Legal Issues in the Digital Age. 2024. Vol. 5. No. 2.
Вопросы права в цифровую эпоху. 2024. Том 5. № 2.

Research article
УДК: 340, 347
DOI:10.17323/2713-2749.2024.2.36.56

Progress in Natural Language Processing Technologies: Regulating Quality and Accessibility of Training Data

Ilya Gennadievich Ilyin

Saint Petersburg State University, 7/22 Liniya V.O., Saint Petersburg 199106, Russia, i.g.ilin@spbu.ru, orcid: https://orcid.org/0000-0003-1076-2765

Abstract

Progress in natural language processing technologies (NLP) is a cardinal factor of major socioeconomic importance behind innovative digital products. However, inadequate legal regulation of quality and accessibility of training data is a major obstacle to this technological development. The paper is focused on regulatory issues affecting the quality and accessibility of data needed for language model training. In analyzing the normative barriers and proposing ways to remove them, the author of the paper argues for the need to develop a comprehensive regulatory system designed to ensure sustainable development of the technology.

Keywords

personal data; data regime; generative neural network; artificial intelligence; natural language processing; large language models; data access; copyright.

Acknowledgments: the paper is published within the project of supporting the publications of the authors of Russian educational and research organizations in the Higher School of Economics academic publications.

For citation: Ilyin I.G. (2024) Progress in Natural Language Processing Technologies: Regulating Quality and Accessibility of Training Data. *Legal Issues in the Digital Age,* vol. 5, no. 2, pp. 36–56. DOI:10.17323/2713-2749.2024.2.36.56

Background

The technology of natural language processing (NLP) is associated with mathematical linguistics and artificial intelligence and allows computers to understand and generate natural language [Hirschberg J., Manning C.D., 2015: 261–266]. As applied to information technologies, language and speech help to promote the engagement between man and computer as exemplified by digital products for processing and analysis of texts (spelling, grammar, duplication, readability checking services, etc.), text translators, voice assistants and other interactive response technologies (chat bots, automated client support systems etc.).

Progress in natural language processing is crucial both from the economic perspective as a key factor for development of artificial intelligence [Feng Z., 2023: 7–8, 25] with a potential for innovative digital products, and also from the social perspective in view of the importance to develop and preserve the natural language as a major aspect of the national and cultural identity.

Meanwhile, despite the innovative nature and socioeconomic value of the technology under discussion, the existing legal framework cannot fully support its sustainable development, a key trouble being normative barriers for access to training data with qualitative and quantitative parameters needed to achieve progress.

From the technical perspective, the urgency of the problem follows from the methods of natural language processing. The technology relies on generative neural networks to create large language models (LLM) [Glauner P., 2024: 24–34]. These models are trained on large data arrays including those structured as a linguistic corpus — a database containing numerous texts (books, transcriptions, translations etc.) and audio files (audio books, broadcasting recordings, podcasts and other audio content) — something that allows them to study the structure of natural language and "understand" different language contexts.

Large language models assume the use of not only available data but also those generated by the neural network on their basis. Such approach, on the one hand, considerably expands the amount of training data but, on the other hand, makes it more difficult to correct algorithmic errors and

defects. For instance, if training data contained defects that could affect the functioning of the algorithm, these defects would corrupt the data generated by the model. In this situation, removing corrupt data is technically difficult. One example of large language models is BERT[1], GPT-3[2] and the underlying digital products like Google Assistant or ChatGPT.

From the regulatory perspective, the issue has been identified in the relevant strategic planning documents, with the 2030 National AI Development Strategy[3] (hereinafter Strategy) as one of the key documents in the field. In the Strategy, normative barriers and a lack of methodological framework for support of AI systems with reliable data are referred to as obstacle for the development of artificial intelligence in Russia.[4] The Strategy calls to develop a comprehensive regulatory system for social relations related to the development and application of AI technologies[5], in particular, to remove excessive normative barriers and create an enabling regulatory environment for development and introduction of AI technologies[6], remove regulatory barriers for development and introduction of large generative models to be trained on large data arrays[7], and provide for regulatory support of AI developers' access to different types of data.[8]

---

[1] Generative Pre-trained Transformer (GPT) is a line of deep learning models developed by OpenAI (United States) and based on the Transformer architecture. Trained without a "trainer", it does not need to be adapted and can be used for a variety of tasks. For detail on GPT see: Yenduri G. et al. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions // arXiv preprint arXiv:2305.10435. 2023. For detail on the Transformer architecture see: Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. 2017. Vol. 30.

[2] Bidirectional Encoder Representations from Transformers (BERT) is a deep learning model designed by Alphabet Inc. (United States). Based of the Transformer architecture, it is trained on bidirectional context meaning an ability to analyze and understand contexts both from left to right and vice versa. For more detail on BERT see: Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding//arXiv preprint arXiv:1810.04805. 2018.

