Collective Licensing for AI Training Data: A Regulatory Design Proposal for Copyright Law

Onyinye Odita

Technology Law Associate,

Punuka Attorneys & Solicitors

Abstract: The rapid deployment of artificial intelligence systems capable of generating text, images, music, and code has intensified legal and policy debates surrounding the use of copyrighted works as training data. At a broad level, existing copyright frameworks were not designed to address large-scale, automated ingestion of protected content for machine learning purposes, resulting in regulatory uncertainty, fragmented litigation, and tension between rights holders and AI developers. Current approaches ranging from unlicensed use claims to narrow exceptions struggle to balance innovation incentives with the protection of creative labor. This article proposes collective licensing as a regulatory design solution for governing the use of copyrighted works in AI training. Drawing on established models from music, broadcasting, and reprographic rights, the analysis situates collective licensing as a middle-ground mechanism capable of reducing transaction costs, ensuring remuneration, and providing legal certainty at scale. The article examines the structural limitations of individual licensing and fair use–based approaches when applied to high-volume, non-expressive uses inherent in machine learning workflows. Narrowing its focus, the article outlines a conceptual framework for collective licensing tailored to AI training contexts. It explores institutional design considerations, including rights aggregation, scope of licensed uses, opt-out mechanisms, remuneration models, transparency obligations, and governance safeguards. Comparative insights from existing collective management organizations inform the proposal, highlighting how regulatory oversight and standardized tariffs could mitigate market power imbalances while supporting technological development. The article argues that a well-designed collective licensing regime can align copyright law with data-driven innovation by preserving incentives for creators, enabling lawful AI development, and reducing adversarial enforcement dynamics. By reframing AI training as a compensable, regulated use rather than an ungoverned exception, the proposal offers a pragmatic pathway for modernizing copyright law in the era of large-scale artificial intelligence.

Keywords: Collective licensing; AI training data; Copyright regulation; Collective management organizations; Data governance; Artificial intelligence

1. INTRODUCTION

1.1 Large-Scale AI Training and the Strain on Copyright Law

The development of contemporary artificial intelligence systems increasingly depends on large-scale training processes that ingest vast volumes of textual, visual, and audio data [1]. These datasets frequently include copyrighted works drawn from books, news articles, images, music, and online repositories [2]. Unlike traditional uses of protected works, AI training involves automated copying, transformation, and statistical analysis at a scale that far exceeds the assumptions underlying conventional copyright licensing frameworks [3].

Copyright law has historically been structured around identifiable acts of reproduction, distribution, and public performance involving discrete works and traceable users [4]. Large-scale AI training disrupts this structure by introducing diffuse, non-expressive uses that are difficult to license individually and challenging to evaluate through traditional infringement analysis [5]. Rights holders face practical barriers to monitoring use, while AI developers encounter uncertainty regarding the legality of training practices [6].

This structural mismatch has generated increasing tension between copyright law and data-driven innovation. Existing licensing mechanisms struggle to accommodate high-volume, low-visibility uses, while doctrinal tools such as fair use offer uncertain and jurisdiction-specific outcomes [7]. As a result, copyright law risks either under-protecting creative labor or over-constraining technological development. Addressing this tension requires rethinking how copyright governance operates when protected works are used as inputs for large-scale computational processes rather than as expressive outputs [8].

1.2 From Litigation and Exceptions to Regulatory Design

Initial legal responses to AI training practices have largely taken the form of litigation and doctrinal argumentation. Rights holders have pursued infringement claims, while AI developers have relied on defenses grounded in fair use, implied license, or non-expressive use theories [9]. These adversarial approaches have produced fragmented outcomes and limited guidance for future conduct.

At the same time, statutory exceptions and limitations have proven ill-suited to address the scale and economic significance of AI training. Fair use analysis, in particular, requires case-specific balancing that offers little ex ante certainty for either creators or developers [10]. Reliance on exceptions also risks normalizing uncompensated uses of creative works, raising concerns about long-term incentives for cultural production.

These shortcomings have prompted calls for regulatory design solutions that move beyond after-the-fact adjudication. Rather than framing AI training as either infringement or exception, regulatory approaches seek to structure lawful use through standardized mechanisms that allocate rights and responsibilities at scale [5]. Collective licensing represents one such approach, offering a framework for legal certainty, remuneration, and reduced transaction costs. This shift from

litigation toward institutional design reflects recognition that systemic challenges require systemic solutions rather than piecemeal doctrinal adaptation [1].

1.3 Objectives, Scope, and Structure of the Article

This article proposes collective licensing as a regulatory design solution for governing the use of copyrighted works in AI training. Its primary objective is to evaluate whether collective licensing can address market failure, reduce legal uncertainty, and balance innovation incentives with protection of creative labor [6]. The analysis treats AI training as a distinct category of copyright-relevant activity requiring tailored regulatory treatment.

The scope of the article is doctrinal, institutional, and policy-oriented. It examines copyright principles related to reproduction and licensing, surveys existing collective management models, and assesses how these frameworks might be adapted to AI training contexts [3]. The article does not advocate for blanket exceptions or unrestricted use but instead explores structured, compensatory mechanisms.

