

DOI: <https://doi.org/10.47344/bbzgtm81>

IRSTI: 10.41.29

## The Rise of AI Training Data Licensing: Reassessing Fair Use and Market Harm

Aisha Zhaksylyk

SDU University, Kaskelen, Kazakhstan

email: [251407001@sdu.edu.kz](mailto:251407001@sdu.edu.kz)

### Abstract

The unlicensed use of copyrighted works to train generative artificial intelligence (AI) systems is among the most contested questions in contemporary copyright law. Courts and scholars remain divided on how the fourth fair use factor should apply to a process spanning data ingestion, model training, and commercial deployment. This article asks how the Factor Four market harm inquiry should be structured as training data licensing markets emerge, and what this means for Kazakhstan, which has adopted an AI training opt-out rule without a developed doctrinal framework. Using doctrinal analysis of United States and European Union copyright law, a comparative method, and a critical synthesis of three scholarly positions, the article reexamines six recent decisions including *Andy Warhol Foundation v. Goldsmith*, *Kadrey v. Meta*, and *Thomson Reuters v. ROSS Intelligence* for what each actually decided. It finds that existing approaches apply one standard of scrutiny to a use occurring in two doctrinally distinct phases, and proposes a phase-sensitive framework: a permissive test for upstream training paired with heightened scrutiny for downstream deployment, supplemented by collective licensing and transparency obligations. Applying this framework to Kazakhstan, the article evaluates Article 23(5) of the Law of the Republic of Kazakhstan “On Artificial Intelligence” and proposes an institutional design built on existing collective-management infrastructure.

**Keywords:** AI training, copyright law, fair use, market harm, licensing.

### Introduction

Generative artificial intelligence (AI) systems are trained by processing very large quantities of text, images, audio, and video, a substantial share of which is protected by copyright and was never licensed for this purpose (Torrance & Tomlinson, 2023). Whether that processing infringes the reproduction right, and whether it can be excused as fair use, has become a central question in contemporary copyright law (Samuelson, 2023). The dispute is no longer only about principle. As dedicated markets for licensing training data have begun to form, the fourth fair use factor, which asks whether a challenged use harms “the market for or value of” the copyrighted work, has moved from a subordinate consideration to the doctrinal battleground on which most AI training disputes are actually fought (Hardman, 2026; Rosati, 2025). An emerging licensing market changes the market harm inquiry in a way earlier case law did not have to confront: a use that seemed purely technical can be redescribed as harmful simply because a market for licensing it now exists, and rightsholders have an obvious incentive to describe it that way (Ge, 2025; Tang, 2021).

*Research problem and thesis.* Courts and commentators have generally analyzed AI training as a single, undifferentiated act. Yet training a generative model is a sequence that begins with functional, non-expressive ingestion and processing and ends, if the model is deployed, in outputs that can compete in the same markets as the works the model was trained on (Liu, 2025; Murray,

2023). Treating this as one question produces two symmetrical errors: a blanket fair-use finding risks ignoring genuine deployment-stage substitution, while a blanket licensing requirement risks treating speculative market definitions as if they were the non-speculative markets Factor Four was designed to protect: the “bootstrapping” problem (Ge, 2025; Tang, 2021; Samuelson, 2024). This article’s central thesis is that existing fair use and text-and-data-mining (TDM) doctrine in the United States (U.S.) and the European Union (EU) apply a single standard to a use that in fact occurs in two doctrinally distinct phases, and that Kazakhstan’s newly enacted machine-readable opt-out in Article 23(5) of the Law “On Artificial Intelligence,” which closely resembles the EU’s Article 4(3) rights-reservation mechanism, has adopted a partial solution without the doctrinal or institutional apparatus needed to make it fully workable.

*Research question and objectives.* This article asks: how should the Factor Four market-harm inquiry be structured to distinguish AI training uses that cause cognizable market harm from those that do not, and what framework should guide Kazakhstan’s own AI training exception? It pursues this by:

- a) establishing, through case-law analysis, what U.S. courts actually decided, as distinct from what commentary often attributes to those decisions;
- b) developing and testing a phase-based framework against the leading positions in the literature;
- c) translating it into institutional recommendations for Kazakhstan.

*Research gap and contribution.* The literature addresses the doctrinal content of Factor Four, the TDM exception, the bootstrapping critique, and licensing and transparency proposals, but has not combined these into a single operational test, nor connected one to the regulatory choices a civil-law jurisdiction such as Kazakhstan must make; Kazakhstani scholarship, in turn, has addressed AI-generated-output authorship and general AI governance (Akshalova & Muhammad, 2024; Aronov & Idrysheva, 2024; Iskakova & Kashkin, 2020; Temirbekov, 2021) but not the input side of the problem. I make three contributions. I reframe the literature’s disagreement as one about which pipeline stage each position implicitly describes; I propose a phase-based framework – a permissive, evidence-based standard for upstream use and heightened scrutiny for downstream use, paired with collective licensing, transparency, and provenance obligations as governance mechanisms; and I apply that framework to Article 23(5), correcting a recurring imprecision in how the provision is described and proposing an institutional design built on Kazakhstan’s existing collective-management infrastructure, not a new regulator.

Section 2 details the methods used. Section 3 sets out the legal framework. Section 4 critically reviews three scholarly approaches. Section 5 develops the comparative analysis. Section 6 analyzes six recent decisions for what each actually held. Section 7 sets out the phase-based framework I propose. Section 8 applies it to Kazakhstan and Section 9 discusses implications and counterarguments. Section 10 concludes.

## Methodology

This is a doctrinal and comparative legal research article. It does not involve empirical data collection, interviews, or statistical analysis. It combines four methods: doctrinal analysis, comparative legal analysis, critical literature synthesis, and structured case-law analysis. Kazakhstan is not a fourth, self-contained case study added at the end. It is the point toward which the first three methods are directed.

### 1. Doctrinal analysis

The doctrinal component examines the primary sources defining the boundaries of permissible AI training: the fair use provision of the U.S. Copyright Act, 17 U.S.C. § 107; the TDM exceptions in Articles 3 and 4 of Directive EU 2019/790, including the Article 4(3) opt-out; and Articles 22 and 23 of the Law of the Republic of Kazakhstan “On Artificial Intelligence” (Law No. 230-VIII, 2025), the first addressing machine-readable forms generally and the second addressing copyright specifically. It tests whether recurring concepts such as transformativeness, market substitution as opposed to speculative licensing harm, and the “bootstrapping” problem, whereby a rightsholder’s assertion of a licensing market is used to manufacture the harm that then defeats a fair use defense, operate consistently across the training and deployment stages, the distinction on which the article turns.

## 2. *Comparative legal analysis*

The United States and the European Union are compared against six criteria: (1) the doctrinal mechanism used to permit or restrict AI training; (2) how the market harm inquiry is structured; (3) the design of any opt-out or rights reservation mechanism; (4) the treatment of licensing markets; (5) transparency and provenance obligations; and (6) the responsible institutional body. These jurisdictions represent the two structurally distinct regulatory models available, and most litigated case law has arisen in one or the other; the comparison is accordingly selected for what it can transplant to Kazakhstan, not exhaustive, and Section 5 draws recommendations directly from the comparative table.

