Scientific Research Publishing

Beijing Law Review, 2025, 16(1), 257-272
https://www.scirp.org/journal/blr
ISSN Online: 2159-4635
ISSN Print: 2159-4627

Intellectual Property in Language Models: Challenges of Ownership in the Integration of Multiple Databases

Thiago de Bessa da Silva[1*], Moisés Cirilo de Brito Souto[1], Max Miller da Silveira[1], Geraldo Andrade de Oliveira[2], Anderson Zanati Dultra[3], Tawfic Awwad Júnior[3], Thiago Marques da Costa[1], Rocco Antonio Rangel Rosso Nelson[4], Thiago Murilo Nobrega Galvão[5]

[1]Competence Center in Open Solutions, Federal Institute of Rio Grande do Norte, Natal, Brazil
[2]Campus Presidente Kennedy, Federal Institute of Espírito Santo, Vitória, Brazil
[3]Electronic Social Communication Secretary, Ministry of Communications, Brasilia, Brazil
[4]Campus Natal-Central, Federal Institute of Rio Grande do Norte, Natal, Brazil
[5]Legal Office, Federal Institute of Rio Grande do Norte, Natal, Brazil

Email: *thiagodebessa@gmail.com, moises.souto@ifrn.edu.br, max.silveira@ifrn.edu.br, geraldoandrade@projetosdepesquisa.org, anderson.dultra@mcom.gov.br, tawfic.awwad@mcom.gov.br, thiago.marques@ifrn.edu.br, rocconelson@hotmail.com, thiagonobrega@icloud.com

How to cite this paper: Silva, T. B., Souto, M. C. B., Silveira, M. M., de Oliveira, G. A., Dultra, A. Z., Júnior, T. A., da Costa, T. M., Nelson, R. A. R. R., & Galvão, T. M. N. (2025). Intellectual Property in Language Models: Challenges of Ownership in the Integration of Multiple Databases. *Beijing Law Review, 16,* 257-272.
https://doi.org/10.4236/blr.2025.161012

Received: February 17, 2025
Accepted: March 11, 2025
Published: March 14, 2025

Abstract

This work is the result of the project Innovative Solution for Generative Artificial Intelligence for Automated Responses, developed within the research group Centro de Competências em Soluções Livres at IFRN (CCSL-IFRN), and addresses the legal challenges related to intellectual property ownership and data protection in the development of language models based on Artificial Intelligence (AI). The study of this topic is relevant due to the insufficiency of current legislation in dealing with issues such as authorship, use of sensitive data, and liability. The objective of this work is to analyze these challenges and propose guidelines for a regulatory framework that balances innovation and the protection of rights. The methodology consists of qualitative research, using a hypothetical-deductive approach, employing bibliographic, documentary, and jurisprudential analysis. Based on the analysis conducted, it is concluded that a specific regulatory framework is necessary to address the peculiarities of language models, ensuring clarity in defining ownership and respecting the rights involved.

Keywords

Intellectual Property, Language Models, Data Protection, Legal Ownership, Artificial Intelligence

1. Introduction

The accelerated development of Artificial Intelligence (AI) currently being experienced was only made possible by a convergence of factors. The evolution of Graphics Processing Units (GPUs) enabled a significant increase in processing capacity, learning efficiency, and energy savings. Additionally, access to a larger volume of data, generated by digitization[1] and the proliferation of connected devices, although under ethical tensions regarding privacy and ownership of information, has been used to feed algorithms and train models. Finally, the availability of computational and storage infrastructure in the cloud democratized access to resources that were previously inaccessible due to high costs and complexity, allowing companies and researchers to rent capacity on demand. The recent technological evolution related to Artificial Intelligence (AI) was driven by a combination of factors, such as the availability of storage and processing solutions; access to data that were previously inaccessible or unobservable; the proliferation of advanced AI algorithms; and favorable economic conditions that encouraged the development and application of these technologies (Cozman et al., 2021: p. 345).

This technological triad redefines legal challenges, requiring adaptive regulation that balances incentives for innovation, protection of fundamental rights, and governance of systemic risks. In this context, the law should not be seen as an obstacle to technological progress but rather as an enabler, ensuring the development of an AI ecosystem that respects the fundamental rights and guarantees that underpin the Democratic Rule of Law.

The issue surrounding the data used for creating and training language models is particularly sensitive for the legal field. This is because it requires the integration of datasets from diverse sources for such training, which can range from public domain content available on the internet to private or sensitive information, obtained through data mining.[2] This raises serious legal challenges, especially related to intellectual property ownership in a broader perspective of these models.[3]

[1] The term "digitalization" was initially conceived to designate the information technologies responsible for processing digital data, as well as the structures created to enable the operation of these technologies. However, its meaning transcends that technical definition, encompassing a profound transformation in living conditions driven by the widespread use of these tools. In this sense, digitalization enabled the adoption of systems that made automated and interconnected productive processes feasible, as seen in Industry 4.0. Additionally, it redefines aspects of people's everyday lives, such as home automation, alters the way we interact through social media platforms, introduces new formats of instant communication, and gives rise to sophisticated surveillance mechanisms operated by both private entities and governmental institutions (Hoffmann-riem, 2021: p. 367).

[2] Data mining refers to an analytical process that aims to discover underlying patterns and develop predictions based on predictive projections. Although the term "data mining" emerged in the 1990s, its conceptual foundations are rooted in three interdependent scientific domains: statistics, which deals with quantitative analysis to identify correlations and perform inferences on datasets; artificial intelligence, which encompasses computational systems with analytical abilities comparable to human cognitive capacities; and machine learning, a specialized field in the development of algorithms capable of learning autonomously and building predictive models from large volumes of information (Mèlo, 2019: p. 116).