The structure proceeds from foundational analysis to normative proposal. Following this introduction, subsequent sections examine copyright doctrine and AI training, analyze collective licensing models, propose a tailored regulatory framework, and assess implementation challenges and policy trade-offs. The article concludes by evaluating the viability of collective licensing as a scalable and principled response to AI-driven transformation in copyright law [8].

2. COPYRIGHT LAW AND AI TRAINING: DOCTRINAL AND INSTITUTIONAL FOUNDATIONS

2.1 Copyright's Treatment of Non-Expressive Uses and Data Processing

Copyright law has traditionally been concerned with expressive uses of protected works rather than computational or analytical processing. Core doctrines such as reproduction and fixation were developed to address copying that enables human perception and consumption, not machine-driven analysis [9]. In the context of AI training, however, protected works are copied, stored, and processed as data inputs rather than as expressive outputs.

Machine learning training typically involves temporary and permanent reproductions of copyrighted material, raising questions about whether such acts fall within the scope of exclusive rights [12]. While fixation doctrine historically requires a work to be embodied in a tangible medium perceptible for more than a transitory duration, digital processing challenges this requirement by blurring distinctions between ephemeral and durable copies [6].

Courts have occasionally recognized that certain non-expressive uses may fall outside the core concerns of copyright, particularly where copying serves an analytical or functional purpose [14]. Nevertheless, copyright law lacks a clear, generally applicable doctrine that categorically addresses large-scale data processing for machine learning. As a result, AI training occupies an ambiguous legal space, neither clearly infringing nor clearly exempt.

This ambiguity is compounded by the scale of AI training, where millions of works may be processed simultaneously [10]. Traditional copyright concepts, developed for individualized acts of copying, struggle to accommodate automated, statistical uses that do not substitute for expressive consumption. This doctrinal uncertainty forms a critical backdrop for evaluating alternative governance mechanisms [15].

2.2 Limitations of Individual Licensing and Fair Use at Scale

Individual licensing has long served as the default mechanism for authorizing uses of copyrighted works. In the context of AI training, however, individual licensing encounters severe practical limitations. The sheer number of works involved, combined with difficulties in identifying rights holders and negotiating terms, creates prohibitive transaction costs [7]. These barriers render individualized consent economically infeasible for both creators and AI developers.

Fair use has emerged as an alternative justification for unlicensed training, particularly where use is characterized as transformative or non-expressive [13]. Yet fair use analysis is inherently fact-specific and retrospective, offering limited ex ante certainty. Developers cannot reliably predict how courts will balance purpose, amount, and market effect in novel AI contexts [6].

Enforcement challenges further undermine both approaches. Rights holders face difficulties detecting use within opaque training pipelines, while remedies are ill-suited to diffuse, large-scale copying [11]. Litigation therefore becomes a blunt and inefficient tool, favoring well-resourced actors and producing inconsistent outcomes.

These limitations indicate structural mismatch rather than doctrinal failure. Copyright's individualized licensing and case-by-case exception models were not designed for high-volume, automated data processing [15]. As AI training becomes integral to economic and cultural production, reliance on these mechanisms alone risks legal uncertainty, uneven compensation, and prolonged conflict between stakeholders [8].

2.3 Regulatory Authority, Market Failure, and the Case for Collective Solutions

The shortcomings of existing copyright mechanisms in AI training contexts point toward a classic case of market failure. Transaction costs, information asymmetries, and enforcement inefficiencies prevent rights holders and users from reaching mutually beneficial arrangements [10]. Where markets cannot self-organize efficiently, regulatory intervention becomes a plausible corrective.

Collective solutions have historically emerged in copyright law to address similar failures. Collective management organizations aggregate rights, standardize licensing terms, and distribute remuneration where individual bargaining is impractical [12]. These institutions operate under statutory or regulatory authorization, balancing private rights with public interest considerations.

In the AI training context, regulatory authority could play a crucial role in enabling collective licensing frameworks. Legislatures or administrative bodies may define the scope of licensable uses, authorize collective entities, and establish oversight mechanisms to ensure transparency and fairness [14]. Such intervention does not displace copyright but reshapes its operation to function at scale.

Collective licensing thus represents a shift from adjudication to governance. Rather than resolving disputes ex post, it structures lawful use ex ante, reducing uncertainty and aligning incentives [9]. By addressing market failure directly, collective solutions offer a principled pathway for reconciling copyright protection with data-driven innovation, setting the stage for more detailed regulatory design [6].
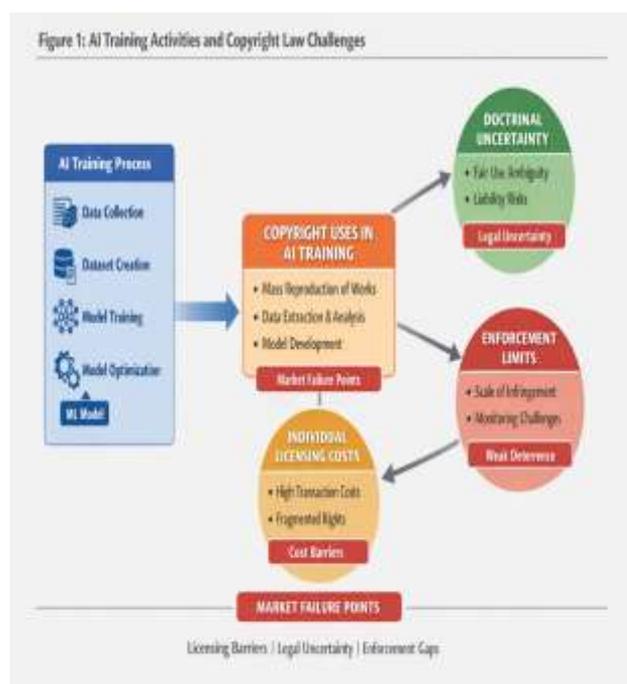