[3] The 2030 National Artificial Intelligence Development Strategy approved by Presidential Decree No. 490 "On the Development of Artificial Intelligence in Russia" of 10 October 2019 ("2030 National AI Development Strategy"). Here and elsewhere all references to documents, regulations, legal practice are taken from SPS Consultant Plus.

[4] Para 17(16) (e), 2030 National AI Development Strategy.

[5] Ibid. Para 24 (f).

[6] Ibid. Para 24 (f).

[7] Ibid. Para 51(11) (c).

[8] Ibid. Para 51(11) (b).

In view of the objectives set by the Strategy, the paper purports to provide a conceptual analysis of the problem to regulate the quality and accessibility of training data, and to identify and propose ways to address the underlying legal constraints.

In terms of its subject matter, the paper has three parts in addition to the background and conclusion. The first part explores the legal aspects related to the impact of data parameters on language models to be developed. The second part is focused on the existing legal arrangements that support the required data quality. The third part is devoted to the issues of accessibility of training data, analysis of normative barriers and discussion of the ways to remove them.

1. Data Parameters: Aspects of Impact on Language Models under Development

1.1. Data and Language Models: Interrelation and Technical Parameters

Progress in natural language processing technologies is largely hinged on the efficient language models developed for a particular language. These models are crucial for subsequent operation of available digital products and affect to what extent a computer is able to "understand" and process texts. A language model is created through a series of consecutive stages.

At first, training data are put together: this stage involves a large amount of textual and other language data from a wide range of sources. Training data for language models will normally include textual data (for instance, written texts, speech transcriptions and annotated lists), speech data (audio recordings, phonetic and prosodic annotations) and multimodal data (image-text, video-text and audio-text pairs) [Dash N.S., et al., 2018: 291].

Once collected, the data is pre-processed. This stage involves removal of noise (for instance, irrelevant information, errors, duplicates), text normalization (bringing to a common format), breaking a text into sentences and words, stop word removal, lemmatization (grouping together inflected forms) and stemming (stripping words down to their stems [Khyani D. et al., 2021: 350–357]. The purpose of pre-processing is to prepare data for mining and language model training [Goldberg Y., 2017: 65–76].

The next stage is training of the language model itself, with regularities, dependencies and peculiarities of the data in question identified through

the use of machine and deep learning algorithms. Language models could be trained to address a number of tasks: text classification, tonality analysis, named entity recognition, machine translation, etc. [Zhou M. et al., 2020: 275–290]. After the training, language models are evaluated on text data to check for efficiency and accuracy. A language model can be fine tuned and optimized depending on the evaluation's results.

Finally, the introduction of a language model assumes its accessibility for integration into the respective digital products. This process will require ongoing monitoring of its functioning with changes and improvements to be made as necessary, for example, to take account of technological innovations and user feedback. Due to ongoing improvement of the model, the stage of introduction is time consuming.

1.2. Functional Errors of Language Models: Legal Defects and Quality Defects

Quality and diversity of training data will directly impact the ability of a language model to be trained and to interpret texts in a given natural language. The structure of data including their arrangement and format, representativeness, amount and other parameters will affect the training process and accuracy of understanding a text's semantics and context. The use of data below the required qualitative/quantitative parameters will hinder further progress of the technology, only to result in negative implications in both technical terms — algorithmic errors due to falsely identified correlations and regularities — and legal terms like illegitimate restriction of rights and liberties (algorithmic discrimination), violation of privacy, personal and family secrets, occurrence of harm etc.

Training data defects could be regarded from two perspectives: firstly, incompatibility with specific technical criteria and metrics (quality defects) such as those of representativeness, amount, purity etc.; secondly, violation of the applicable legal regime (legal defects) such as personal data protection when data are processed as part of a language model.

It has a sense first discuss in more detail the implications of training data quality defects. It should be noted above all that quality defects will not inevitably bring negative outcomes. For example, a minor inaccuracy, insufficiency, irrelevance of training data, while not having a major bearing on common dependencies to be identified, could impact the findings of data analysis with regard to specific individuals [Hacker P., 2021: 260,

263]. The set of required quality parameters and respective metrical values should apparently differ in technical terms depending on the purpose of a given language model and its area of application.

In general terms, the data quality defect as applied to the natural language processing technology could contribute to the digital divide [Lythreatis S. et. al, 2022: 1–11] and cause language discrimination.