[3] The subject of the economic impacts of intellectual property can be analyzed on three distinct levels. At the microeconomic level, intellectual property is understood as a strategic tool used by economic agents, whether they are innovators or not. It functions as a mechanism to ensure the appropriation of the results stemming from innovative efforts, with its economic effects initially observed within companies. These impacts are significant both for the definition of organizational strategies and for decisions related to investments and innovation activities. From a broader perspective, the role of intellectual property is discussed from a social point of view, evaluating whether it strengthens or reduces the market power of innovators, whether it contributes positively or negatively to social welfare, whether it favorably influences the pace of innovation in certain sectors, and whether it affects the development of countries. Finally, in a normative approach, intellectual property is considered a potential instrument of public policy, employed to encourage or discourage specific behaviors by economic agents. In this context, it is especially relevant to discuss the limits of the effectiveness of this mechanism in achieving the desired objectives (Mello, 2010: p. 373).

This occurs because traditional legal concepts related to intellectual property, developed before the popularization of generative AI, face difficulties in addressing the collaborative and multifaceted process of developing these models. Thus, the central question arises: how is ownership determined within the Brazilian legal context for a language model that uses data from different sources? This problem unfolds into other important questions, such as: defining the legal nature of the data used; protecting the copyrights of original content creators; and the impacts of using sensitive or personal data.

Thus, this study analyzes these legal challenges related to intellectual property ownership in language models, focusing on the integration of multiple data sources. Specifically, it will examine: 1) the legal aspects of intellectual property in AI models; 2) the issue of personal data protection and liability in the use of databases; and 3) ownership rights in language models and possible solutions for determining ownership in projects involving varied data sources.

The relevance of this research lies in the growing importance of artificial intelligence systems in the current technological landscape and the urgency of establishing clear legal guidelines to guide their development. The lack of precise definitions regarding the ownership of these models could hinder innovation in the sector and generate legal conflicts among the parties involved.

In light of the above, the present research employs a qualitative methodology, using a hypothetical-deductive approach, with descriptive and analytical characteristics, adopting bibliographic and documentary research techniques, examining legislation, doctrine, and case law.

2. Legal Aspects of Intellectual Property in Language Models

The current scenario of technological innovation presents a unique challenge to the intellectual property legal regime applied to artificial intelligence. The question is how to protect and promote the development of generative artificial intelligence technologies without compromising the fundamental principles of the legal system that govern intellectual creations. This issue becomes particularly relevant when considering the nature of AI systems, which challenge traditional notions of authorship and inventiveness.[4]

In the international context, intellectual property laws, with particular emphasis on the TRIPS Agreement (Trade-Related Aspects of Intellectual Property Rights),[5] of the World Trade Organization (WTO) and the Berne Convention for the Protection of Literary and Artistic Works,[6] impose minimum protection

[4] "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment." (OECD, 2025).

[5] Incorporated into the legal framework through Decree No. 9,289, of February 21, 2018, which promulgates the Protocol of Amendment to the Agreement on Trade-Related Aspects of Intellectual Property Rights, adopted by the General Council of the World Trade Organization on December 6, 2005.

[6] Promulgated by Decree No. 75,699, of May 6, 1975 (Brasil, 2018a).

standards on Brazil as a signatory. These standards directly influence the Brazilian regulatory approach since, once incorporated into the domestic legal framework, national legislation must reflect these guidelines to ensure compliance with international obligations.

The TRIPS Agreement, for instance, establishes minimum protection criteria for patents, copyrights—including the protection of software as a literary work, which is fundamental for AI—and trade secrets, impacting the regulation of artificial intelligence. The Berne Convention, by providing for the automatic protection of literary and artistic works, influences Brazil's approach to the ownership and exploitation of content generated by autonomous systems. However, the issue of authorship of AI-generated works remains an open debate, as will be discussed throughout this article.

The incorporation of these standards demands a thorough analysis, especially regarding the development of artificial intelligence, such as defining authorship in AI-generated works and protecting intangible assets such as algorithms, AI models, know-how, and databases. In the Brazilian legal system, as well as in various international legal systems, the protection of intellectual creations is structured around the anthropocentric concept of the author, meaning the central figure is the human author.[7] This conception is enshrined in the list of fundamental rights, specifically in Article 5, items XXVII and XXIX, of the Constitution,[8] and in the introductory clause of Article 11 of the Copyright Law (Law No. 9,610/1998),[9] which establishes the link between authorship and the individual. However, the Superior Court of Justice (STJ), in the ruling of Special Appeal No. 1,473,392-SP,[10] expanded the traditional understanding by recognizing that legal entities can also hold copyright.

Even so, these premises prove insufficient with the emergence of works generated by generative AI, since these models operate autonomously, without direct

[7] "The designation of authorship presupposes recognizing the condition of incompleteness that allows humans to observe their own existence and always call it into question. Technological development in the field of AI does not follow this direction, so that the machine-author appears as a product of an anthropocentric vanity that historically subjugates a large part of humanity to serve interests in whose deliberation it does not participate." (Arrabal, 2024: p. 118).

[8] "Art. 5º, XXVII—Authors shall have the exclusive right to use, publish, or reproduce their works, which shall be transferable to their heirs for the period fixed by law." "Art. 5º, XXIX—The law shall ensure to the authors of industrial inventions a temporary privilege for their use, as well as protection for industrial creations, trademarks, company names, and other distinctive signs, having regard to the social interest and the technological and economic development of the Country." (Brasil, 1988b).

