397.2024.03.10.
[10] Wang, L. M., & Bao, D. Y. R. (2025). On the analogous application of the 'notice' rule in infringement cases involving generative artificial intelligence works. *Journal of Comparative Law, 4*, 1-13. DOI: 10.3969/j.issn.1004-8561.2025.4.bjfyj202504001.
[11] Cheng, X. (2026). On why rules governing online infringement should not apply to infringement by generative artificial intelligence services. *Journal of Comparative Law, 1*, 125-137. ISSN: 1004-8561.
[12] Frank Peterson v. Google LLC, YouTube Inc. and Elsevier Inc. v. Cyando AG, Joined Cases C-682/18 and C-683/18 (2021) ECLI:EU:C:2021:503, 22 June 2021.
[13] Yu, B. (2025). The theoretical justification for online platforms' obligation to restore content and the improvement of the liability regime. *Modern Law Science, 47*(4), 132-145. DOI: 10.3969/j.issn.1001-2397.2025.04.09.
[14] Lead Group for the Implementation of the Civil Code of the Supreme People's Court. (2020). *Understanding and application of the tort liability part of the civil code of the People's Republic of China*. People's Court Press. ISBN: 978-7-5216-1120-5.
[15] Biderman, S., Prashanth, U. S., Sutawika, L., et al. (2024). Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems, 37*, 1-19.
[16] 17 U.S.C. § 512(c)(1)(A) (2018).
[17] Viacom International, Inc. v. YouTube, Inc. (2012) 676 F.3d 19, 30-31 (2d Cir.).
[18] Capitol Records, LLC v. Vimeo, LLC (2016) 826 F.3d 78, 93-94 (2d Cir.).
[19] Appeal in the case of Tencent Technology (Beijing) Co., Ltd. et al. v. Chongqing Tianji Meike Technology Co., Ltd. concerning infringement of the right to disseminate works via information networks and unfair competition (2024) Chongqing Higher People's Court, Civil Judgement No. (2024) Yu Min Zhong 172.
[20] Retrial case concerning the infringement of the right to disseminate works via information networks between Beijing Baidu Online Network Technology Co., Ltd. and the Guangzhou Branch of Shanghai Feicui Dongfang Communication Co., Ltd. (2022) Guangdong Provincial High People's Court, Civil Judgement No. (2022) Yue Min Zai 59.
[21] Case of ByteDance et al. v. Tencent et al. regarding infringement of the right to disseminate works via information networks (2022) Hangzhou Intermediate People's Court, Zhejiang Province, Civil Judgement No. (2022) Zhe 01 Min Zhong 3098.
[22] Sha, L. (2023, April 27). Governance of algorithmic recommendations and the determination of platform liability for infringement: The Hangzhou Intermediate People's Court's judgement in the case of ByteDance et al. v. Tencent et al. regarding infringement of the right to disseminate works via information networks. *People's Court Daily*. Retrieved from https://www.chinacourt.cn/article/detail/2023/04/id/7265759.shtml.
[23] Appeal in the copyright infringement and unfair competition dispute between a certain cultural development Co., Ltd. in Shanghai and a certain intelligent technology Co., Ltd. in Hangzhou (2024) Hangzhou Intermediate People's Court, Zhejiang Province, Civil Judgement No. (2024) Zhe 01 Min Zhong 10332.
[24] Getty Images (US), Inc. v. Stability AI Ltd (2025) EWHC 2863 (Ch), paras 599-601.
[25] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*.
[26] OpenAI. (2022). New and improved content moderation tooling. *OpenAI Technical Report*. Retrieved from https://openai.com/blog/new-and-improved-content-moderation-tooling.
[27] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744.

DOI: 10.54254/2753-7048/2026.JO39103

[28] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv Preprint*, arXiv:2212.08073.

[29] Quintais, J. P. (2025). Generative AI, copyright and the AI Act. *Computer Law and Security Review, 56*, 1–17. DOI: 10.1016/j.clsr.2025.106107.

[30] Case C-401/19, Republic of Poland v. European Parliament and Council of the European Union, ECLI:EU:C:2022:297.

[31] Article 8(2) of the 'Provisions of the Supreme People's Court on Several Issues Concerning the Application of Law in the Trial of Civil Disputes Involving Infringement of the Right to Dissemination via Information Networks' (amended in 2020).

[32] Roadshow Films Pty Ltd v. iiNet Ltd (2012) 248 CLR 42 (High Court of Australia).

[33] Andersen v. Stability AI Ltd., No. 3:23-cv-00201, Order on Motions to Dismiss (N.D. Cal. Oct. 30, 2023).