Digital divide is a kind of social inequality identified as impossibility for individuals or social groups to have equal access to information and communication technologies, as well as equal level of skill to use them [Rogers S.E., 2016: 197–199]. The urgency to address this problem has been underlined at the national[9] and international level.[10] In terms of law, the problem of digital divide will primarily affect the relations of constitutional law, in particular, the legal status of individuals, human and civil rights and liberties guaranteed by the state [Mushakov V.E., 2022: 69–73] including equal civil and human rights and liberties irrespective of the language.

Digital divide can manifest itself as language discrimination resulting in limited access of specific social groups to a technology due to impossibility to use it in a native language (limited choice of supported languages) or incorrect functioning due to specific dialect and peculiarities of the language spoken by the social group in question.

Article 2 of the Universal Declaration of Human Rights (1948)[11] prohibits discrimination including on the basis of language. A similar provision is set by Article 1 (3) of the UN Charter[12] as also reflected in paragraph 2, Article 19 of the Russian Constitution[13] whereby the state guarantees equal civil and human rights and liberties irrespective of language.

---

[9] Federal Government Resolution No. 313 "On approving the Information Society public program of the Russian Federation" of 15 April 2014 // SPS Consultant Plus.

[10] United Nations Declaration of Principles Building the Information Society of 12 December 2003. Available at: https://www.un.org/ru/events/pastevents/pdf/dec_wsis.pdf (accessed: 19.04.2024); UN Tunis Agenda for the Information Society of 15 November 2005. Available at: https://www.un.org/ru/events/pastevents/pdf/agenda_wsis.pdf (accessed: 19.04.2024)

[11] Universal Declaration of Human Rights (passed by the UN General Assembly 10.12.1948). Available at: https://www.un.org/ru/documents/decl_conv/declarations/declhr.shtml (accessed: 10.06.2024)

[12] United Nations Charter (passed in San Francisco 26.06.1945). Available at: https://www.un.org/ru/about-us/un-charter/full-text (accessed: 10.06.2024)

[13] Constitution of Russia (approved by universal vote on 12.12.1993 as amended in the course of all-Russia popular vote on 01.07.2020).

Progress in natural language processing technologies adds up a new form of discrimination where it occurs through inadequate digitization of languages rather than someone's guilty action.

A language model to be developed will require access to training data in a given language. Meanwhile, digital data for development of robust and accurate language models are not available for all languages. For example, if the training data set was limited and did not cover all dialects of a language, the functioning of the language model may be incorrect or inaccurate or fail altogether when processing a natural language incorporating such dialects. The differences of pronunciation, vocabulary and grammar can result in defective text or speech recognition and analysis. Moreover, such problems will not arise for a language with a high level of digitization and therefore high representativeness. A similar issue is also observed in respect of minor languages. Thus, while digitization of specific major languages (like English, Russian) is high, many digital products are still not available to speakers of minor languages, for example, Udmurt, Buryat, Tuvan. For this reason, technical and legal support of access to the relevant linguistic corpuses is critical for digitization of the said languages and thus for development of the technology in question and elimination of digital divides.

Data quality can be undermined both for objective reasons (for instance, insufficient digitization level) and because of wishful action to corrupt training data and thus change the language model's training outcomes. In practice, such action is called data poisoning [Russo A., Proutiere A., 2021: 3234–3241]. False examples introduced into the training data set could result in wrong outcomes produced by the model like corrupt and incorrect translation of documents by automatic translation systems, only to affect the accuracy and meaning of the information to be transmitted. In chat bots, this can result in wrong answers to user queries to dump down user experience, undermine trust in the technology and bring about related legal implications, such as violation of consumer rights to quality products/services[14], right to information[15], etc. Errors in text analysis systems can result in wrong interpretation of text tonality or content, something especially critical in analysis of public opinion or monitoring of social networks and fraught with major implications including wrong legal qualification of one's

[14] Article 4 of Federal Law No. 2300-1 "On Protecting Consumers' Rights" of 07 February 1992 (hereinafter "Law on Protecting Consumers' Rights").

[15] Ibid. Article 8.

actions that could be wrongly qualified as incitement of hatred or humiliation of human dignity.[16]

Where natural language processing is used in critically important sectors such as medicine, the implications of data poisoning can be especially harmful and cause considerable damage, for example, through a wrong diagnosis due to wrong interpretation of medical data, thus jeopardizing human life and health.

Meanwhile, correcting technical errors and removing poor quality training data from language models will cause the issue of algorithmic shadow left by such data [Li T.C., 2022: 480–505]. In the general sense, this problem means that even removed data will still impact the created language models. Thus, for example, removing personal data from a training data set does not fully prevent their further influence on the language model: algorithmic shadow will be still observed in its operation. This is fraught with violating the data subject's rights and questions the operational legitimacy of such model as a whole.