[9] "Art. 11. An author is the natural person who creates a literary, artistic, or scientific work." (Brasil, 1998b)

[10] "RECURSO ESPECIAL. DIREITOS AUTORAIS. CONTRATO SOB ENCOMENDA. PESSOA JURÍDICA. TITULAR DE DIREITOS DO AUTOR. POSSIBILIDADE. DIREITO À INDENIZAÇÃO. OBRA UTILIZADA SEM A DEVIDA AUTORIZAÇÃO. 1. Nos contratos sob encomenda de obras intelectuais, a pessoa jurídica, que figura como encomendada na relação contratual, pode ser titular dos direitos autorais, conforme interpretação do art. 11, parágrafo único, da Lei 9.610/98. 2. Assim, ocorrendo a utilização posterior da obra encomendada, sem a devida autorização, caberá à pessoa jurídica contratada pleitear a reparação dos danos sofridos. 3. Recurso especial não provido." (Brasil, 2016).

and continuous human intervention, generating content that cannot be directly attributed to a specific human author.[11] This regulatory gap,[12] challenges the pillars of copyright law, which presuppose the existence of an identifiable individual or collective creator (Samuelson, 2016).

Moreover, the collaborative and decentralized nature of AI model development, which often involves multiple databases, algorithms, and technical resources, complicates the issue of determining the legal nature of ownership for models trained with such data. Different categories of information require distinct legal treatments.

In many cases, ownership of the AI model may be attributed to the developers or the company that funded the project. In these situations, the rights over the AI model are often regulated by intellectual property agreements, which define ownership, benefit distribution, and responsibilities among the parties involved. Additionally, they depend on the legal nature of the data used in its training, such as copyrights, patents, and trade secrets.

Within the context of copyright law, the protection granted to computer programs,[13] is based on the same regulatory framework intended for literary works, as established by Brazilian legal order.[14] This protection, primarily based on Law No. 9,609/98 (Software Law) in conjunction with Law No. 9,610/98 (Copyright Law), recognizes the intellectual character of these creations, analogous to literary works in their protective essence, despite their technical and functional peculiarities.

However, there remains a significant regulatory gap in the legal system regarding autonomous creations derived from the operation of computer programs, such as Generative AIs. These algorithmic creations, generated independently of direct and continuous human intervention, demand a differentiated legal regime due to the absence of clear parameters for attributing authorship, ownership, and

[11] "The issue of authorship, as demonstrated, represents a major barrier to the construction of an idea of an AI as a subject of author's rights, since the theme is strongly linked to the element of creativity, which is inherent to human beings. The heart of the matter, therefore, is: to what extent can an AI be, in fact, "creative"? Considering the fact that we are referring to algorithms programmed by human beings, and that the information that the AI collects from the Internet is also, for the most part, created by human users, could this basis constitute a field for the recognition of copyright" (Souza, 2022).

[12] In this context it is important to mention that: "The function of the legal scholar is like that of the watchmaker when fitting all the pieces together, in a rational manner, one in relation to the others, so that the watch can function and tell the time, which in the legal field would be equivalent to connecting the norms, as if they were sets of wheels, of toothed pieces, of springs and other connections, which serve for the functioning of a watch, forming a systematic whole." (Diniz, 1999: pp. 26-27).

[13] Law No. 9,609, of February 19, 1998, defines in its Article 1 a computer program as "the expression of an organized set of instructions in natural or coded language, contained in a physical medium of any nature, necessarily used in automatic information processing machines, devices, instruments, or peripheral equipment, based on digital or analog techniques, to make them function in a specific manner and for determined purposes." (Brasil, 1998a).

[14] "Article 2—The intellectual property protection regime for computer programs is the same as that granted to literary works under the copyright and related rights legislation in force in the country, subject to the provisions of this Law." (Brasil, 1998a).

limits of protection in the current legislation.[15]

Regarding the patent system, regulated by the Industrial Property Law (Law No. 9,279/1996), it faces particular challenges when applied to AI-based innovations. The automation of the creative process questions traditional criteria for evaluating inventive activity, requiring a reinterpretation of patentability requirements. In this context, protection through trade secrets has emerged as a strategic alternative for safeguarding algorithms and development methodologies.

Given this scenario, it becomes evident that the current legal system protecting intellectual property requires a different legal approach. The advent of artificial intelligence not only challenges traditional concepts of authorship and inventiveness but also raises questions about how to protect and incentivize innovation. The complexity of AI-generated creations demands a reformulation of the legal framework to balance the protection of individual and collective rights with the need to promote technological and economic development.

As an example of the regulatory insufficiency of the Brazilian legal framework in addressing the challenges posed by generative AI, the Copyright Law (Law No. 9,610/1998) and the Software Law (Law No. 9,609/1998) can be cited. These laws were conceived in a context where human intervention is essential for the creation of protected works but fail to address scenarios in which generative AI systems produce works without direct human involvement in the ideation or expression of the work.

For instance, generative AI models can create texts, images, or music, trained on large datasets, without clear guidelines on who holds the copyright over these creations. The absence of legal provisions to recognize authorship and even originality of AI-generated works generates significant legal uncertainties, challenging fundamental principles such as the very definition of creativity.

The regulatory challenges imposed by AI also extend to critical legal issues, threatening principles such as privacy and free competition, especially in cases where models rely on large volumes of data to operate. The massive collection and processing of personal data to train AI systems often conflict with regulations such as the General Data Protection Law (Law No. 13,709/2018). This raises concerns about consent, algorithmic discrimination due to biases in datasets, and, crucially, the lack of explainability and transparency in AI decision-making.