Figure 1: *Conceptual mapping of AI training activities within copyright law and market failure points.*

3. COLLECTIVE LICENSING IN COPYRIGHT LAW: MODELS AND LESSONS

3.1 Historical Development of Collective Management Organizations

Collective management organizations (CMOs) emerged as institutional responses to structural limitations in individual copyright licensing. Their development was driven by sectors where copyrighted works were used frequently, diffusely, and by multiple users, making direct negotiation impractical [15]. The music industry provides the earliest and most influential example. Performing rights organizations were established to license public performances of musical works, collect royalties, and distribute revenues to composers and publishers on a collective basis.

Broadcasting further accelerated the need for collective licensing. Radio and television transmissions involved simultaneous use of vast repertoires, rendering individual authorization infeasible [19]. CMOs enabled broadcasters to obtain blanket licenses while ensuring remuneration for rights holders. Similar dynamics arose in reprography, where photocopying of books and journals in educational and institutional settings led to the creation of collective schemes to manage secondary copying [13].

With the rise of digital technologies, collective licensing expanded to cover online uses, cable retransmission, and private copying levies [17]. These regimes adapted traditional collective principles to new technological contexts, often under statutory mandates. Throughout this evolution, collective licensing has served as a pragmatic compromise between exclusive rights and widespread access.

The historical trajectory of CMOs reveals a recurring pattern: where individual licensing fails due to scale, frequency, or opacity of use, collective solutions emerge [20]. This pattern is directly relevant to AI training, which similarly involves high-volume, automated use of copyrighted works beyond the reach of conventional licensing mechanisms.

3.2 Institutional Design, Governance, and Oversight Mechanisms

The effectiveness of collective licensing depends heavily on institutional design, governance structures, and regulatory oversight. At the core of collective management is rights aggregation, whereby individual authors or publishers mandate a collective entity to license specified uses on their behalf [16]. This aggregation enables blanket licensing and reduces transaction costs for users.

Tariff setting is a central governance function. CMOs typically establish standardized rates based on usage categories, revenue metrics, or sectoral norms [18]. These tariffs must balance fair remuneration for rights holders with affordability for licensees. In many jurisdictions, tariffs are subject to approval or review by regulatory bodies to prevent abuse of market power.

Transparency and accountability are equally critical. CMOs manage large revenue flows and must maintain accurate reporting, distribution rules, and dispute resolution mechanisms [14]. Regulatory supervision often includes auditing requirements, governance standards, and obligations to represent rights holders equitably.

Oversight frameworks vary across jurisdictions. Some systems rely on competition law and general corporate regulation, while others establish specialized copyright

tribunals or administrative agencies [19]. These institutions play a key role in resolving disputes over tariffs, licensing scope, and revenue distribution.

Institutional design thus determines whether collective licensing functions as a legitimate intermediary or devolves into a monopolistic bottleneck. The experience of existing CMOs demonstrates that collective solutions can succeed only when supported by clear mandates, robust governance, and effective regulatory supervision [20].

3.3 Strengths and Weaknesses of Existing Collective Licensing Regimes

Collective licensing regimes offer several well-documented strengths. Most notably, they reduce transaction costs by replacing thousands of individual negotiations with standardized licenses [13]. This efficiency enables lawful use at scale while ensuring that creators receive remuneration. Collective systems also provide legal certainty, as users can rely on blanket licenses rather than uncertain fair use defenses.

From a distributional perspective, collective licensing can support creators who lack bargaining power in individual negotiations [17]. By pooling rights, CMOs enhance leverage and facilitate revenue flows that might otherwise be foregone. In sectors such as music and broadcasting, collective licensing has become an indispensable infrastructure.

However, collective regimes also exhibit weaknesses. Concentration of rights aggregation can create market power, raising concerns about excessive tariffs or exclusionary practices [18]. Transparency deficits and governance failures have, in some cases, undermined trust among rights holders. Distribution methodologies may disadvantage certain creators, particularly where usage data is imperfect.

Another limitation lies in adaptability. Collective regimes developed for specific industries may struggle to accommodate novel uses or technologies without legislative intervention [16]. This rigidity can slow innovation or produce regulatory lag.

These strengths and weaknesses offer critical lessons for AI training contexts. While collective licensing promises scalability and certainty, its success depends on careful design to mitigate risks of overreach, inefficiency, and inequitable outcomes [20].