## 3. *Literature synthesis and the three scholarly approaches*

The synthesis draws on articles, working papers, and legal commentary published primarily in 2021-2026, supplemented by earlier fair use scholarship that remains doctrinally relevant. Sources were selected for direct engagement with Factor Four as applied to AI training, the TDM exception, licensing market bootstrapping, or collective licensing and transparency proposals; general AI law scholarship without a copyright argument was excluded, as was Kazakhstani scholarship on AI generated output authorship, which addresses a different question (ownership of output) than the one I address (lawfulness of input). The literature is sorted into three approaches using five criteria: characterization of training as expressive or non-expressive; weight given to transformative use reasoning; treatment of Factor Four; position on bootstrapping; and preferred regulatory solution. An author is placed in the innovation-oriented category where most criteria favor functional characterization (e.g., Liu, 2025; Torrance & Tomlinson, 2023); in the rightsholder oriented category where they point the other way (e.g., Rosati, 2025; Opderbeck, 2024); and in the nuanced category where a position diverges by pipeline stage (e.g., Fernandes, 2024; Hardman, 2026; de la Durantaye, 2025). The camps are not internally uniform.

## 4. *Case-law methodology*

Six decisions (Section 6) were selected because each addresses a different point on the training to deployment spectrum, from a foundational non-AI transformative use precedent through the first LLM training merits ruling to an ongoing dispute over downstream substitution. Each is doctrinally important in a different way: *Warhol* supplies the post-2023 framing of Factors One and Four; *Authors Guild* the functional-copying analogy; *Kadrey* the first merits ruling on LLM training; *ROSS* the clearest case of direct market substitution; *Getty* the limits of the “model as copy” theory; and the *New York Times* litigation the leading downstream-substitution claim. Cases concerning only output-similarity without a developed market-harm holding (e.g., *Andersen v. Stability AI*) were excluded from detailed analysis and are noted only in passing. Each selected case is analyzed for its facts, legal issue, actual holding, relevance to fair use and market harm, relevance to AI training, and the limits of the analogy, so cases are compared on equal terms.

*Kazakhstan's role in the comparative method*

Kazakhstan enters the analysis at three points. In the comparative table (Section 5), Article 23(5) is placed alongside the U.S. and EU mechanisms on each criterion, showing where it resembles the EU model and where its connection to the Law's own general provisions, such as Article 22, remains unconfirmed. In the case-law analysis (Section 6), each case's limits identify which doctrinal assumptions cannot be assumed to transfer to Kazakhstan. The phase-based framework (Section 7) is built so its upstream/downstream distinction can be implemented through an opt-out mechanism of the kind Kazakhstan has already adopted. Section 8 then addresses the specific institutional questions such as standardization of the opt-out format, the responsible authority, the role of collective management organizations (CMOs), and the treatment of foreign developers that this analysis is designed to inform.

## Legal Framework: AI Training, Copyright, and the Structure of Fair Use

*The training process and the stages of copying*

A generative AI model is built by exposing a machine-learning architecture to a large corpus of data so it can adjust internal parameters and later generate new content (Liu, 2025). Three stages matter for copyright analysis: **ingestion**, collecting and reproducing data into a working corpus (Rosati, 2025; Opderbeck, 2024); **storage**, retaining works or intermediate representations for processing; and **processing**, in which the model's parameters are updated as numerical weights encoding statistical regularities, not retrievable copies (Murray, 2023; Quang, 2021). Memorization complicates this picture. A model can sometimes reproduce near-verbatim fragments, which is why courts have treated the possibility of extraction, not only ingestion, as relevant to market harm. Every stage involves reproduction within the meaning of copyright law, even where intermediate (Rosati, 2025; Opderbeck, 2024). The decisive question is not whether copying occurs, but whether an exception excuses it.

*Text and Data Mining exceptions and the limits of the TDM analogy*

Text and Data Mining (TDM), as codified in Articles 3 and 4 of Directive EU 2019/790, refers to techniques for extracting facts, patterns, or correlations from data sets for analytical purposes, without producing new expressive content competing with the works mined. Article 3 provides a mandatory, non-waivable exception for research organizations (Directive EU 2019/790, art. 3); Article 4 provides a broader exception for any user, subject to a rightsholder's ability to reserve rights "in an appropriate manner," including machine-readable means for online content (Directive EU 2019/790, art. 4(3)). This is the mechanism Kazakhstan's Article 23(5) closely resembles. Generative AI training strains this framework because it is not purely analytical. The resulting model is built to generate new instances resembling the training data's styles and structures (Dornis & Stober, 2025; Ge, 2025). That is why a substantial body of scholarship questions whether Article 4 was ever meant to reach commercial foundation-model training (Buick, 2025; Rosati, 2025).

*Fair use and the structure of Factor Four*

Under 17 U.S.C. § 107, courts weigh four non-exclusive factors: (1) the purpose and character of the use; (2) the nature of the work; (3) the amount used; and (4) the effect on the potential market for or value of the work (17 U.S.C. § 107(1) (4)). Factor Four § 107(4) has increasingly operated as the point on which AI training disputes turn, since it is where arguments about licensing markets and substitution are channeled (Hardman, 2026). If we treat copying as

harmful merely because a license could have been bought, we are no longer finding fair use; we are assuming it away (Tang, 2021; Samuelson, 2024). A related trap is bootstrapping: because rightsholders can create a licensing market simply by offering one, allowing that market alone to defeat fair use lets rightsholders define their way out of any exception (Ge, 2025; Tang, 2021). Both points are settled outside the AI context, but their application to AI training is what recent litigation has had to work out case by case (Section 6).

*Andy Warhol Foundation v. Goldsmith* (2023), though not an AI case, supplies the doctrinal backdrop for this litigation by reorienting Factors One and Four toward market proximity and commercial purpose. Every subsequent AI training decision has had to address that shift and to state its limits.

## Literature Review

This research builds on the U.S. Copyright Act, Directive EU 2019/790, and case law on transformativeness, market substitution, and the boundaries of TDM exceptions. Kazakhstani scholarship has developed along a different axis, addressing AI's general legal personality and governance (Temirbekov, 2021; Iskakova & Kashkin, 2020) and the authorship of AI-generated output (Akshalova & Muhammad, 2024; Aronov & Idrysheva, 2024), but not the input side of the problem I examine. That is the gap Section 8 builds on directly.

Beyond Kazakhstan, the literature organizes into three positions, sorted using the criteria set out in the Methodology. The three positions do not converge, and I preserve rather than resolve their disagreement here, because Section 7 argues that each camp is strongest precisely where it describes its own stage of the pipeline.