3. Data Protection and Liability in the Use of Databases

The Federal Constitution, in its current wording after Constitutional Amendment No. 115, of February 10, 2022, enshrined the right to personal data protection, including in digital contexts, as a fundamental right (Brasil, 1988a). This constitutional norm emerged in alignment with the evolving jurisprudence and regulatory framework that had already been developing in the country.

[15] "This gap exists because the concept of authorship in Brazilian law was developed with a focus on natural persons and the so-called 'creations of the mind'. However, we have also seen that the LDA itself extends the concept of authorship to legal entities, and within these exceptions, we could, by analogy, find some room for autonomous AI to be considered an author." (Rici, 2023).

Indeed, even before the entry into force of the aforementioned Constitutional Amendment,[16] the Brazilian legal system already included the General Data Protection Law (LGPD), Law No. 13,709/2018, which established a significant milestone by defining specific parameters for the handling of personal data. It requires data subjects to provide clear, express, and voluntary consent for their personal information to be used, after being properly informed of the specific purpose for which it is intended.

In this context, the concept of data goes beyond the mere notion of information that can be collected, stored, processed, or transmitted, regardless of format or medium. The model established by the LGPD for regulating the processing of such information classifies it according to its nature and degree of sensitivity, demonstrating this.

Personal data are classified as any information that allows the direct or indirect identification of an individual. At a higher level of protection are sensitive data, which, due to their potential to generate discrimination, require qualified protection. These include information about racial or ethnic origin, religious beliefs, political preferences, health, sexual life, and biometric or genetic characteristics. There are also anonymized data, which have undergone irreversible techniques that prevent the identification of the data subject,[17] ensuring their safe use in research and statistical analyses.[18]

[16] "Thus, it is observed that the legal framework already contained, prior to the LGPD, reasonably specific regulations protecting personal data, with only an apparent antinomy between the norms, which is resolved by the application of the LGPD to personal data processing activities carried out after its enactment." (Grossi, 2022: p. 85).

[17] "Currently, there are several techniques and best practices for protecting sensitive data in AI projects. Among the most commonly used are: Anonymization and Pseudonymization: Removes or replaces identifiable personal information, such as names, addresses, and identification numbers, with non-personal identifiers to ensure that the data cannot be directly linked to a specific individual. Encryption: Applies encryption techniques to protect sensitive data during storage, transmission, and processing, ensuring that only authorized individuals can access the data. Data Minimization: Collects only the data strictly necessary for the AI project's purpose, reducing the amount of sensitive information that needs protection. Restricted Access: Limits access to sensitive data only to authorized individuals through appropriate access control, authentication, and authorization measures. Network Security: Protects the networks used in the AI project to prevent unauthorized access and data interception, utilizing firewalls, VPNs, and other security measures to strengthen network infrastructure. Training in a Secure Environment: Involves the use of federated learning techniques, which allow models to be trained without real data being shared between servers, ensuring that sensitive data is not accidentally exposed. Risk and Privacy Impact Assessment: Conducts risk and privacy impact assessments to identify and mitigate potential threats to the security of sensitive data and users' personal information. Auditing and Monitoring: Implements auditing and monitoring systems to track access to and use of sensitive data, enabling the detection of suspicious behavior or unauthorized activities. Regulatory Compliance: Ensures compliance with data privacy laws and regulations relevant to the AI project. Education and Awareness: Proposes training for all members of the AI project team to foster awareness and adherence to data protection principles." (Silva, 2023: pp. 52-53).

[18] The concepts of personal data, sensitive personal data, and anonymized data are codified in Article 5 of the LGPD, as follows: "Article 5—For the purposes of this Law, the following definitions apply: I—Personal data: information related to an identified or identifiable natural person; II—Sensitive personal data: personal data concerning racial or ethnic origin, religious belief, political opinion, trade union membership or affiliation with a religious, philosophical, or political organization, as well as data concerning health, sex life, genetic, or biometric data when linked to a natural person; III—Anonymized data: data related to a data subject that cannot be identified, considering the use of reasonable and available technical means at the time of processing." (Brasil, 2018b).

Specifically, in the context of developing Artificial Intelligence (AI) language models, the LGPD imposes a set of requirements that must be observed to ensure legal compliance. Among these requirements, the principles of purpose, necessity, and transparency stand out. Their objective is to limit the use of personal data to what is strictly necessary for the development and operation of AI systems, aiming to balance technological advancement with the protection of individuals' fundamental rights, particularly regarding privacy and the responsible use of personal data.[19]

Adherence to these principles not only affects how data is acquired and used but also has significant implications for the intellectual property of language models. Ownership and usage rights over these models can be significantly impacted by the need to document the origin and legality of each dataset used in the training process.

Moreover, the LGPD does not yet have specific regulations addressing the reuse of anonymized data, creating legal uncertainties regarding ownership and licensing of such data. Although Article 12 of the LGPD excludes anonymized data from the concept of personal data,[20] the lack of clear guidelines on their reuse and the possibility of reversing the anonymization process presents challenges that need to be addressed by legislators.

It is important to highlight that data anonymization is often presented as a solution to reconcile the use of large databases with the requirements for personal data protection. However, this technique faces considerable challenges, especially in the context of language models. A practical example of these challenges occurs when language models are trained with texts that, even when anonymized, may contain linguistic biases that allow for the indirect identification of individuals.[21]

In the specific context of language models, the issue of re-identification takes on even more complex contours. For example, when a model is trained with multiple anonymized datasets, the cross-referencing of information may reveal patterns that allow for the identification of individuals, even if each dataset is individually considered secure. This has direct implications for intellectual property, as it may necessitate the removal or modification of parts of the training set, potentially affecting ownership and rights over the model. This problem also finds parallels in the European context, where its general regulatory framework, the General Data Protection Regulation—GDPR (European Union, 2016), also faces the challenge of establishing secure criteria for effective anonymization.