Algorithmic destruction — elimination of data through special algorithms — is among technological solutions advanced in modern studies of this domain to address the algorithmic shadow problem [Rahman A., 2020: 575–577]; [Schneier B., 2015: 448]. Some researchers believe technology can be successfully applied to deal with algorithmic shadow to guarantee the data removal right to data subjects, for example, as regards personal data processing [Li T.C., 2022: 505]. However, it is worth noting that the development of specific algorithms to remove corrupt data will come at a significant economic and technological cost. It suggests that using this method across the board to deal with algorithmic shadow, just as making it legally binding is premature and requires further study from both legal and technological perspectives.

2. Legal Mechanisms of Data Quality Assurance

2.1. Dualism of Approaches

Extreme importance of qualitative data parameters and potential impact on operation of language models suggest the need to assure these parameters in legal and technological terms. In view of the discussed regulatory

[16] Article 282 of the Criminal Code of Russia No. 63-FZ of 13 June 1996; Article 20.3.1, Administrative Code of Russia No. 195-FZ of 30 December 2001.

methods, two approaches to address this task can be proposed: normative approach based on imperative (centralized) method; and contractual approach based on dispositive (decentralized) method.[17]

Normative approach assumes that quality parameters will be established and assured via legally guaranteed mandatory technical requirements, standards, certification and control procedures, as well as directly by law. This will put in place general rules for all parties involved in AI development thus allowing to introduce stricter control. A downside of this approach may be its insufficient flexibility to adapt to changes, something likely to become critical in the context of rapid advance of information technologies.

Contractual approach, in its turn, relies on decentralized relations between the parties, with consensual data quality standards to enhance flexibility and adaptivity to varying demands and situations. However, that approach requires more complex engagement between the parties to legal relationships and cannot invariably guarantee that their interests are mutually observed (such as in case of an inadequate counterclaim under a paid service agreement, abuse by a stronger contracting party, etc.). With both approaches having upsides and downsides, the problem is likely to be efficiently addressed through a comprehensive solution combining certain elements of the approaches. It is useful discuss each of them in detail.

2.2. Normative Approach: Data Accuracy Principle

The number of regulations governing data quality is currently extremely limited, one regulatory source to be considered being the Federal Personal Data Law.[18] It establishes the principle of "data accuracy"[19] whereby data should be accurate, adequate and relevant for processing purposes. Moreover, the data that fall short of these criteria should be either deleted or corrected. This principle is echoed by the data subject's right to correct the underlying data.[20] Meanwhile, implementation of the said principle is problematic.

[17] The issue of qualification of regulatory methods is beyond the scope of the paper. Meanwhile, it should be noted that classification of regulatory methods is a subject of debate in doctrine. For example, the following methods are proposed: incentives and punishment, authorization (licensing), prohibition and enforcement.

[18] Federal Law No. 152-FZ "On Personal Data" of 27 July 2006 (as amended on 06 February 2023) (hereinafter "Federal Personal Data Law").

[19] Ibid. Para 6, Article 5.

[20] Ibid. Para 1, Article 14.

Firstly, the law does not specify to what extent personal data could fail to meet the criteria mentioned. Moreover, as was told above, a minor inaccuracy, inadequacy or irrelevance of data will not have a major impact under certain conditions.

Secondly, it is not clear how one can assess and measure the accuracy, adequacy and relevance of personal data with regard to processing purposes. For example, other countries' law will sometimes establish stricter requirements to data depending on processing purposes. Thus, Germany's Data Protection Act has a special provision on personal data processing for scoring — assessment of creditworthiness in the financial sector — that allows to use and process only the data obtained through a "scientifically acknowledged procedure of mathematical statistics".[21]

Implementation of this principle should apparently rely on the risk-oriented approach to allow for possibility to process in some cases the data that do not fully meet the required criteria while in other cases, on the contrary, specify and introduce stricter criteria for data processing.

Normative definition of data quality parameters through the said principle is also restricted by its inapplicability to all types of data since the Personal Data Law applies only to personal data processing.[22] Therefore, the said principle is applicable only to personal data processing. Moreover, now data cannot be invariably and unambiguously qualified as personal data, with difficulties concerning both the form of expression and qualification likely to arise at some processing stage. Overall, the issue is that the current definition of personal data[23] assumes a binary approach, that is, data can be either personal or otherwise. This approach does not take into account data for different individuals can be identifiable to a variable extent, for example, due to accessibility of other datasets [Oostveen M., 2016: 306], and that the current progress in information and computer sciences reveals different level of possible identifiability and related sets of risk [Kolain M., Grafenauer C., Ebers M., 2021:174]. In addition, it is noteworthy that data being processed could lose and acquire the relevant identifiability markers, that is, be dynamic rather than static. Therefore, data can be qualified as personal only at a specific stage of the language model's development. The

[21] § 31(1) Federal Data Protection Act (BDSG). Germany. Official English translation is available at: https://www.gesetze-im-internet.de/englisch_bdsg/englisch_bdsg.html#p0256 (accessed: 10.06.2024)

[22] Para 1, Federal Personal Data Law.