**The innovation-oriented approach** recharacterizes training related copying as upstream, functional processing rather than exploitation of expression, treating *Authors Guild v. Google* (2015) as a template for the ingestion and processing stages and insisting that Factor Four requires evidence of cognizable substitution, not a bare foregone-licensing claim (Liu, 2025; Torrance & Tomlinson, 2023; Rodriguez Maffioli, 2023; Murray, 2023). It is the camp most committed to the bootstrapping critique (Ge, 2025; Tang, 2021; Hardman, 2026), and *Kadrey v. Meta* (2025), discussed in Section 6, lends it some support by treating the plaintiffs' failure to prove market dilution as decisive. Where it has least to say is the deployment stage, where outputs can plausibly compete with the works trained on. It acknowledges the input/output distinction but does not develop a correspondingly rigorous test for when output-stage substitution becomes cognizable, and it struggles to explain cases of direct commercial competition such as *Thomson Reuters v. ROSS Intelligence* (2025). Applied to Kazakhstan without qualification, it would counsel against giving much weight to Article 23(5) at the pre-training stage but offers no guidance on deployment by commercial products built on Kazakh-language data.

**The rightsholder-oriented approach** treats training-related copying as squarely within the reproduction right, arguing that TDM style exceptions were built for analytical, non-generative uses and do not comfortably extend to systems trained to produce competing expressive output, and that both lost sales and lost training-data licensing revenue should be cognizable harm (Rosati, 2025; Opderbeck, 2024; Fernandes, 2024; Dornis & Stober, 2025). It is best supported in cases of direct commercial competition, such as *ROSS*, and takes seriously *Warhol's* insistence that commercial purpose and market proximity matter independently of transformativeness. The bootstrapping problem is what this approach never fully resolves. If a licensing market that rightsholders can

create unilaterally is itself enough to establish harm, Factor Four stops doing any independent analytical work. The objection matters precisely because the position is otherwise strong. An unqualified version of this position would require individually licensing training corpora that can run to billions of works, which the literature itself recognizes as unworkable at scale (Fernandes, 2024; Rodriguez Maffioli, 2023). It is the position closest in spirit to Article 23(5) of the Law of the Republic of Kazakhstan “On Artificial Intelligence”, but applied without institutional support it risks the same transaction cost problem it identifies as unworkable elsewhere.

Between these two sits a third position. **The nuanced approach** argues that the standard should track the pipeline stage: permissive treatment for non-expressive ingestion, closer scrutiny for deployment where substitution risk is concrete (Fernandes, 2024; Hardman, 2026). It treats the uncertainty evident in *Kadrey and Andersen v. Stability AI* as real rather than already resolved by existing doctrine (Guadamuz, 2024). It favors collective licensing, statutory levies, or opt-out registries paired with transparency obligations (de la Durantaye, 2025; Guadamuz, 2024), and seeks a middle path on Factor Four that separates substitutive from complementary effects (Samuelson, 2024; Buick, 2025). Only this position treats "AI training" as more than a single act. However, its diagnosis outruns its design: who would run a registry, in what format, and with what hold over foreign developers all remain open.

**The research gap.** Across all three camps, two gaps recur: none offers a specific, evidence-based test for distinguishing cognizable substitution from bootstrapped licensing revenue, and none develops a concrete institutional design for a jurisdiction legislating this question without either an open-ended fair use tradition or a mature collective management sector, Kazakhstan’s actual position. Section 7 addresses the first gap by building a phase-based test with explicit evidentiary content; Section 8 addresses the second by specifying the institutional mechanisms the literature identifies as necessary but does not itself design.

**Comparative Analysis: United States, European Union, and Kazakhstan**

**Table 1.**  
*Comparative Framework for Copyright and AI Training in the United States, European Union, and Kazakhstan*

| Criterion                                        | United States                                                                                                                                                                                      | European Union                                                                                                                                                                                                                                         | Kazakhstan                                                                                                                                                                                                                                |
|--------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>General legal approach and core exception</b> | Open-ended fair use, 17 U.S.C. § 107; no AI-specific statute; assessed case by case under Factors One and Four ( <i>Kadrey v. Meta</i> , 2025; <i>Thomson Reuters v. ROSS Intelligence</i> , 2025) | Enumerated TDM exceptions: Art. 3 (research organizations, non-waivable) and Art. 4 (general purpose, waivable by opt-out), layered with general-purpose AI (GPAI) duties under the AI Act (Directive EU 2019/790, arts. 3-4; Regulation EU 2024/1689) | New sui generis provision: use of works to train AI models is permitted absent a machine-readable prohibition (Law No. 230-VIII, art. 23(5)), operating alongside, but not integrated into, the general Copyright Law (Law No. 6-I, 1996) |
| <b>Market harm analysis</b>                      | Factor Four; post- <i>Warhol</i> emphasis on                                                                                                                                                       | Not framed as “market harm”; addressed                                                                                                                                                                                                                 | No statutory market-harm test; Article 23(5)                                                                                                                                                                                              |

|                                     |                                                                                                                                                                 |                                                                                                                                                                 |                                                                                                                                                                                                                                                                    |
|-------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                     | commercial purpose and market proximity; U.S. Copyright Office (2025) treats market dilution and lost licensing opportunity as cognizable but context-dependent | indirectly through the opt-out's availability rather than a substitution test (Directive EU 2019/790, art. 4(3))                                                | turns solely on the presence or absence of an opt-out, not on evidence of harm                                                                                                                                                                                     |
| <b>Licensing mechanisms</b>         | Emerging voluntary, individually negotiated licenses; U.S. Copyright Office (2025) favors organic market development                                            | No AI-specific collective licensing regime; general collective management is governed by Directive 2014/26/EU, which does not extend to training-data licensing | CMOs exist for music and other repertoires (Law No. 6-I, arts. 43, 46-2), but none currently administers AI training-data licensing                                                                                                                                |
| <b>Opt-out mechanism</b>            | None at the federal statutory level                                                                                                                             | Art. 4(3): rights reservation "in an appropriate manner," including machine-readable means for online content                                                   | Art. 23(5): prohibition expressed "in machine-readable form"; Art. 22 delegates the machine-readable-forms procedure generally to the authorized body, but the Law does not confirm this extends to the Art. 23(5) signal specifically (Law No. 230-VIII, art. 22) |
| <b>Transparency obligations</b>     | No federal mandate; California's AB 2013 requires disclosure of copyrighted training content, effective January 1, 2026 (Cal. Assemb. B. 2013, 2024)            | AI Act Art. 53(1)(c)–(d): GPAI providers must maintain a copyright-compliance policy and publish a training-content summary (Regulation EU 2024/1689, art. 53)  | No comparable obligation; Art. 22's procedure serves transparency purposes generally but does not require training-data disclosure                                                                                                                                 |
| <b>Institutional implementation</b> | Federal courts; U.S. Copyright Office issuing non-binding reports (U.S. Copyright Office, 2025)                                                                 | European Commission/AI Office administering GPAI obligations; national courts applying Directive EU 2019/790                                                    | The authorized body determines the machine-readable-forms procedure generally (Law No. 230-VIII, arts. 1, 22); a separate authorized body accredits CMOs (Law No. 6-I, art. 46-2);                                                                                 |

|                               |                                                                                                        |                                                                                                                                         |                                                                                                                                                      |
|-------------------------------|--------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|
|                               |                                                                                                        |                                                                                                                                         | neither law names a single body for AI-training copyright questions specifically                                                                     |
| <b>Main legal uncertainty</b> | Whether Factor Four treats licensing-market dilution as cognizable absent direct substitution evidence | Whether Art. 4's exception, drafted for analytical extraction, covers commercial generative training at all (Buick, 2025; Rosati, 2025) | Whether Art. 22's procedure extends to the Art. 23(5) signal; scope over foreign developers; absence of a market-harm test comparable to Factor Four |

*Note.* Table compiled by the author from the primary sources cited in each cell.