[19] Article 6 of the LGPD lists and defines the principles that, together with the principle of good faith, must be observed in the processing of personal data. Among them, the principles of purpose, necessity, and transparency were chosen as they reflect specific challenges inherent to language model technology, which requires massive data processing, often including personal information (Brasil, 2018b).

[20] Article 12 of the LGPD: "Anonymized data shall not be considered personal data for the purposes of this Law, except when the anonymization process to which they have been subjected is reversible, using exclusively proprietary means, or when it can be reversed with reasonable efforts." (Brasil, 2018b).

[21] "Given that biases present themselves as an intrinsic characteristic of human thinking, it can be concluded, similarly, that an algorithm created by biased human beings will likely suffer from the same 'ailment', not intentionally, but as a result of the information provided to the system. In this way, the so-called algorithmic biases arise, which occur when machines behave in ways that reflect the implicit human values involved in the programming." (Nunes et al., 2018).

In the global context, the legal frameworks proposed by Brazil to regulate artificial intelligence (AI) clearly draw inspiration from the European Union's General Data Protection Regulation (GDPR), particularly through the General Data Protection Law (LGPD). However, while the GDPR already establishes specific and robust guidelines for AI usage—including requirements for transparency, the right to explanation, and restrictions on sensitive data processing—Brazil still faces significant gaps in its regulatory approach.

Although the LGPD is aligned with the principles of the GDPR, it lacks specific regulations addressing artificial intelligence. In contrast, the GDPR defines clear criteria for explainability and transparency in automated systems and personal data protection in AI contexts. For example, the GDPR mandates that individuals affected by automated decisions have the right to obtain clear explanations of how these decisions were made, as well as mechanisms for contestation and human review. This detailed approach strengthens AI operators' accountability and fosters greater public trust in AI-driven technologies. In Brazil, however, the LGPD remains more generic, without specific obligations to address AI-related challenges, leaving regulatory gaps in the legal framework.

Furthermore, regarding the security and confidentiality of data, it is observed that the LGPD requires data controllers to adopt technical and administrative measures to prevent security incidents related to the processing of personal data.[22] However, the legislation does not specify the legal consequences in cases of involuntary re-identification, creating a regulatory gap that is particularly problematic for AI models that integrate multiple datasets. This gap makes the attribution of responsibility uncertain in cases of privacy violations, especially when the processing chain involves various actors, from the original data providers to the model developers and end users.

Another particularly complex issue is the legal liability of language models that use multiple datasets. Since the chain of obligations may involve various actors, from the original data providers to the model developers and end users, a clear structure for assigning responsibilities in cases of privacy violations or misuse of data becomes necessary.

To mitigate these risks, the LGPD provides that the National Data Protection Authority (ANPD) may require data controllers to prepare a Data Protection Impact Report (DPIA), which maps potential risks when personal data is processed, particularly concerning the fundamental rights and freedoms of individuals. It specifies which actions, tools, and protections will be implemented to reduce or avoid the identified risks.[23]

[22] This is what the head of article 46 of the LGPD says: "The processing agents must adopt security measures, technical and administrative measures capable of protecting personal data from unauthorized access and from accidental or unlawful situations of destruction, loss, alteration, communication or any form of inappropriate or unlawful processing." (Brasil, 2018b).

[23] Art. 5º, XVII: "report on the impact of personal data protection—documentation of the controller containing the description of personal data processing processes that may generate risks to civil liberties and fundamental rights, as well as risk mitigation measures, safeguards and mechanisms" (Brasil, 2018b).

Thus, the development of AI language models in Brazil faces significant challenges at the intersection of personal data protection and intellectual property. To ensure compliance with the LGPD, it is essential to adopt a planned approach, especially regarding anonymization and data reuse.

In this context, the DPIA emerges as a tool to identify risks and implement effective protective measures. However, complex issues such as legal liability and the possibility of data re-identification still require more precise regulation to ensure the legal certainty necessary for the safe and ethical development of these technologies in Brazil.

A strategy for reconciling the development of AI models based on large volumes of data with data protection laws can be achieved through the use of anonymization and pseudonymization techniques. These tools enable AI training data to be processed in a way that preserves individual privacy. Anonymization removes or alters information that could directly identify an individual, while pseudonymization replaces personal data with fictitious identifiers, allowing reversion only under specific conditions. By implementing these techniques, companies and developers could leverage large datasets to train AI models without violating regulations such as the LGPD, ensuring legal compliance without compromising the data quality essential for technological innovation.

4. Ownership of Language Models Trained on Multiple Sources

As demonstrated thus far, the issue of copyright protection, intellectual property rights, and trade secrets in language models trained on multiple sources must be resolved through the enactment of new legislation that provides objective criteria aligned with market needs and the protection of fundamental rights, ensuring the development of competitive language models in Brazil compared to those developed abroad.

The current regulatory framework for copyright in Brazil, outlined by Law No. 9,610/1998, proves insufficient to address the challenges posed by contemporary artificial intelligence systems, particularly language models trained from diverse data sources. This regulatory gap arises because the existing legislation was conceived in a distinct technological context, based on the assumption that only natural persons can be considered authors of intellectual works. This paradigm, although coherent at the time of its conception, does not adequately accommodate the peculiarities of AI systems, which are capable of generating content with a high degree of autonomy, often without direct or significant human intervention.