[23] Ibid. Para 1, Article 3.

data accuracy principle is thus applicable only to the data qualified as personal at the given stage than to all data processed at different stages of the language model's development.

2.3. Contractual Definition of Data Quality. Application of GOST

Regulating data quality through contractual terms is another approach. In this case, qualitative parameters could be described either explicitly with the help of the chosen technical criteria and specifications or with reference to the corresponding standards like GOST, or else via another applicable technical regulation.

Two types of contracts can be identified in the proposed context: those entered to settle the relationships with regard to data accessibility and use (such as a licensing agreement to deposit or use a database) and those not explicitly aimed at regulating the use of data but whose qualitative parameters are likely to impact significantly the relationships in question (such as a licensing agreement with the end user of a digital product).

In the first case, the parties will explicitly set the qualitative parameters of data in the relevant agreement. Thus, in order to deposit language data in the Common Language Resources and Technology Infrastructure (CLARIN)[24], the depositor will sign a licensing agreement describing qualitative parameters and forms of data to be uploaded, assigning responsibilities and also establishing the terms of payment and distribution of data based on sample licenses designed by the organization [Kelli A., Vider K., Lindén K., 2016: 13–24].

In the second case, the described qualitative parameters, terms of use and distribution will normally apply not to data but the underlying digital products. For example, before starting to use Yandex Speech Kit[25], users are required to accept the terms defining the procedure of use.[26] This situation will raise the question of whether the data (including qualitative pa-

[24] International infrastructure for support of research in the area of humanities and social sciences by providing access to various language resources and tools. For detail see: https://www.clarin.eu (accessed: 10.06.2024)

[25] A Yandex service allowing to transform text into speech (speech synthesis) and vice versa (speech recognition). See: URL: https://yandex.cloud/ru/services/speechkit?utm_referrer=https%3A%2F%2Fwww.google.com%2F (accessed: 10.06.2024)

[26] Speech Kit terms of use / Yandex Speech Kit. Available at: URL: https://yandex.ru/legal/cloud_terms_speechkit/ (accessed: 10.06.2024)

rameters) is relevant for the underlying digital product. Will the data lose independence, only to become its qualitative parameter? Who will be then responsible for the product's defects caused by questionable data: model developer or product developer? The answer to these questions is likely to be of principal importance both for performance under the said contracts and generally for the problem of contractual assurance of training data quality. Further, it should be noted that the question of assigning responsibility for harm caused by AI systems is debatable among researchers. It is generally proposed, firstly, to design risk minimization mechanisms already at the stage of AI system development; secondly, more specifically define who can assume responsibility for such harm; and, thirdly, apply a concept similar to that of "major hazard" in respect of AI systems [Kharitonova Yu.S., Savina V.S., Panyini F., 2022: 683-708].

One way to define quality data via contractual terms is to apply relevant technical standards such as intergovernmental standards (GOST). With regard to data, the fundamental document is GOST R ISO 8000-100-2019 Data quality[27] as well as GOST R ISO/MEK 20546-2021 Information technologies. Big data. Overview and glossary.[28] Key requirements to data quality such as accuracy, adequacy, relevance and consistency are defined in GOST R ISO 8000-100-2019 while GOST R ISO/MEK 20546-2021 provides an extensive overview and unification of the terms related to big data, something that helps to standardize the data processing approaches and establish a common conceptual framework for regulating the relations involved in language model training.

Technical Committee for Standardization No. 164 Artificial Intelligence (TK164) is currently in charge of developing relevant GOST applicable to AI and data.[29] The Committee is crucial for the development of regulatory framework for AI technologies in Russia, in particular, the rules that allow researchers and developers to have access to the required amount of data for efficient training of models and lower risk of unauthorized use of information. One standard under development in the discussed domain is Data quality for analytics and machine learning. The draft standard consisting of

[27] Rosstandard Order No. 836-st «On approving a national standard of Russia" of 29 October 2019.

[28] Rosstandard Order No. 632-st "On approving a national standard of Russia" of 13 July 2021.