Two things stand out. The United States and the European Union have converged, from opposite starting points, on the same practical need for a mechanism to signal and price consent at scale, since case-by-case adjudication and a rights-reservation exception with no institutional backing both meet the same transaction-cost problem. The EU has begun closing that gap through the AI Act's transparency duties; the United States has not yet done so by statute, leaving the Copyright Office's advisory 2025 report as the primary institutional statement apart from case law. Kazakhstan's Article 23(5) opt-out, by contrast, closely resembles the EU's Article 4(3) mechanism in doctrinal form, and Article 22 addresses technical standardization in general terms, but the Law does not confirm that this procedure covers the Article 23(5) signal specifically, and it supplies none of the remaining infrastructure the EU has found necessary: no collective-registry design, no transparency duty comparable to AI Act Article 53, and no market-harm test, since the opt-out signal alone is dispositive. A bright-line opt-out has the virtue of predictability an open-ended fair use inquiry lacks, but it currently has no answer to the deployment-stage harms U.S. and EU experience shows rightsholders actually litigate over – the finding Sections 7 and 8 build on directly.

## Case Law and Factor Four

These six decisions do not establish a single rule on AI training and copyright.

### 1. *Andy Warhol Foundation v. Goldsmith*

Lynn Goldsmith photographed Prince in 1981; after Prince's death, the Andy Warhol Foundation licensed a work from its Warhol series based on that photograph to Condé Nast for a magazine cover, without a further license from Goldsmith. The Supreme Court held that Factor One weighed against the Foundation because this 2016 licensing served substantially the same commercial, editorial purpose as Goldsmith's photograph, which she had also licensed. Transformative character must be assessed against the specific challenged use, not the work in the abstract, and a new meaning does not outweigh a shared commercial purpose in an established licensing market. The Court was explicit that its holding addressed only the 2016 transaction, not the 1984 creation of the series (*Andy Warhol Foundation for Visual Arts, Inc. v. Goldsmith*, 2023).

*Warhol* is not an AI case; it supplies the doctrinal backdrop by reorienting Factors One and Four toward market proximity and commercial purpose. Its facts concern a single human-created

derivative work licensed once, not mass computational processing, so its reasoning transfers to the deployment stage, where outputs compete in a market the rightsholder already occupies, not to ingestion or processing, which it did not address.

## 2. *Authors Guild v. Google*

Google scanned more than twenty million books to build a full-text search index, displaying limited snippets without a license. The Second Circuit held for Google: the purpose was highly transformative because it created a research tool, not a reading substitute; snippet view prevented reconstructing a usable portion of any book. And the court found no cognizable market harm, declining to treat a hypothetical licensing fee for the indexing as the kind of harm Factor Four protects against (*Authors Guild v. Google, Inc.*, 2015).

The case is the leading authority for treating large-scale, functional copying as transformative, and its refusal to recognize a speculative licensing market bears directly on the bootstrapping problem. It is widely cited as an analogy for AI training's ingestion stage, but the analogy has a limit often elided in that commentary: Google Books did not generate new expressive content, whereas generative AI systems are trained to produce new outputs resembling the training data. It does not, and could not, resolve whether an AI model's outputs raise a separate, downstream substitution risk.

## 3. *Kadrey v. Meta Platforms*

Book authors alleged that Meta trained its Llama models on their books, obtained from shadow-library sources. In June 2025, Judge Chhabria granted Meta summary judgment on fair use: the training was highly transformative, and the models could reproduce at most a small fragment of any book, limiting substitution risk. Decisively, the plaintiffs had failed to develop a market-dilution theory that a flood of AI-generated works could suppress demand for human-authored works in a genre without word-for-word copying while Meta had shown no effect on book sales. The court was explicit that this reflected these plaintiffs' evidentiary choices, not a general holding that training on copyrighted books is lawful, and that a different record could reach a different result (*Kadrey et al v. Meta Platforms, Inc.*, 2025).

*Kadrey* is the clearest illustration that Factor Four now turns on the quality of the market-harm evidence before the court. Its market-dilution theory, in which harm does not depend on literal output substitution, directly informs the evidentiary threshold Section 7 proposes for Phase 2. The ruling is narrow and fact-bound, and I do not cite it as authority that AI training is generally fair use.

## 4. *Thomson Reuters v. ROSS Intelligence*

ROSS Intelligence, building a competing legal-search product after Thomson Reuters refused it a license, obtained roughly 25,000 "Bulk Memos" derived from Westlaw headnotes from a third party and used them for training. On renewed summary judgment, Judge Bibas held all four factors favored Thomson Reuters: the use was commercial and not transformative because ROSS's tool returned the same kind of answers the headnotes provide, making it a direct market substitute, and the intermediate character of the copying did not save the use once the ultimate product competed directly with Westlaw. The court explicitly noted this was not a generative AI case, since ROSS's tool retrieved existing text instead of generating new content (*Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc.*, 2025).

ROSS is the strongest authority that direct commercial competition combined with an actual data-licensing market defeats fair use regardless of how intermediate the copying appears, but it is frequently cited, imprecisely, as if it resolved fair use for generative AI training generally. It did not: the parties were direct rivals building the same product from the same content, a relationship

starker than that between a foundation-model developer and the universe of authors it trained on. It supports classifying phase by the purpose and market relationship of the copying, not its position in the technical pipeline.

#### 5. *Getty Images v. Stability AI*

Getty alleged that Stability AI's Stable Diffusion model was trained, in part, on watermarked Getty images without a license, bringing claims for copyright, trade mark, and database right infringement. Getty abandoned its primary training-infringement claim before judgment because it could not establish the training occurred within the UK, a territorial limitation, not a ruling on the training's lawfulness. The secondary-infringement claim also failed: the High Court held that a model's weights encode learned statistical patterns rather than retrievable copies, so the model is not itself a copy of the training images. The court found only a narrow and, per the judgment, historic and extremely limited instance of trade mark infringement concerning legacy versions, and rejected the broader claims (*Getty Images v. Stability AI*, 2025).

The point most often lost in secondary commentary is that *Getty* does not resolve whether unlicensed training for image-generation models causes market harm. Its central claim was never decided on the merits, and its holding is confined to the "model as copy" theory under UK doctrine, with no exact U.S. or Kazakhstani equivalent; it should not be cited as authority that AI training is lawful or unlawful in general. Its institutional lesson, taken up in Section 8, is that a rule addressing the legality of training conduct (such as Article 23(5)) is distinct from one addressing whether a distributed model itself infringes.