This gap between the current technological landscape and the existing legal framework highlights the urgent need for specific regulation. The creation of an updated regulatory framework will not only align Brazilian legislation with the demands of technological development but also ensure balanced protection of the rights involved, promoting both innovation and legal certainty.

The National Congress has already initiated discussions on this matter through

Bill No. 2,338 of 2023, which addresses the use of artificial intelligence in the country. The bill is currently pending review by the Chamber of Deputies (House of Representatives)[24] but has already been approved by the Senate.[25]

The bill addresses ownership issues in the chapter on fostering sustainable innovation, specifically in the section on authorship and related rights. It establishes that those responsible for AI systems utilizing content protected by copyright and related rights have a duty to disclose such usage through the publication of summaries on electronic platforms.[26]

It further stipulates that automated use of protected content does not constitute infringement when carried out by scientific, research, and educational organizations, as well as museums, public archives, and libraries, provided that access to the data occurs lawfully, without commercial purposes, and the use is limited to what is necessary to achieve the proposed objectives, respecting the economic interests of the rights holders and without interfering with the normal exploitation of the works or protected content.[27]

Additionally, the bill provides mechanisms for economic compensation to copyright holders whose works are incorporated into the training process of artificial intelligence systems. This provision establishes guiding criteria for determining such compensation, based on principles of reasonableness and proportionality, as well as the necessary analysis of competitive impact, aiming to preserve market balance and the economic viability of technological ventures.[28]

[24] According to information from the Federal Senate website (https://www25.senado.leg.br/web/atividade/materias/-/materia/157233#tramitacao_10494842), Bill No. 2338, of 2023, has already been approved by the Plenary of the Federal Senate and the draft of the final wording will still be considered by the Reviewing House (Chamber of Deputies).

[25] The final text approved by the Initiating House (Federal Senate) of Bill No. 2338, of 2023 is available at: https://legis.senado.leg.br/sdleg-getter/documento?dm=9881643&ts=1735605226813&disposition=inline

[26] The wording of the first article of this section of Bill No. 2,338, of 2023 states that: "Art. 62. The AI developer who uses content protected by copyright and related rights must disclose the protected content used in the development processes of AI systems, by publishing a summary on an easily accessible website, observing commercial and industrial secrecy, in accordance with specific regulations. Sole paragraph. For the purposes of this Chapter, development comprises the stages of mining, training, retraining, testing, validation and application of AI systems." (Brasil, 2023).

[27] According to article 63 of Bill No. 2,338 of 2023: "Art. 63. It does not constitute an infringement of copyright and related rights the automated use of protected content in text and data mining processes for the purposes of research and development of AI systems by scientific, research and educational organizations and institutions, museums, public archives and libraries, provided that the following conditions are observed: I—access has been lawful; II—it has no commercial purpose; III—the use of content protected by copyright and related rights is made to the extent necessary for the objective to be achieved, without prejudice to the economic interests of the rights holders and without competition with the normal exploitation of the protected works and content." (Brasil, 2023).

[28] According to article 65 of Bill No. 2,338 of 2023: "Art. 65. The AI agent that uses content protected by copyright and related rights in mining, training or developing AI systems must remunerate the holders of such content by virtue of such use, ensuring: I—that the holders of copyright and related rights have effective conditions to collectively negotiate, under the terms of Title VI of Law No. 9,610, of February 19, 1998 (Copyright Law), or directly the use of the content of which they are holders, and may do so free of charge or for a fee; II—that the calculation of the remuneration referred to in the caput consider the principles of reasonableness and proportionality and relevant elements, such as the size of the AI agent and the competitive effects of the results in relation to the original content used; III—free negotiation in the use of protected content, aiming to promote a research and experimentation environment that allows the development of innovative practices, and that do not restrict the freedom of agreement between the parties involved, under the terms of arts. 156, 157, 421, 422, 478 and 479 of Law No. 10,406, of January 10, 2002 (Civil Code), and art. 4º of Law No. 9,610, of February 19, 1998 (Copyright Law). § 1 The remuneration referred to in the caput of this article is due only to: I—holders of national or foreign copyright and related rights domiciled in Brazil; II—persons domiciled in a country that ensures reciprocity in protection, in terms equivalent to this article, to copyright and related rights of Brazilians, as provided for in arts. 2, sole paragraph, and 97, § 4º, of Law No. 9,610, of February 19, 1998 (Copyright Law), and collection is prohibited in cases where reciprocity is not assured. § 2 The holder of the right to remuneration provided for in the caput who opts for negotiation and direct authorization, under the terms of item I of the caput, may exercise it regardless of further regulation." (Brasil, 2023).

This represents a significant advancement in the legal treatment of the matter, although aspects requiring further elaboration and detailed regulation remain, such as ownership of language models trained on multiple databases. The legislative proposal under consideration recognizes the need to harmonize the interests of copyright holders with the technical and economic demands of developing language models, establishing a differentiated regime for institutions operating without commercial intent.

The mandatory transparency in the use of protected content, through the publication of electronic summaries, represents progress, as it allows copyright holders to monitor and, if necessary, challenge the use of their works, providing them with effective means to safeguard their interests.

Thus, a balanced regulation between the protection of fundamental rights, especially copyright, and the promotion of innovation in the use of artificial intelligence is sufficient to ensure copyright protection without hindering technological advancement. Bill No. 2,338/2023 represents an important step by introducing measures of transparency and compensation for content creators, but there are still points that need to be better defined, such as the criteria for remuneration and the ownership of models trained with multiple databases. The challenge for lawmakers will be to find a balance between innovation and legal certainty, ensuring that Brazil remains competitive in AI development without compromising the rights of authors and other stakeholders.