[29] Set up by Rosstandard Order No. 1732 "On establishing the technical committee for standardization Artificial Intelligence" of 25 July 2019/ SPS Consultant Plus.

several parts follows ISO/IEC series 5259 international standard that equally consists of several parts that describe the principal concepts, terms and examples of defining data quality for analytics and machine learning, propose data quality model, measurement methodologies and guidance on data quality reports, and outline the data quality management process including risk management aspects and ways to meet the requirements to quality.[30]

Development and approval of the above GOSTs are supposed to greatly facilitate the issue of defining quality data in terms of both data parameters themselves and the applicable metrics and conceptual framework. Meanwhile, it should be noted that GOSTs will often contain rigid and detailed requirements that can be inappropriate or excessive in a particular case, only to complicate the adaptation of contractual terms to the parties' specific needs.

3. Regulating Access to Training Data

3.1. Personal Data

While defective data quality is normally related to technical shortcomings, legal defects primarily involve the problem of compliance with a legal regime of using data for training. Moreover, the problem itself is expressed in the form of conflict between the interests of developers critically in need of more or less free access to large amounts of data and the third-party interests protected by specific legal regime constraining such access.

The issue of compliance with data regime while developing AI models applicable, particularly, to personal data and other restricted information, as well as protection of intellectual property rights is recognized in the Strategy as one of the "challenges" faced by Russia in the area of AI development.[31]

As the issue of implications of personal data regime for language models to be developed and marketed was explored by the author in detail in previous studies, this paper will present only the main findings. One way to determine the extent of impact of personal data regime is to analyze physical, time-bound and territorial scope of the underlying regulation. Under this approach, physical impact can be determined in respect of different development stages of digital products and the extent of personal data use at each stage [Kelli A. et. al., 2021:154–159], while time limits by the effec-

[30] For detail on ISO/IEC series 5259 standard see: https://www.iso.org/ru/search.html?PROD_isoorg_ru%5Bquery%5D=ISO%2FIEC%205259 (accessed: 10.06.2024)

[31] Para 17(16), subparagraph (g), 2030 National AI Development Strategy.

tive term of data subjects' right to personal data protection, and territorial scope by national jurisdictions where the respective models are developed or marketed.

It is noteworthy that this approach reveals a number of shortcomings. For example, attempts to determine physical impact via different development stages show that data could lose and, on the contrary, acquire identifiability markers at some stage, only to considerably complicate their qualification as personal data.

Determination of time limits raises the issue of effective term of deceased persons' right to personal data protection. While not establishing such term, the law only prescribes that data in this case can be processed only if consented by successors where such consent was not given during the person's lifetime.[32] In absence of such term, no time limits of the underlying legal regime could be determined. It is equally noteworthy that, apart from the term problem, there is no order of priority in respect of successors who could give such consent, only to cause legal uncertainty in situations where some successors will withhold it while others not.

As for the territorial scope, the problem is in the need to comply with different national regimes at a time which is often impractical as, for example, in the case of the General Data Protection Regulation[33] and Russian personal data protection law. Meanwhile, it could become necessary due to both exterritorial effect of regulation itself, specific and related fields, and because of technical necessity to collect and process data in the national territory of other countries.

3.2. Protecting Intellectual Assets in Designing and Training Language Models

Whereas the impact of personal data regime on development of language models was explored by the author in detail in previous studies, the issue of the underlying use of intellectual assets received less attention. The urgency of this problem is confirmed by numerous cases of litigation be-

[32] Para 7, Article 9, Federal Personal Data Law.

[33] Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 of the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). In force since 25 May 2018. Available at: URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed: 10.06.2024)

tween language model developers and authors such as, for example, claims against OpenAI[34], language model developer for Chat GPT. Let us discuss the problem in more detail.

In the context of intellectual property law, data used for language models such as texts and audio files can be represented as items of copyright and related rights. Meanwhile, it is noteworthy that they will not be protectable across the board. Thus, copyright protectability criteria include creative components and objective form of a work,[35] with related rights exercisable to the extent that the copyright to the work used to create the item of related rights was observed[36] etc.

Depending on extent of protectability of copyright and related rights, the data used to develop a language model could be divided into three groups: "unprotected" (such as acts of legislation, official documents etc.), "safe" (such as manuals, technical specifications, expert opinions etc., all those generally not subject to protection) and works subject to copyright and related rights [Truyens M., Van Eecke P., 2014: 153–170]. A functional language model will require the components from all the three groups: the use of only "unprotected" and "safe" groups will not suffice. Meanwhile, it is technically problematic to draw a line for associating specific components with a particular group. Thus, though not all language model data will be subject to copyright and related rights, one cannot exclude the use of protected items for sure.