#### 6. *The New York Times v. Microsoft Corp. and OpenAI*

The New York Times alleges that Microsoft and OpenAI used millions of Times articles without authorization to train GPT-based models, and that resulting products can reproduce Times content closely enough to substitute for visiting the original article. In March 2025 the court denied in substantial part the defendants' motions to dismiss, allowing the central copyright claims to proceed; the parties have continued to litigate through 2025 and into 2026 (*The New York Times Company v. Microsoft Corporation et al.*, 2025). In September 2026 the United States filed a Statement of Interest under 28 U.S.C. § 517 in the consolidated multidistrict proceeding, arguing that training a large language model on copyrighted text is transformative and that the market-dilution theory is doctrinally flawed (Statement of Interest of the United States, 2026). No court has ruled on the fair use merits.

This is the sharpest available example of a downstream-substitution argument, situated against an actual licensing market since other publishers have already entered direct licensing agreements. Because there is no merits ruling, the case cannot be cited for any specific legal conclusion; it illustrates the structure of a downstream-substitution argument as currently litigated, and I do not treat it as settled precedent.

Read individually, not as interchangeable "AI cases," these six decisions point toward a distinction rather than a single rule: fair use or something close to it where the use is functional and unaccompanied by substitution evidence (*Authors Guild v. Google*; *Kadrey* on its record), and findings against the secondary user where the use competes directly (ROSS) or an established licensing market already exists (*Warhol*). *Getty* and the pending *New York Times* litigation, the two most often mischaracterized publicly, have on close reading decided far less than commonly assumed. I build the framework in Section 7 on this pattern.

### Results: The Proposed Phase-Based Framework

Sections 5 and 6 point toward a framework that treats the two phases differently. I offer it as my own proposal: the case law supports a phase-sensitive distinction, but no court has established one.

*One standard, two phases*

As Sections 4 and 6 show, the literature's three camps disagree less about fair use in the abstract than about which pipeline stage they implicitly describe, and the case law's apparent inconsistency tracks the stage of use before each court.

*Phase 1: pre-training and upstream use*

Phase 1 covers ingestion, storage, and processing undertaken to build a model's parameters, not to generate market-facing content. I propose that, consistent with *Authors Guild v. Google* and *Kadrey*'s treatment of training as highly transformative, ingestion and processing be treated as presumptively lawful, subject to two conditions. The use must not exceed what is technically necessary to train the model, and it must respect a properly implemented opt-out where one exists. Under this proposal, Factor Four plays a minimal role at Phase 1: only a demonstrated, non-speculative effect on an existing training-data licensing market should override the presumption, to avoid the bootstrapping problem identified in Sections 3 and 4.

*Phase 2: deployment, output, and downstream use*

Phase 2 begins once a model is deployed commercially and can generate outputs offered to end users, where market proximity becomes the central question, as *Warhol* required for Factor One and *ROSS* for Factor Four. I would evaluate Phase 2 uses under a heightened standard requiring one of three showings before market harm is treated as cognizable:

- a) demonstrated output substitution, as alleged in the pending *New York Times* litigation;
- b) established market dilution, the theory *Kadrey* treated as viable but unproven;
- c) determined direct competition, as in *ROSS*.

The deliberate effect is to exclude a fourth: the bare fact that a license could, in principle, be sold, which every camp in the literature including the rightsholder-oriented one recognizes as an unresolved bootstrapping risk if treated as sufficient on its own.

*ROSS* also shows that phase classification is functional, not chronological: copying that looks intermediate but is undertaken specifically to build a directly competing product should be classified as Phase 2 from the outset, regardless of its position in the technical pipeline.

*Horizontal governance mechanisms*

A phase distinction is administrable only if paired with mechanisms reducing transaction costs and evidentiary uncertainty. A **machine-readable opt-out**, the mechanism the EU has built into Article 4(3) and Kazakhstan into Article 23(5), lets rightsholders remove works from the Phase 1 presumption without negotiation; Section 8 addresses why Kazakhstan's Article 22 procedure has not yet been confirmed to extend to this signal. **Collective licensing** offers a realistic vehicle for Phase 2 licensing at scale, since individual licensing of the volume of works involved is not administratively feasible. The rightsholder-oriented and nuanced camps agree on that point, even as they disagree on how much weight licensing should carry. **Transparency and provenance obligations**, modeled on AI Act Article 53(1)(c)-(d), supply the evidentiary infrastructure the Phase 2 test requires. **Evidence requirements tied to the three showings above** give both parties a concrete evidentiary target, reducing the kind of underdeveloped record decisive against the plaintiffs in *Kadrey*.

*Measured against the alternatives*

Unlike an unqualified innovation-oriented rule, the framework I propose accounts for *Warhol* and *ROSS*; unlike an unqualified rightsholder-oriented rule, it avoids extending the bootstrapping problem to a stage where no market-facing product yet exists. It also improves on the EU's Article 4 opt-out standing alone, which has no answer to Phase 2 substitution or dilution questions, and on Article 23(5) as currently drafted, which gives no weight to deployment-stage harm and places the entire burden of protection on rightsholders with no corresponding developer disclosure obligation. Pairing the opt-out with Phase 2 scrutiny and transparency obligations addresses these gaps without abandoning the predictability an opt-out rule provides.

## Implications for Kazakhstan

The comparative and case-law analysis in Sections 5 and 6 was constructed to answer institutional questions Kazakhstan's legislature has raised but not yet resolved.

### *Two statutes and what Article 23(5) actually says*

Kazakhstan regulates this problem through two instruments drafted separately. The Law "On Copyright and Related Rights" (Law No. 6-I, 1996) supplies the reproduction right and, under Article 43, the statutory basis for collective management of property rights; Article 46-2 governs the voluntary accreditation of CMOs. The Law "On Artificial Intelligence" separately addresses, in Article 23, copyright questions raised by AI systems, including, in paragraph 5, the use of works to train AI models; Article 22 delegates to the authorized body the procedure for machine-readable forms in the field of AI generally (Law No. 230-VIII, 2025). These statutes are not formally integrated, and even within the 2025 Law, Article 22's procedure is not expressly linked to the Article 23(5) signal. Kazakhstan therefore does not need new institutions. It needs to connect an existing training-data rule, standardization procedure, and collective-management infrastructure that currently operate without reference to one another.

Article 23(5) provides that "the use of works for training artificial intelligence models is permitted only in the absence of a prohibition, expressed in machine-readable form, by the author or the rightholder." It applies to works generally, not to "derivative works" specifically; it closely tracks Article 4(3) of Directive EU 2019/790, without being copied from it. Read against Section 7's framework, Article 23(5) is a workable Phase 1 mechanism, but it says nothing about Phase 2: a model trained in full compliance could still be deployed in ways generating output substitution, dilution, or direct competition. Compliance should therefore be treated as necessary but not sufficient for lawfulness, leaving deployment-stage harm to a separate Phase 2 inquiry.

### *Standardizing the opt-out*

An opt-out is meaningful only if technically implementable across rightsholders from large publishers to individual authors with no technical infrastructure: it should use a simple, machine-parseable protocol interoperable with approaches emerging around the EU's Article 4(3); it should be assertable collectively through a CMO-maintained registry; and it should attach to a work through portable metadata, not the domain it is scraped from, since Kazakh-language content is frequently mirrored across platforms the author does not control.