Furthermore, Bill No. 2,338/2023 represents a milestone in Brazil's attempt to enter the international legislative debate on Artificial Intelligence (AI), demonstrating alignment with instruments such as the EU AI Act[29] and the guidelines of the Organization for Economic Cooperation and Development (OECD).[30] The Brazilian proposal incorporates fundamental principles of global AI regulation, including transparency, explainability, non-discrimination, and risk mitigation, reflecting the risk-based approach of the AI Act. The concern with data protection, harmonizing with both the General Data Protection Law (LGPD) and the

[29] The AI Act, officially Regulation (EU) 2024/1689, is recognized as the world's first comprehensive regulatory framework on artificial intelligence. Proposed by the European Commission and later approved by the European Parliament and the Council of the European Union, this regulation establishes rules for the development, implementation, and use of AI systems across all EU Member States. Its primary objective is to ensure that AI technologies used in Europe are safe, transparent, ethical, and compliant with fundamental rights, thereby fostering responsible innovation and strengthening trust among both citizens and businesses.

[30] The OECD Guidelines on Artificial Intelligence (OECD, 2019), establish principles and recommendations for the responsible development and deployment of artificial intelligence (AI) among OECD member countries and beyond. These guidelines promote an AI ecosystem that is inclusive, sustainable, and human-centered, ensuring that AI technologies respect human rights, democratic values, and the rule of law. The five core principles outlined in the OECD guidelines are: Inclusive and Sustainable Growth—AI should contribute to economic growth and social well-being; Human-Centered Values and Fairness—AI systems must respect human rights and ensure fairness; Transparency and Explainability—AI decision-making should be understandable and accountable; Robustness, Security, and Safety—AI should function reliably throughout its lifecycle. Accountability—AI actors must be responsible for AI-related decisions and impacts. These principles influenced global AI policies, including the EU AI Act and national regulatory efforts, such as Brazil's Bill No. 2,338/2023, reinforcing a risk-based approach to AI governance.

European GDPR, also highlights this convergence.

However, the transposition of international models poses challenges. The definition of liability for damages caused by AI systems, which is crucial for legal certainty, lacks detailed provisions in Bill No. 2,338/2023 compared to the AI Act, which establishes clearer mechanisms for assigning responsibility. Additionally, adapting these regulations to the Brazilian reality—considering the current stage of AI ecosystem development and the need to foster innovation, particularly for small enterprises—remains a significant challenge. The key lies in creating a regulatory balance that avoids excessive burdens while fostering a favorable environment for technological development, without neglecting the protection of fundamental rights.

5. Conclusion

The use of language models based on artificial intelligence has brought legal challenges that must be addressed through a new regulatory framework, as the existing ones are insufficient for their regulation. Issues such as determining the ownership of these models and protecting the data used in their training remain inadequately defined in the Brazilian legal system, contributing to a scenario of legal uncertainty.

In the current legal framework, particularly the Copyright Law, was not designed to encompass creations derived from automated processes. The concept of authorship presupposes direct human intervention, which becomes problematic when dealing with generative AI models that operate by integrating large volumes of data. This reality raises discussions about the ownership of the models and the rights of third parties whose information may have been used in their development.

Beyond the issue of authorship, there are specific concerns regarding the use of databases. The General Data Protection Law (LGPD) established parameters for handling personal information, but its application in the context of artificial intelligence still requires further clarification. An example of this is the issue of reusing anonymized data and the risks of re-identification, which are critical points that demand a more detailed approach in legislation.

In this regard, Bill No. 2,338/2023, currently under discussion in the National Congress, represents an effort to regulate the use of artificial intelligence in the country. However, its text still needs refinement to address with greater precision aspects such as the ownership of models trained with diverse databases, criteria for compensating copyright holders, and accountability for the misuse of information.

Given this scenario, it is essential that lawmakers establish clear rules to ensure the protection of the rights of all parties involved without hindering technological innovation. The regulation must balance the interests of copyright holders and those responsible for developing AI models, ensuring equilibrium between legal protection and technological advancement. The law should act not as an obstacle to innovation but as an enabler that facilitates progress while safeguarding fundamental rights and promoting economic development through technological advancement.

To ensure that the regulatory framework keeps pace with the rapid technological advancements in artificial intelligence (AI), it is essential to adopt dynamic and adaptable regulatory instruments that foster innovation, legal certainty, and the protection of fundamental rights. A crucial first step would be the creation of a specialized AI regulatory agency, similar to Brazil's National Data Protection Authority (ANPD). This agency could monitor the development and use of AI systems, issue technical guidelines, and establish compliance standards. It would also play a key role in integrating international regulations, aligning Brazil with global frameworks such as the EU AI Act and the OECD guidelines. Moreover, its technical expertise would enable a more agile and proactive approach, allowing for quick responses to technological changes, reducing regulatory gaps, and ensuring greater effectiveness in enforcing AI-related regulations. By incorporating mechanisms for continuous legislative updates, fostering innovation, and aligning with international standards, Brazil could position itself as a leader in AI regulation, ensuring that technological advancements translate into inclusive and sustainable social benefits.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Arrabal, A. K. (2024). Autor-máquina. *Revista de Informação Legislativa: RIL, 61,* 101-122. https://www12.senado.leg.br/ril/edicoes/61/243/ril_v61_n243_p101