One caveat is in order regarding the concept of "using" the said items to develop a language model. Some researchers believe that the use of works for data mining — a stage of the model's development — does not involve copyright since it protects the creative form of expression while in data mining works are viewed as a database and are thus outside the available remedies [Kolsdorf M., 2021: 142–164]. This assumption is, in our view,

---

[34] See, for example, collective lawsuit, case No. 1:24-cv-00084, Nicholas Gage v Microsoft, OpenAI, United States District Court for the Southern District of New York. Available at: https://fingfx.thomsonreuters.com/gfx/legaldocs/klvydkdklpg/OPENAI%20COPYRIGHT%20LAWSUIT%20basbanescomplaint.pdf (accessed: 10.06.2024). Lawsuit, case 1:23-cv-11195, the New York Times company v. Microsoft, OpenAI, United States District Court for the Southern District of New York. Available at: https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf (accessed: 10.06.2024)

[35] Para 80–82, Federal Supreme Court Plenum Resolution No. 10 "On the application of Part IV of the Civil Code of Russia" of 23 April 2019.

[36] Para 3, Article 1303, Civil Code of Russia.

questionable. In the first place, the idea that copyright protects only the creative form is open to debate. Despite that this approach is explicitly reflected in law[37], research papers invoke a need to extend the scope of copyright to the work's content [Gavrilov E.P., 2009: 31–38] or else attempts to use the existing legal mechanisms to overcome the said constraint, for instance, by delineating the concepts of external and internal forms of a work [Kashanin A.V., 2010: 68–138]. Moreover, the use of works for data mining should be considered in conjunction with other related operations including those preceding mining such as copying, collection, transmission and classification of data. Except for temporary copying required for technological process and not amounting to the use of works[38], the said operations can involve intellectual property rights. The above is equally applicable to language models where data mining is just a development stage.

Using the items of copyright and related rights to design language models will require to comply with the author's personal non-property rights as well as the underlying exclusive rights. As such, the use of copyrighted items can rely on two patterns, the first based on the author's (other copyright holder's) prior consent (in the form of licensing agreement or that for assignment of exclusive rights), the second (doctrine of free use) restricting the author's (other copyright holder's) rights. While neither of the said patterns fully satisfies the industry's needs, they involve risks related to illegitimate use of intellectual property assets meaning violation of copyright and related rights.

The first pattern based on the author's prior consent to use copyrighted items for linguistic resources is apparently the least risky in terms of violation of copyright and exclusive rights. However, it raises an issue primarily related to identification of the author or other copyright holder who is often impossible to identify. It is further complicated by the question of how to go about the works created automatically or with minimum human involvement. Another trouble is that of time and cost of negotiations to conclude the respective agreements.

It is worth noting that large technological dotcom companies providing a wide range of digital services will often resort to such pattern. For example, the licensing agreement for Alisa voice assistant allows Yandex to

[37] Para 5, Article 1259, Civil Code of Russia (Part IV) No. 230-FZ of 18 December 2006 ("Civil Code of Russia").

[38] Para 2, Article 1270, Civil Code of Russia.

use voice prints borrowed not only from the application but also from the company's numerous other services.

While the second pattern based on the free use doctrine does not involve any time or cost in terms of author's consent and payment of royalties, its use in Russia is restricted to specific cases listed in law.[39] As developing a language model — language digitization and data mining — is potentially important for both science and culture, the model's use for "information, research, education and culture" is likely to be the most suitable of all cases of free use established by law.[40] However, the following analysis will reveal a number of complications to apply this exception.

With the invoked purposes to fit together, free use for research, education or culture also requires to specify the author and a borrowing source, allowing to use a work to the extent that fits the citing purposes [Gracheva D.A., 2023: 50]. These criteria are impractical to meet with regard to a language model.

Firstly, as was noted above, it is not always possible to exactly identify the authors of all works being used and thus make the respective references.

Secondly, the law does not say how to determine whether a work is used within the extent of the respective citing purposes. A functional language model will require a considerable amount of data to inevitably include protected works, with their number and extent of their citing likely to differ depending on the underlying technology and purposes. In absence of the criteria to determine the extent of possible citing, there will always be risk that in a given case the use of a work may be recognized as excessive in relation to purposes.

Thirdly, development of a model does not always serve only scientific and cultural purposes. In this particular case, the issue lies in the ratio of business and scientific/cultural purposes. The natural language processing technology has a scientific and social value, something that does not rule out its high economic potential. In this regard, the question is whether one could rely on the doctrine of free use to develop models for subsequent commercialization. It is logical to assume that where such model was originally developed by a business entity, this doctrine would generally be of no avail. Meanwhile, this situation is causing the number of potential produc-

[39] Sub-para 1, para 2, Article 1270, Articles 1273–1280, Civil Code of Russia.

[40] Article 1274, Civil Code of Russia.

ers to dwindle by excluding those without enough resources to rely on the first pattern based on prior consent, only to constrain the development of the entire sector.

Thus, the application of both first and second patterns to use protected works for designing language models and developing NLP technologies is now thwarted.