Table 1: Comparison of Major Collective Licensing Models and Regulatory Features

Collective Licensing Model	Typical Sectors of Use	Legal Basis	Rights Aggregation Method	Tariff Setting Mechanism	Opt-Out Availability	Regulatory Oversight	Key Strengths	Key Limitations
Voluntary Collective Licensing	Music performance, broadcasting	Contractual mandates from rights holders	Individual rights holders voluntarily assign rights	Negotiated tariffs set by CMO	Yes	Competition law, general corporate regulation	High flexibility; respects author autonomy	Limited coverage; fragmentation risk
Extended Collective Licensing (ECL)	Education, libraries, reprography	Statutory extension of collective agreements	Mandatory extension to non-members with opt-out	Tariffs approved or reviewed by regulator	Yes	Dedicated copyright authority or tribunal	Broad coverage; reduced transaction costs	Requires strong legitimacy and oversight
Compulsory Licensing	Broadcasting, retransmission, cable	Statutory mandate	Rights automatically licensed by law	Rates set by statute or regulator	No	High regulatory control	Legal certainty; universal access	Weakens exclusivity; inflexible pricing
Levy-Based Collective Systems	Private copying, media devices	Statutory levy schemes	Indirect aggregation via levies on devices/media	Levy rates set by legislation or regulator	No	Government supervision	Simple administration; predictable revenue	Weak link between use and remuneration
Hybrid Collective Models	Digital copying, online uses	Mixed statutory and contractual basis	Combination of mandates and statutory	Negotiated with regulatory backstop	Limited	Sector-specific regulatory oversight	Adaptable to new technologies	Governance complexity

Collective Licensing Model	Typical Sectors of Use	Legal Basis	Rights Aggregation Method	Tariff Setting Mechanism	Opt-Out Availability	Regulatory Oversight	Key Strengths	Key Limitations
			coverage					
Proposed AI Training Collective Licensing	AI and machine learning training	Statutory authorization with opt-out	Aggregation of training-use rights	Regulated tariffs with adaptive metrics	Yes	Specialized regulatory body		

4. DESIGNING A COLLECTIVE LICENSING FRAMEWORK FOR AI TRAINING DATA

4.1 Defining the Scope of Licensed AI Training Uses

A workable collective licensing framework for AI training must begin with a precise definition of the uses it covers. AI training involves acts of copying, storage, and computational analysis undertaken to develop or improve machine learning models, rather than to communicate expressive content to human audiences [22]. Licensing scope should therefore focus on reproductions made for training, validation, and model optimization, distinguishing these activities from downstream outputs or expressive dissemination.

Clear boundaries are essential to avoid overreach. Licensed uses should encompass ingestion of protected works into training datasets, intermediate copies created during preprocessing, and retention necessary for model refinement [19]. By contrast, outputs generated by trained models should generally fall outside the license, as they raise separate questions of authorship, infringement, and liability [24]. This distinction preserves doctrinal clarity and prevents collective licensing from becoming a de facto authorization for derivative exploitation.

Exclusions are equally important. Private, non-commercial experimentation, uses covered by existing statutory exceptions, and purely internal testing with licensed datasets may warrant differentiated treatment [21]. Tailoring scope prevents unnecessary burdens on innovation while ensuring compensation where economically significant use occurs.

Defining scope in functional rather than technological terms enhances adaptability. As training techniques evolve, focusing on the purpose and effect of use—statistical learning rather than expressive substitution—allows the framework to remain relevant [25]. A carefully delimited scope thus provides legal certainty, respects existing copyright boundaries, and anchors collective licensing as a targeted response to AI training practices.

4.2 Rights Aggregation, Opt-Outs, and Representativeness

Rights aggregation is the institutional core of collective licensing, enabling authorization of large-scale uses that individual rights holders cannot feasibly license on their own. In the AI training context, aggregation would involve creators or publishers mandating a collective entity to license specified training uses on their behalf [20]. This approach reduces transaction costs while maintaining the integrity of exclusive rights.

Preserving creator autonomy requires meaningful opt-out mechanisms. Rights holders should be able to exclude their works from collective schemes or limit participation to defined categories of use [23]. Opt-out design must balance individual choice with collective effectiveness, as excessive fragmentation can undermine representativeness and legal certainty for licensees.

Representativeness is critical to legitimacy. A collective body must encompass a sufficiently broad repertoire to justify blanket licensing and minimize the risk of unlicensed use [25]. Regulatory authorization can help ensure that collective entities meet thresholds of participation and act in the interests of the rights holder community as a whole.

Transparency and governance safeguards reinforce trust. Clear rules regarding membership, voting rights, and decision-making processes reduce concerns about capture or unequal treatment [19]. Special consideration may be required for marginalized or independent creators who lack institutional leverage.

By combining aggregation with opt-out protections and representativeness requirements, collective licensing can reconcile efficiency with autonomy. This balance is essential for deploying collective solutions in AI training without replicating the shortcomings that have undermined confidence in some existing collective regimes [22].

4.3 Remuneration Models and Economic Allocation Principles

Remuneration lies at the heart of any collective licensing framework, shaping both incentives and perceived fairness. Several models are available for compensating rights holders whose works are used in AI training. Flat-fee models offer simplicity and predictability, allowing licensees to pay standardized rates based on factors such as model size, training frequency, or organizational scale [24]. However, flat fees risk undercompensating high-value contributors or overburdening smaller developers.

Usage-based models aim for closer alignment between compensation and actual use. Metrics such as dataset composition, frequency of inclusion, or proxy indicators of contribution could inform distribution [21]. Yet precise

measurement poses technical and privacy challenges, particularly where training data pipelines are opaque.