Brasil (1988a). *Emenda Constitucional nº 115, de 10 de fevereiro de 2022. Altera a Constituição Federal para incluir a proteção de dados pessoais como direito fundamental. Diário Oficial da União: Seção 1, Brasília, DF, p. 1, 11 fev. 2022.* https://www.planalto.gov.br/ccivil_03/constituicao/emendas/emc/emc115.htm

Brasil (1988b). *Constituição da República Federativa do Brasil: Atualizada até a Emenda Constitucional n° 134. Brasília.* http://www.planalto.gov.br/ccivil_03/constituicao/ConstituicaoCompilado.htm

Brasil (1998a). *Lei nº 9.609, de 19 de fevereiro de 1998. Dispõe sobre a proteção da propriedade intelectual de programa de computador, sua comercialização no País, e dá outras providências. Diário Oficial da União, Brasília.* http://www.planalto.gov.br/ccivil_03/leis/l9609.htm

Brasil (1998b). *Lei nº 9.610, de 19 de fevereiro de 1998. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências. Diário Oficial da União. Brasília.* http://www.planalto.gov.br/ccivil_03/leis/l9610.htm

Brasil (2016). *Superior Tribunal de Justiça. Recurso Especial n. 1.473.392/SP. Direitos autorais. Contrato sob encomenda. Pessoa jurídica. Titular de direitos do autor. Possibilidade. Direito à indenização. Obra utilizada sem a devida autorização. Relator: Ministro Luis Felipe Salomão. Quarta Turma, julgado em 11 out. 2016. Diário da Justiça Eletrônico, Brasília.*

https://scon.stj.jus.br/SCON/GetInteiroTeorDoAcordao?num_registro=201303778189&dt_publicacao=21/11/2016

Brasil (2018a). *Decreto nº 9.289, de 21 de fevereiro de 2018. Promulga o Protocolo de Emenda ao Acordo sobre Aspectos dos Direitos de Propriedade Intelectual Relacionados ao Comércio, adotado pelo Conselho-Geral da Organização Mundial do Comércio, em 6 de dezembro de 2005. Diário Oficial da União, Brasília.* http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/decreto/d9289.htm

Brasil (2018b). *Lei n.º 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, Brasília, DF, 15 ago.* https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm

Brasil (2023). *Projeto de Lei n° 2.338, de 2023. Dispõe sobre o uso da Inteligência Artificial. Brasília.* https://legis.senado.leg.br/sdleg-getter/documento?dm=9347622&ts=1733707591128&disposition=inline

Cozman, F. G., Plonski, G. A., & Neri, H. (2021). *Inteligência artificial: Avanços e tendências.* Universidade de São Paulo. Instituto de Estudos Avançados. https://www.livrosabertos.abcd.usp.br/portaldelivrosUSP/catalog/book/650

Diniz, M. H. (1999). *As Lacunas no Direito* (5th ed.). Saraiva.

European Union (2016). *Regulation (EU) 2016/679* of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation). *Official Journal of the European Union, 119,* 1-88. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

Grossi, B. M. (2022). *O desafio da regulação dos dados pessoais: Entre a autonomia e a heteronomia.* Master's Thesis, Pontifícia Universidade Católica de Minas Gerais. https://www.researchgate.net/publication/367075319_O_desafio_da_regulacao_de_dados_pessoais_entre_autonomia_e_heteronomia/link/63c0380a7ecd35045c4217e5/download?_tp=eyJjb250ZXh0Ijp7InBhZ2UiOiJwdWJsaWNhdGlvbiIsInByZXZpb3VzUGFnZSI6bnVsbH19

Hoffmann-Riem, W. (2021). *Teoria Geral do Direito Digital: Transformação digital desafios para o direito.* Forense.

Mello, M. T. L. (2010). Propriedade Intelectual e Concorrência. *Revista Brasileira de Inovação, 8,* 371-402. https://periodicos.sbu.unicamp.br/ojs/index.php/rbi/article/view/8648985 https://doi.org/10.20396/rbi.v8i2.8648985

Mèlo, A. (2019). *Proteção de Dados Pessoais na Era da Informação.* Juruá Editora.

Nunes, D., & Marques, A. L. P. C. (2018). Inteligência artificial e direito processual: Vieses algorítmicos e os riscos de atribuição de função decisória às máquinas. *Revista de Processo, 285,* 421-447.

OECD (2019). *Artificial Intelligence in Society.* OECD Publishing. https://www.oecd.org/en/publications/artificial-intelligence-in-society_eedfee77-en.html

OECD (2025). *AI Principles Overview.* https://oecd.ai/en/ai-principles

Rici, M. D. C. (2023). A autoria na era da inteligência artificial: Quem é o autor das obras geradas por atos autônomos de inteligência artificial? *Revista de Direito e as Novas Tecnologias*, 18, p. 4.

Samuelson, P. (2016). Allocating Ownership Rights in Computer-Generated Works. *University of Pittsburgh Law Review*, 47, 1185-1206. https://www.law.berkeley.edu/wp-content/uploads/2024/01/Pam-Samuelson-

Allocating-Ownership-Rights-in-Computer-Generated-Works.pdf

Silva, F. C. (2023). *Proteção de dados pessoais na era da inteligência artificial.* Trabalho de Conclusão de Curso (MBA)—Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. https://bdta.abcd.usp.br/directbitstream/7274f87b-0f7a-48f1-a070-05e2b5898f06/Felipe%20Casali%20Silva.pdf

Souza, B. D. A., & Jacob, R. R. C. (2022). Propriedade intelectual na criação de artes com inteligência artificial: O caso Midjourney. *Revista de Direito e as Novas Tecnologias, 16,* p. 6.