Article 22 already assigns this procedure to "the authorized body in the field of artificial intelligence" (Law No. 230-VIII, arts. 1, 22), a body the Law defines functionally. It is presumptively the Ministry of Artificial Intelligence and Digital Development, established in 2025, though the Law does not name that ministry by title. The existing framework could provide a basis for closing the remaining gap: that body could exercise its Article 22 mandate to confirm, by

implementing regulation, that the procedure extends to Article 23(5) and publish a technical specification comparable to the AI Act Article 53(1)(d) template, while the body accrediting CMOs under Article 46-2 could certify that a registry-based mechanism satisfies it, subject to appropriate implementing rules and, if necessary, targeted legislative clarification.

*Collective management organizations, foreign developers, and transparency*

CMOs could, without new legislation beyond an implementing regulation under Article 43, operate the opt-out registry, negotiate standardized Phase 2 licensing terms (Section 4), and serve as an evidentiary intermediary for the market-dilution and substitution showings proposed in Section 7, since an organization aggregating data across its membership is better positioned than an individual author to demonstrate a market-wide effect.

Most foundation-model training occurs outside Kazakhstan, by developers with no local presence. This is the same territorial problem that caused Getty's UK training claim to fail on jurisdictional grounds. Two more realistic points of leverage exist instead. First, since Article 23(5)'s permission is conditioned on the absence of an opt-out, a developer that trains without querying a registry cannot claim the provision's benefit for works that were opted out, since the condition attaches to the work, not the training location. Second, deployment-stage obligations are more enforceable. A developer offering a commercial product to Kazakhstani users can be required to demonstrate compliance with transparency obligations modeled on AI Act Article 53(1)(c)-(d): a copyright-compliance policy and a training-content summary in a format issued by the authorized body under its Article 22 mandate, supplying the evidentiary basis the Phase 2 test requires.

*Remaining practical problems*

Three problems remain. There is no reliable way, with current technical means, to verify that a foreign-trained model actually queried a registry before training, which is a further reason to place enforcement leverage at deployment. Because accreditation under Article 46-2 is voluntary, not every collective management organization holds it, so the opt-out function would need to be consolidated within accredited organizations to avoid a fragmented registry. And because the Law already permits training absent an opt-out, any later standardization would need transitional provisions for models already trained, an issue my copyright-focused analysis cannot resolve.

In summary, I recommend:

1. Implementing regulation confirming that Article 22 extends to the Article 23(5) signal and specifying a standardized opt-out format;
2. Authorization for accredited CMOs to operate registries and negotiate Phase 2 licensing under their Article 43 functions;
3. A transparency and compliance-disclosure requirement modeled on AI Act Article 53;
4. Confirmation that Article 23(5) compliance governs training lawfulness but not deployment-stage market-harm claims;
5. A preserved, non-waivable research exception paralleling Article 3 of Directive EU 2019/790.

None requires a new regulatory body; each attaches a previously unconfirmed extension to a mandate that already exists (Turetskiy, 2025).

## Discussion

### *a) More than a policy preference*

Section 6's pattern supports a claim stronger than "a phased approach seems reasonable". Courts can be read as already applying something like a phase distinction without naming it, since

the four decisions turning on ingestion versus deployment (Section 6) align consistently with that boundary. This doctrinal continuity across different factual contexts is evidence the distinction tracks something the case law treats as significant, though it remains my own synthesis rather than a rule any court has adopted.

*b) Two counterarguments and their limits*

The most serious counterargument is that a broader ex ante licensing requirement would protect creators more reliably than a permissive Phase 1 standard depending on litigation-stage evidence. That concern has real force, since building an evidentiary record is costly and, as *Kadrey* shows, can fail even where a plausible theory of harm exists. The response is not that this concern is misplaced, but that requiring individual licenses for the full universe of training works would relocate the transaction-cost problem instead of eliminating it, risking the same market-concentration outcome the innovation-oriented camp warns against (Section 4). The reverse concern is equally real. An unqualified fair-use presumption risks ignoring the downstream substitution *Warhol* and *ROSS* recognized as significant. Section 7's framework does not split the difference. It assigns each concern to the pipeline stage where it is decidable. The collective-registry mechanism addresses consent at Phase 1 without individual licensing, while the three Phase 2 showings, made provable through transparency obligations, address speculative harm without ignoring genuine substitution.

*c) Implications for Kazakhstan and what remains open*

For Kazakhstan, Article 23(5), evaluated against Phase 1 concerns alone, is a defensible design choice, not the underdeveloped provision it is sometimes taken to be. Its real deficiency is the absence of a Phase 2 layer and of confirmation that the institutional infrastructure the Law already contains in general form, such as the Article 22 procedure and the Article 46-2 accreditation mechanism, has been connected to the Article 23(5) signal specifically. This reframes the legislative task facing Kazakhstan as completing Article 23(5), not replacing it, in the ways detailed in Section 8.

I do not claim that the three Phase 2 showings are exhaustive, or that courts applying them would reach uniform results; *Kadrey* itself shows how much turns on the adequacy of a specific evidentiary record. Nor do I claim that a machine-readable opt-out, however well standardized, can solve the identification problems that make individual consent difficult to obtain at scale. A rightsholder who never learns a registry exists cannot use it. These are acknowledged limits of the framework, not defects to be argued away.

## Conclusion

This article asked how the fourth fair use factor should be structured to distinguish AI training uses causing cognizable market harm from those that do not, now that training-data licensing markets have begun to emerge, and what the answer means for Kazakhstan's own AI training exception. The disagreement among courts and scholars is best explained not as one about whether AI training is lawful, but as one produced by applying a single standard to a use occurring in two doctrinally distinct phases. Read individually, four of the six cases examined can be read as already tracking this distinction in substance, even though none articulates it as a general rule, while *Getty* and the pending *New York Times* litigation show how easily a holding can be overstated and how much remains unresolved.

My contribution is the phase-based framework developed in Section 7, proposed as an

analytical synthesis of the comparative and case-law analysis, not as a rule already settled in any jurisdiction's case law. Collective licensing and transparency obligations are not peripheral to it; they make the phase distinction administrable by giving rightsholders a realistic way to exercise the Phase 1 opt-out and giving both rightsholders and courts the evidentiary basis the Phase 2 standard requires. For Kazakhstan, Article 23(5) is a workable Phase 1 mechanism already in place, closely resembling the EU's Article 4(3), but it operates without confirmation that Article 22 extends to it, without a settled institutional assignment for the registry function, and without a Phase 2 counterpart; Section 8's recommendations use institutions Kazakhstan already has, though implementing regulation, and in some respects legislative clarification, would still be needed to connect them.