Economic allocation must also address equity concerns. Concentration of remuneration among large rights holders may exacerbate existing inequalities in creative markets [19]. Allocation formulas should therefore incorporate mechanisms to support smaller creators and ensure broad participation benefits.

Valuation challenges are inherent. AI training extracts diffuse value from vast datasets, making it difficult to attribute model performance to individual works [25]. Regulatory oversight and stakeholder consultation can help establish reasonable proxies and adjust rates over time.

Ultimately, remuneration design should reflect the collective nature of AI training. By prioritizing fairness, transparency, and adaptability, collective licensing can provide sustainable compensation without imposing infeasible measurement burdens, aligning economic incentives with the realities of machine learning development [22].

Figure 2: *Proposed collective licensing architecture for AI training data.*

5. INSTITUTIONAL IMPLEMENTATION AND REGULATORY SAFEGUARDS

5.1 Role of the State: Authorization, Oversight, and Enforcement

The effective operation of collective licensing for AI training data depends on an active but carefully delimited role for the state. Legislatures or designated administrative agencies are uniquely positioned to authorize collective management organizations, define their legal mandates, and ensure that licensing schemes operate in the public interest [27]. Without formal authorization, collective licensing risks legal uncertainty and fragmented participation, undermining its intended scalability.

Authorization may take the form of statutory recognition, compulsory licensing frameworks, or extended collective licensing regimes that bind non-members while preserving opt-out rights [24]. Each model requires legislative clarity regarding scope, eligibility, and legal effect. Administrative agencies can further specify operational standards, approve tariffs, and resolve disputes between licensors and licensees [30].

Oversight is essential to legitimacy. Regulatory supervision can address information asymmetries, prevent self-dealing, and ensure equitable treatment of rights holders [23]. Common oversight mechanisms include reporting obligations, periodic audits, and review of governance practices. Enforcement powers, including sanctions for non-compliance, reinforce accountability and deter abuse.

The state's role is not to manage licensing directly but to create an enabling framework within which collective solutions can function credibly [29]. By providing legal recognition, monitoring performance, and enforcing compliance, public institutions help align private ordering with broader policy objectives. This regulatory involvement distinguishes collective licensing from purely contractual arrangements and anchors it within copyright's public law dimension [25].

5.2 Transparency, Auditability, and Data Governance Obligations

Transparency and auditability are foundational safeguards in any collective licensing regime, particularly where AI training processes are opaque and technically complex. Collective management organizations must provide clear information on licensing terms, tariff structures, and revenue distribution methodologies to maintain trust among rights holders [26]. Without transparency, collective systems risk legitimacy deficits and resistance from creators.

Auditability complements transparency by enabling verification of compliance and accuracy. Rights holders should have access to mechanisms that allow independent review of royalty calculations and distribution practices [30]. Similarly, regulators require audit trails to assess whether licensing bodies adhere to statutory mandates and governance standards.

AI developers participating in collective licensing schemes also bear data governance obligations. Accurate reporting of licensed use, dataset composition, and training activities is necessary to support remuneration allocation and oversight [23]. While proprietary concerns may limit disclosure, standardized reporting frameworks can balance confidentiality with accountability.

Data governance obligations extend to security and integrity. Collective entities managing usage data and financial flows must implement safeguards against misuse, unauthorized

access, and data manipulation [28]. Clear retention policies and data minimization principles further reduce risk.

By embedding transparency, auditability, and data governance into regulatory design, collective licensing can function as a trusted intermediary between creators and AI developers. These safeguards not only protect individual interests but also support systemic confidence in collective solutions as legitimate instruments of copyright governance [25].

5.3 Competition, Antitrust, and Market Power Considerations

Collective licensing inevitably raises competition and antitrust concerns due to the concentration of rights aggregation and pricing authority. When a single collective entity controls access to large repertoires, it may acquire significant market power, potentially leading to excessive tariffs or exclusionary practices [29]. Addressing these risks is essential to maintaining lawful and efficient licensing markets.

Competition law provides an important backstop. Regulatory frameworks should ensure that collective entities operate within clearly defined mandates and are subject to antitrust scrutiny where appropriate [24]. Safeguards may include non-discriminatory licensing obligations, transparent tariff methodologies, and limits on exclusive dealing arrangements.

Institutional design can further mitigate market power. Allowing multiple collective entities to operate in parallel, subject to coordination rules, may introduce competitive pressure [26]. Alternatively, regulatory oversight of tariff setting can constrain pricing abuse while preserving efficiency.

Attention must also be given to the position of AI developers. Collective licensing should not function as a barrier to entry that disadvantages smaller or emerging firms [30]. Tiered pricing structures and proportional fees can reduce exclusionary effects.

Balancing efficiency with competition is a recurring challenge in copyright governance. By integrating antitrust principles into regulatory design, collective licensing frameworks can avoid replicating monopolistic dynamics while delivering the scalability and certainty needed for AI training contexts [23].

Table 2: Regulatory Safeguards for Collective Licensing in AI Training Contexts

Regulatory Safeguard	Purpose	Application in AI Training Contexts	Responsible Authority	Risk Mitigated
Statutory Authorization	Provide legal certainty and mandate	Formal recognition of collective licensing for AI	Legislature / Copyright Authority	Legal ambiguity; unenforceable licenses
		training uses		
Defined Licensing Scope	Prevent overreach	Clear delineation between training uses and AI outputs	Regulator	Unauthorized derivative exploitation
Opt-Out Mechanisms	Preserve creator autonomy	Allow rights holders to exclude works from collective schemes	Collective Management Organization (CMO)	Compulsory use concerns
Tariff Review and Approval	Prevent excessive pricing	Regulatory oversight of licensing fees and rate-setting models	Copyright Tribunal / Regulator	Abuse of market power
Transparency Obligations	Ensure accountability	Disclosure of tariffs, distribution rules, and governance practices	CMO / Regulator	Trust deficits; opaque decision-making
Audit and Reporting Requirements	Enable verification	Independent audits of usage reporting and revenue distribution	Regulator / External Auditors	Misallocation of royalties
Data Governance Standards	Protect sensitive information	Secure handling of training-use data and reporting metadata	CMO / AI Developers	Data misuse; security breaches
Non-Discriminatory Licensing	Maintain fair access	Equal licensing terms for similarly situated AI	Regulator	Exclusionary practices

Regulatory Safeguard	Purpose	Application in AI Training Contexts	Responsible Authority	Risk Mitigated
		developers		
Competition and Antitrust Oversight	Preserve market fairness	Monitoring concentration and pricing behavior of CMOs	Competition Authority	Monopolization risks
Dispute Resolution Mechanisms	Resolve conflicts efficiently	Specialized tribunals or administrative review processes	Copyright Authority	Prolonged litigation
Periodic Review and Adaptation	Ensure long-term relevance	Scheduled reassessment of licensing scope and tariffs	Legislature / Regulator	Regulatory obsolescence
Sanctions and Enforcement Powers	Ensure compliance	Penalties for non-compliance by CMOs or licensees	Regulator / Courts	Regulatory evasion

6. COMPARATIVE AND INTERNATIONAL PERSPECTIVES

6.1 Emerging Approaches to AI Training and Copyright Worldwide

Jurisdictions worldwide have begun to respond to AI training practices through divergent regulatory and policy approaches, reflecting differing legal traditions and innovation priorities. Some systems emphasize copyright exceptions for text and data mining, permitting AI training under defined conditions while preserving opt-out rights for rightholders [31]. These approaches aim to reduce friction for research and innovation but vary significantly in scope and enforceability.

Other jurisdictions rely more heavily on judicial interpretation, assessing AI training through existing doctrines such as fair use or implied license [28]. This case-by-case method offers flexibility but generates uncertainty, particularly for cross-border AI development. In contrast, a smaller number of policy proposals explicitly recognize AI training as a compensable use, calling for licensing-based solutions that preserve author remuneration [34].

Administrative guidance has also emerged as a regulatory tool. Copyright offices and cultural agencies increasingly issue interpretive statements clarifying disclosure obligations,

authorship limits, or permissible data uses [29]. While non-binding, such guidance influences industry behavior and signals future legislative direction.

These varied trajectories demonstrate the absence of a settled global approach. As AI development accelerates, regulatory divergence risks increasing legal complexity for both creators and developers, underscoring the need for coordination mechanisms capable of operating across jurisdictions [33].

6.2 Harmonization, Fragmentation, and Cross-Border AI Development

Fragmentation in copyright treatment of AI training has significant implications for cross-border AI development. AI systems are often trained using datasets aggregated from multiple jurisdictions, each with distinct legal requirements governing reproduction and data use [30]. Divergent national rules complicate compliance and increase legal risk for globally operating developers.

Lack of harmonization may also incentivize regulatory arbitrage. Developers may locate training activities in jurisdictions with permissive regimes, potentially undermining creator protections elsewhere [35]. This dynamic can distort competition and weaken the effectiveness of national copyright policies.

Conversely, overly restrictive regimes risk isolating domestic creators and developers from global innovation ecosystems [28]. Inconsistent enforcement further exacerbates uncertainty, as rights holders struggle to assert claims against opaque, transnational training pipelines.

These challenges reveal structural limits of purely national solutions. While international copyright treaties provide baseline standards, they offer limited guidance on AI-specific uses [32]. Absent coordination, fragmentation is likely to persist, increasing transaction costs and discouraging lawful engagement.

Addressing cross-border complexity requires regulatory mechanisms that can scale beyond individual jurisdictions while respecting national legal autonomy [31].

6.3 Collective Licensing as a Coordinating Mechanism

Collective licensing offers a potential mechanism for mitigating international fragmentation by standardizing authorization and remuneration for AI training uses. By aggregating rights and providing blanket licenses, collective schemes can reduce uncertainty for developers operating across borders [34]. When coupled with reciprocal agreements between collective management organizations, licensing coverage may extend internationally.

Such coordination is not unprecedented. International music licensing and reprographic rights systems demonstrate how collective frameworks can operate transnationally while remaining grounded in national law [29]. Applying similar

principles to AI training could facilitate lawful data access while ensuring creator compensation.

Collective licensing also supports regulatory convergence without requiring uniform substantive law. Jurisdictions may adopt compatible licensing architectures tailored to domestic policy priorities while enabling interoperability [33]. This flexibility enhances political feasibility.