What remains uncertain is not a minor residue: the evidentiary threshold for market dilution remains unsettled even in the United States, as *Kadrey* shows; the territorial reach of any Kazakhstani rule over foreign-trained models raises jurisdictional problems UK courts have not resolved domestically, as *Getty* illustrates; and the pending *New York Times* litigation may yet reshape how downstream substitution is analyzed. Future research should test the proposed Phase 2 showings against evidentiary records as such litigation develops, examine whether an Article 22 implementing regulation is adopted and how it defines its scope, and assess empirically whether Kazakhstan's collective-management sector has the capacity to operate the registry function proposed here. A phase-sensitive, evidence-based framework does not remove these uncertainties; it offers a more defensible place to locate them than either an undifferentiated fair-use presumption or an undifferentiated licensing requirement.

## References

- Akshalova, R.D., & Muhammad, I. (2024). Subjects of intellectual property rights, created by artificial intelligence. *Eurasian Journal of International Law (EAJIL)*, 2024, #1, 51-58. <https://www.doi.org/10.32523/2791-0954-2024-8-1-51-58>
- Andersen et al v. Stability AI Ltd. et al. (2024) No. 3:2023cv00201 - Document 166. United States District Court northern District Of California <https://law.justia.com/cases/federal/district-courts/california/candce/3:2023cv00201/407208/166/>
- Andy Warhol Foundation for Visual Arts, Inc. v. Goldsmith, 598 U.S. (2023). Docket No.21-869. Supreme Court Of The United States. <https://supreme.justia.com/cases/federal/us/598/21-869/>
- Aronov A. K. & Idrysheva S. K. (2024). Copyright Protection On Works Generated By Artificial Intelligence. *Bulletin of Institute of Legislation and Legal Information of the RK*, №2(77)-2024, 146-157. [https://doi.org/10.52026/2788-5291\\_2024\\_77\\_2\\_146](https://doi.org/10.52026/2788-5291_2024_77_2_146)
- Authors Guild v. Google, Inc. (2015). No. 13-4829. United States Court Of Appeals For The Second Circuit. <https://law.justia.com/cases/federal/appellate-courts/ca2/13-4829/13-4829-2015-10-16.html>
- Buick, A. (2025). Copyright and AI training data – transparency to the rescue? *Journal of Intellectual Property Law & Practice*, Volume 20, Issue 3, March 2025, Pages 182-192, <https://doi.org/10.1093/jiplp/jpae102>
- Cal. Assemb. B. 2013 (Generative AI Training Data Transparency Act), 2024 Reg. Sess. (Cal. 2024) (effective Jan. 1, 2026). [https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill\\_id=202320240AB2013](https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2013)

- de la Durantaye, K. (2025). Control and Compensation. A Comparative Analysis of Copyright Exceptions for Training Generative AI. *International Review of Intellectual Property and Competition Law (IIC)*, 56, 737-770 (2025). <https://doi.org/10.1007/s40319-025-01569-6>
- Directive EU 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market. <https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng>
- Directive 2014/26/EU of the European Parliament and of the Council of 26 February 2014 on Collective Management of Copyright and Related Rights and Multi-Territorial Licensing of Rights in Musical Works for Online Use in the Internal Market. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32014L0026>
- Dornis, T. W., & Stober, S. (2025). Generative AI Training and Copyright Law. *Transactions of the International Society for Music Information Retrieval*. <https://doi.org/10.48550/arXiv.2502.15858>
- Fernandes, P. M. (2024). AI Training and Copyright: Should Intellectual Property Law Allow Machines to Learn?. *Bioethica*, 10(2), 8-21. <https://doi.org/10.12681/bioeth.39041>
- Ge, Ch. (2025). Beyond Fair Use and Opt-Out: Forging a Hybrid Copyright Path for Generative AI in China. *Innovative Applications of AI*, Vol.2 Issue 3(2025), 141-146. <https://doi.org/10.70695/10.70695/IAAI202503A9>
- Getty Images v. Stability AI, (2025). EWHC 2863 (Ch). Royal Courts of Justice. <https://www.judiciary.uk/wp-content/uploads/2025/11/Getty-Images-v-Stability-AI.pdf>
- Guadamuz, A. (2024). A Scanner Darkly: Copyright Infringement in Artificial Intelligence Inputs and Outputs. *GRUR International*, Volume 73, Issue 2, February 2024, Pages 111-127., <http://dx.doi.org/10.2139/ssrn.4371204>
- Hardman, B. (2026). When Market Disruption Is Not Market Harm: AI Training, Causation, and the Boundaries of Fair Use (January 21, 2026). Available at SSRN: <http://dx.doi.org/10.2139/ssrn.6111926>
- Iskakova, J.T., & Kashkin, S.Y. (2020). Modern Copyright Law And The Problems Of Artificial Intelligence Development. *Вестник Университета имени О. Е. Кутафина*, 2/2020, 43-52. <https://doi.org/10.17803/2311-5998.2020.66.2.043-052>
- Kadrey et al v. Meta Platforms, Inc., (2025). No. 3:2023cv03417 - Document 598. United States District Court northern District Of California. <https://law.justia.com/cases/federal/district-courts/california/candce/3:2023cv03417/415175/598/>
- Law of the Republic of Kazakhstan No. 6-I “On Copyright and Related Rights” (June 10, 1996, as amended). <https://adilet.zan.kz/rus/docs/Z9600000006>
- Law of the Republic of Kazakhstan No. 230-VIII “On Artificial Intelligence” (November 17, 2025). <https://adilet.zan.kz/rus/docs/Z25000000230>
- Liu, G. (2025) Generative Artificial Intelligence Generates Images: Copyright Infringement or Fair Use. *Advances in Education, Humanities and Social Science Research*, Volume 13, Issue 1, 59. <https://doi.org/10.56028/aehtsr.13.1.59.2025>
- Murray, M.D. (2023). Generative AI Art: Copyright Infringement and Fair Use. *SMU Science and Technology Law Review*, Vol.26/2, 258-316. <https://doi.org/10.25172/smustlr.26.2.4>
- Opderbeck, D. W. (2024). Copyright in AI Training Data: A Human-Centered Approach. *Oklahoma Law Review*, Vol. 76, 2024 (Forthcoming), *Seton Hall Law School Legal Studies Research Forthcoming*, Available at SSRN: <http://dx.doi.org/10.2139/ssrn.4679299>
- Quang, J. (2021). Does Training AI Violate Copyright Law? *Berkeley Technology Law Journal*, Volume 36, Issue 4, 1407-1436. DOI: <https://doi.org/10.15779/Z38XW47X3K>

- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- Rodriguez Maffioli, D. (2023). Copyright in Generative AI training: Balancing Fair Use through Standardization and Transparency. Available at SSRN: <http://dx.doi.org/10.2139/ssrn.4579322>
- Rosati, E. (2025). Copyright Exceptions and Fair Use Defences for AI Training Done for “Research” and “Learning,” or the Inescapable Licensing Horizon. *European Journal of Risk Regulation*, Vol. 16, Issue 3, 961-984. <https://doi.org/10.1017/err.2025.10035>
- Samuelson, P. (2023). Generative AI meets copyright. *Science Journal*, Vol. 381/6654, 158-161. <https://www.science.org/doi/10.1126/science.adi0656>
- Samuelson, P. (2024) How to Think about Remedies in the Generative AI Copyright Cases. *Communications of the ACM*, Volume 67, Issue 7, 27-30. <https://doi.org/10.1145/3654699>
- Section 107, Title 17 of the United States Code (Limitations on Exclusive Rights: Fair Use). <https://www.law.cornell.edu/uscode/text/17/107>
- Statement of Interest of the United States, *In re OpenAI, Inc., Copyright Infringement Litigation*, No. 1:25-md-03143-SHS-OTW (S.D.N.Y. Sept. 1, 2026). <https://storage.courtlistener.com/recap/gov.uscourts.nysd.640396/gov.uscourts.nysd.640396.1669.0.pdf>
- Tang, X. (2021). Can Copyright Holders Do Harm to Their Own Works? A Reverse Theory of Fair Use Market Harm. 54 U.C. Davis L. Rev. 1245 (2021), *UCLA School of Law*, Public Law Research Paper No. 21-13, Available at SSRN: <http://dx.doi.org/10.2139/ssrn.3264238>
- Temirbekov, Zh. (2021). Artificial intelligence and law: brief overview. *Law and State*, 3(92), 2021-01-01, 142-156. DOI: 10.51634/2307-5201\_2021\_3\_142
- The New York Times Company v. Microsoft Corporation et al. (2025). No. 1:2023cv11195 - Document 514 (S.D.N.Y. 2025). United States District Courtsouthern District Of New York <https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2023cv11195/612697/514/>
- Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc., (2025). No. 1:20-cv-613-SB, 2025 WL 458520 (D. Del. Feb. 11, 2025). <https://www.ded.uscourts.gov/opinion/thomson-reuters-enterprise-centre-gmbh-et-al-v-ross-intelligence-inc-5>
- Torrance, A. W., & Tomlinson, B. (2023) Training Is Everything: Artificial Intelligence, Copyright, and Fair Training. *ArXiv*. <https://doi.org/10.48550/arXiv.2305.03720>
- Turetskiy, N.N. (2025). About The Law Of The Republic Of Kazakhstan “On Artificial Intelligence”. *Bulletin Of Institute Of Legislation And Legal Information Of The Republic Of Kazakhstan*, Volume 80, No4-2025, 398-408. [https://doi.org/10.52026/2788-5291\\_2025\\_80\\_4\\_398](https://doi.org/10.52026/2788-5291_2025_80_4_398)
- U.S. Copyright Office. (2025). *Copyright and artificial intelligence, part 3: Generative AI training* (Pre-publication version). <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf>

FTAMP: 10.41.29

## **ЖИ-ді оқыту деректерін лицензиялаудың дамуы: әділ пайдалану мен нарықтық зиянды қайта бағалау**

Жақсылық Айша  
СДУ Университеті, Қаскелең, Қазақстан  
email: [251407001@sdu.edu.kz](mailto:251407001@sdu.edu.kz)

### **Аңдатпа**

Генеративті жасанды интеллект (ЖИ) жүйелерін оқыту үшін авторлық құқықпен қорғалатын туындыларды рұқсатсыз пайдалану – қазіргі авторлық құқықтағы ең даулы мәселелердің бірі. Соттар мен ғалымдар әділ пайдаланудың төртінші факторын деректерді жинау, модельді оқыту және коммерциялық енгізу кезеңдерін қамтитын процеске қалай қолдану керектігі туралы бірыңғай пікірге келе алмай отыр. Мақалада оқыту деректерін лицензиялау нарықтары қалыптасқан сайын әділ пайдаланудың төртінші факторы шеңберінде нарыққа келетін зиянды бағалау қалай құрылымдалуы тиіс екені және мұны машиналық оқылатын опт-аут ережесін қабылдаған Қазақстан үшін не білдіретіні зерттеледі. АҚШ пен ЕО авторлық құқығына доктриналық талдау, салыстырмалы-құқықтық әдіс және үш ғылыми ұстанымның сыни синтезін қолдана отырып, мақала соңғы алты сот шешімін, оның ішінде *Andy Warhol Foundation v. Goldsmith*, *Kadrey v. Meta* және *Thomson Reuters v. ROSS Intelligence* істерін, нақты немен шешілгені тұрғысынан қайта талдайды. Талдау көрсеткендей, қолданыстағы тәсілдер доктриналық тұрғыдан ерекшеленетін екі кезеңдегі пайдалануға бірыңғай стандарт қолданады. Сондықтан мақала кезеңдік модельді ұсынады: оқыту процесінің бастапқы кезеңіне қатысты неғұрлым икемді бағалау стандартын, ал коммерциялық енгізу кезеңінде қолданылатын қатаңырақ тексеру стандарттарымен ұштастырады, әрі бұл модель ұжымдық лицензиялау мен ашықтық міндеттемелерімен толықтырылады. Осы модельді Қазақстанға қолдана отырып, мақала Қазақстан Республикасының «Жасанды Интеллект туралы» Заңының 23-бабының 5-тармағында көзделген машиналық оқылатын опт-аут ережесін бағалайды және қолданыстағы ұжымдық басқару инфрақұрылымына негізделген институционалдық жүйе ұсынады.

**Кілт сөздер:** ЖИ оқыту, авторлық құқық, әділ пайдалану, нарықтық зиян, лицензиялау

МРНТИ: 10.41.29

## **Развитие лицензирования данных для обучения ИИ: переоценка добросовестного использования и рыночного ущерба**

Жақсылық Айша  
Университет СДУ, Каскелен, Казахстан  
email: [251407001@sdu.edu.kz](mailto:251407001@sdu.edu.kz)

### **Аннотация**

Нелицензированное использование охраняемых авторским правом произведений для обучения генеративных систем искусственного интеллекта (ИИ) – один из наиболее спорных вопросов авторского права. Суды и исследователи расходятся во мнениях относительно того, как четвертый фактор добросовестного использования должен применяться к процессу сбора данных, обучения модели и коммерческого внедрения. В статье исследуется, как должен быть структурирован

анализ рыночного ущерба по четвёртому фактору в условиях формирования рынков лицензирования обучающих данных, и что это означает для Казахстана, принявшего норму об отказе (opt-out) от использования произведений для обучения ИИ без развитой доктринальной основы. На основе доктринального анализа права США и ЕС, сравнительно-правового метода и критического синтеза трёх научных позиций статья заново анализирует шесть недавних решений, включая *Andy Warhol Foundation v. Goldsmith*, *Kadrey v. Meta* и *Thomson Reuters v. ROSS Intelligence*, с точки зрения того, что было решено. Показано, что существующие подходы применяют единый стандарт проверки к использованию в двух доктринально различных фазах, и предлагается поэтапная модель: разрешительный стандарт оценки для начального этапа обучения и усиленный контроль для этапа коммерческого внедрения, дополненные коллективным лицензированием и прозрачностью. Применяя эту модель к Казахстану, статья оценивает машиночитаемый механизм отказа, предусмотренный пунктом 5 статьи 23 Закона Республики Казахстан «Об искусственном интеллекте», и предлагает институциональный механизм на основе существующей инфраструктуры коллективного управления.

**Ключевые слова:** обучение ИИ, авторское право, добросовестное использование, рыночный ущерб, лицензирование

Received: 28 April, 2026