Although challenges remain including governance alignment and revenue distribution—collective licensing provides a pragmatic pathway for coordinating AI copyright governance internationally. It reframes AI training from a jurisdiction-specific compliance problem into a managed, compensable global activity [30].

7. NORMATIVE ASSESSMENT AND POLICY TRADE-OFFS

7.1 Innovation, Access, and Incentive Balancing

Collective licensing seeks to balance competing policy objectives at the intersection of copyright and AI development. From an innovation perspective, standardized licenses reduce legal uncertainty and transaction costs, enabling developers to access data at scale [32]. This predictability supports research, competition, and technological progress.

At the same time, collective licensing preserves incentives for creative labor by recognizing AI training as a compensable use [35]. By distributing remuneration across rightholders, it mitigates the risk that large-scale data extraction erodes the economic foundation of creative industries.

Access considerations further support collective solutions. Blanket licenses expand lawful access while avoiding exclusionary effects inherent in individual bargaining [28]. When carefully designed, collective licensing aligns innovation incentives with sustainable cultural production, advancing copyright's dual objectives of progress and protection [34].

7.2 Risks, Objections, and Alternative Regulatory Paths

Despite its advantages, collective licensing faces objections. Critics argue that mandatory or extended collective schemes resemble compulsory licensing, potentially undermining exclusive rights [31]. Others highlight administrative complexity, governance risks, and the possibility of rent-seeking behavior by collective entities [29].

Alternative approaches include expanded statutory exceptions, opt-out data mining regimes, or reliance on fair use adjudication [33]. While these models reduce regulatory overhead, they often fail to address compensation and long-term incentive concerns.

No single solution is without trade-offs. The normative challenge lies in designing collective licensing frameworks that minimize coercion, ensure accountability, and remain adaptable [30]. When embedded within robust oversight

structures, collective licensing represents a balanced policy response rather than an overreach.

Figure 3: *Policy trade-offs between innovation, creator compensation, and legal certainty.*

8. CONCLUSION: COLLECTIVE LICENSING AS A SCALABLE COPYRIGHT RESPONSE TO AI

The rapid expansion of data-intensive AI training has exposed structural limitations within traditional copyright law, particularly its reliance on individualized licensing, expressive-use doctrines, and ex post adjudication. This article has shown that doctrinal uncertainty, market failure, and enforcement impracticalities make existing frameworks ill-suited to govern large-scale, automated uses of copyrighted works for machine learning. At the same time, unqualified reliance on exceptions or litigation risks undermining creator incentives and generating persistent legal instability.

Drawing together doctrinal, economic, and institutional analysis, the article argues that collective licensing offers a pragmatic and balanced regulatory response. By aggregating rights, standardizing authorization, and distributing remuneration, collective licensing addresses transaction cost barriers while preserving the core normative commitment of copyright to human creativity and fair compensation. Importantly, collective models shift governance from adversarial dispute resolution toward structured, ex ante regulation capable of operating at technological scale.

The analysis further demonstrates that collective licensing is not a rigid or monolithic solution. When combined with opt-out mechanisms, regulatory oversight, transparency obligations, and competition safeguards, it can respect creator autonomy, mitigate market power, and remain adaptable to

evolving AI practices. While not eliminating all tensions between copyright and artificial intelligence, collective licensing reframes AI training as a managed, compensable activity rather than a legal anomaly. In doing so, it offers a scalable pathway for aligning innovation, cultural sustainability, and legal certainty in the age of artificial intelligence.

9. REFERENCE

1. Quang J. Does training AI violate copyright law?. Berkeley Tech. LJ. 2021;36:1407.

2. Hugenholtz PB, Quintais JP. Copyright and artificial creation: does EU copyright law protect AI-assisted output?. IIC-International Review of Intellectual Property and Competition Law. 2021 Oct;52(9):1190-216.

3. Bonadio E, McDonagh L. Artificial intelligence as producer and consumer of copyright works: evaluating the consequences of algorithmic creativity. Intellectual Property Quarterly. 2020 Apr 1;2020(2):112-37.

4. Bonadio E, McDonagh L. Artificial intelligence as producer and consumer of copyright works: evaluating the consequences of algorithmic creativity. Intellectual Property Quarterly. 2020 Apr 1;2020(2):112-37.

5. Contractor D, McDuff D, Haines JK, Lee J, Hines C, Hecht B, Vincent N, Li H. Behavioral use licensing for responsible AI. InProceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 2022 Jun 21 (pp. 778-788).

6. Kop M. AI & intellectual property: Towards an articulated public domain. Tex. Intell. Prop. LJ. 2019;28:297.

7. Eze Dan-Ekeh. DEVELOPING ENTERPRISE-SCALE MARKET EXPANSION STRATEGIES COMBINING TECHNICAL PROBLEM-SOLVING AND EXECUTIVE-LEVEL NEGOTIATIONS TO SECURE TRANSFORMATIVE INTERNATIONAL ENERGY PARTNERSHIPS. International Journal Of Engineering Technology Research & Management (IJETRM). 2018Dec21;02(12):165–77.

8. Levendowski A. How copyright law can fix artificial intelligence's implicit bias problem. Wash. L. Rev.. 2018;93:579.

9. Lim D. AI & IP: innovation & creativity in an age of accelerated change. Akron L. Rev.. 2018;52:813.

10. Margoni T, Kretschmer M. A deeper look into the EU text and data mining exceptions: harmonisation, data ownership, and the future of technology. GRUR international. 2022 Aug 1;71(8):685-701.

11. Sobel BL. Artificial intelligence's fair use crisis. Colum. JL & Arts. 2017;41:45.

12. Udeh NC. *Building sustainable SME banking strategies that expand market access, boost client retention, and support economic inclusion.*

13. Kop M. The right to process data for machine learning purposes in the EU. Harvard Law School, Harvard Journal of Law & Technology (JOLT) Volume. 2020 Jun 22;34.

14. Black J, Murray AD. Regulating AI and machine learning: setting the regulatory agenda. European journal of law and technology. 2019 Dec 30;10(3).

15. Gillotte JL. Copyright infringement in AI-generated artworks. UC Davis L. Rev.. 2019;53:2655.

16. Chen G, Xie P, Dong J, Wang T. Understanding programmatic creative: The role of AI. Journal of advertising. 2019 Aug 8;48(4):347-55.

17. Ogunyemi FM. Developing robust accounting models for quantifying scope 3 emissions and climate-related liabilities in energy-intensive corporate value chains. *Int J Financ Manage Econ*. 2019;2(2):97–105. doi:10.33545/26179210.2019.v2.i2.658

18. Veale M, Borgesius FZ. Demystifying the draft EU artificial intelligence act. arXiv preprint arXiv:2107.03721. 2021 Jul 8.

19. Bently L, Sherman B, Gangjee D, Johnson P. Intellectual property law. Oxford university press; 2022.

20. Nwenekama Charles-Udeh. Leveraging financial innovation and stakeholder alignment to execute high-impact growth strategies across diverse market environments. Int J Res Finance Manage 2019;2(2):138-146. DOI: 10.33545/26175754.2019.v2.i2a.617

21. Kemp R. Legal Aspects of Artificial Intelligence (v. 2.0). Kemp IT Law.–2016.–URL: https://www. kempitlaw. com/wp-content/uploads/2016/11/Legal-Aspects-of-AI-Kemp-IT-Law-v2. 0-Nov-2016-. pdf. 2018 Sep.

22. Bosworth D, Yang D. Intellectual property law, technology flow and licensing opportunities in the People's Republic of China. International Business Review. 2000 Aug 1;9(4):453-77.

23. Sag M. The new legal landscape for text mining and machine learning. J. Copyright Soc'y USA. 2018;66:291.

24. Eze Dan-Ekeh. Engineering high-value commercialization frameworks integrating technical innovation with strategic sales leadership to drive multimillion-dollar growth in global energy markets. World J Adv Res Rev. 2019;4(2):256-268. doi:10.30574/wjarr.2019.4.2.0152

25. Rodrigues R. Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. Journal of Responsible Technology. 2020 Dec 1;4:100005.

26. Hacker P, Passoth JH. Varieties of AI explanations under the law. From the GDPR to the AIA, and beyond. InInternational workshop on extending explainable AI beyond deep models and classifiers

2020 Jul 18 (pp. 343-373). Cham: Springer International Publishing.

27. Feyisayo Michael Ogunyemi. Green financing mechanisms for closing the trillion-dollar climate investment gap: A multi-stakeholder framework integrating public policy, private capital, and carbon accounting standards. Int J Res Finance Manage 2020;3(2):99-108.
DOI: 10.33545/26175754.2020.v3.i2a.6091 citation

28. Ebers M, Hoch VR, Rosenkranz F, Ruschemeier H, Steinrötter B. The european commission's proposal for an artificial intelligence act—a critical assessment by members of the robotics and ai law society (rails). J. 2021 Oct 8;4(4):589-603.

29. Zeilinger M. Tactical entanglements: AI art, creative agency, and the limits of intellectual property. meson press; 2021.

30. Badmus A, Adebayo M. Compliance-aware DevOps for generative AI: Integrating legal risk management, data controls, and model governance to mitigate deepfake and data privacy risks in synthetic media deployment. Journal of AI Governance. 2020;1(2):45-61.

31. Omogiate PM, Legal CA. Developing standardized metadata protocols enabling transparent provenance tracking for AI-created media within federal intellectual property regulatory systems nationwide. International Journal of Computer Applications Technology and Research. 2022;11(12):711-23.

32. Paullada A, Raji ID, Bender EM, Denton E, Hanna A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns. 2021 Nov 12;2(11).

33. Bulayenko O, Quintais JP, Gervais DJ, Poort J. Ai music outputs: Challenges to the copyright legal framework. Available at SSRN 4072806. 2022 Feb 28.

34. Feyisayo Michael Ogunyemi. Integrated ESG-financial risk accounting: A framework for embedding ESG metrics into enterprise risk management and financial statements. Int J Res Hum Resour Manage 2021;3(2):144-154.
DOI: 10.33545/26633213.2021.v3.i2b.386

35. Hacker P. A legal framework for AI training data—from first principles to the Artificial Intelligence Act. Law, innovation and technology. 2021 Jul 3;13(2):257-301.