3.3. Data Dissemination: Repositories and Re-use

A possible solution to the above issue is to encourage higher education institutions to engage in the development and creation of language databases (linguistic corpuses) for further dissemination via a licensing agreement system. As language digitization has a high social value, the involvement of universities in this process appears logical and reasonable. There are examples of partnership between business entities and universities for development of natural language processing technologies such as ABBY chair at the Moscow Institute of Physics and Technology (MIFT), or the joint academic program of the Tsentr Rechevykh Tekhologiy company group and the ITMO National Research University. However, the development of linguistic corpuses on the basis of universities, while addressing the problem of targeted use of data to directly create linguistic corpuses and develop language models, leaves out the issues of their further dissemination. Thus, what will happen if a university loses interest in further dissemination of a linguistic corpus for some reason or other, or does not have enough funds to do it? On the contrary, can a university create a linguistic corpus through free use of works and then commercialize the outcomes relying on the concept of entrepreneurial university through a spin-off company? All these questions are currently open and urgent and require further in-depth study and analysis from the perspective of both jurisprudence and other sciences.

Another possible solution to the data accessibility problem is to make the data at state information systems (SIS) available to developers, that is, allow to re-use the already accumulated data. Re-use of SIS data for designing language models can considerably expedite the process of development and introduction of new technologies as well as enhance their effectiveness and adaptivity to various areas of application. As stated by the Federal Accounting Chamber in an analytical report[41], there were over 800 federal SIS

[41] Available at: https://ach.gov.ru/upload/pdf/Оценка%20открытости%20ГИС%202020.pdf (accessed: 10.06.2024)

in Russia in 2020 for support of information exchanges between public authorities in various social spheres. These systems contain data ranging from statistics to education, health and other socially important sectors. The use of SIS data in the interest of technological development thus appears to be quite promising.

Despite a varying degree of maturity of such systems, there is a reason to assume that SIS data will be of sufficient quality while their diversity will ensure representativeness. This will lay down a robust foundation for designing high quality, comprehensively trained language models capable of addressing widely diverse tasks. However, this will only be possible if the specifics of each type of data and their adequacy for the given purpose are carefully accounted for.

Meanwhile, data re-use is fraught with a number of legal and ethical risks related to both compliance with legal regimes (such as tax secret, personal data) and transparency and safety. Preventing the said risks will apparently require to develop common regulatory principles and approaches to data re-use including clear legal provisions and standards of data protection and data subject rights, as well as generally enhance control and audit mechanisms for the use of data to develop AI systems.

Conclusion

The paper was designed to provide a conceptual analysis of the regulatory problem for quality assurance and accessibility of training data in the context of the Strategy's objectives.[42]

Firstly, with regard to data quality assurance, likely implications of using corrupt data were explored and discussed from the perspective of undermining both technical parameters of data (quality defect) and legal regime (legal defect). Secondly, two approaches to data quality assurance were analyzed: normative and contractual. Despite their inherent downsides, it is feasible to use and apply both approaches in developing relevant regulation.

With regard to data accessibility, the research has allowed to identify and describe a number of constraints to use data for training. These constraints come in the first place from normative barriers that impede access to data due to a need to comply with the underlying legal regimes, as well as from a lack of adequate legal mechanisms to override them. These constraints to

[42] 2030 National AI Development Strategy.

a large extent slow down the process of development and introduction of language models to undermine the technology's progress as a whole as well as digital transformation of various economic and social sectors.

Progress of the technology will largely depend, on the one hand, on cooperation between all of the sector's stakeholders and, on the other hand, on the availability of modern regulation to support its sustainable development.

References

1. Dash N.S., Arulmozi S. (2018) History, features, and typology of language corpora. Singapore: Springer, p. 291.

2. Feng Z. (2023) Formal analysis for natural language processing: a handbook. Berlin: Springer Nature, pp. 7,8, 25.

3. Gavrilov E.P. (2009) Copyright and the content of artistic work. *Patenty i litsenzii*=Patents and Licenses, no. 7, pp. 31–38 (in Russ.)

4. Glauner P. (2024) Technical foundations of generative AI models. Legal Tech — Zeitschrift für die digitale Anwendung, pp. 24–34.

5. Goldberg Y. (2017) Features for textual data. In: Neural network methods for natural language processing. Cham: Springer, pp. 65–76.

6. Gracheva D.A. (2023) Free use of copyright and related rights in the context of development of digital technologies in Russia. *Trudy po intellektualnoy sobstvennosti*=Works on Intellectual Property, vol. 45, no. 2, pp. 44–52 (in Russ.)

7. Hacker P. (2021) A legal framework for AI training data—from first principles to the Artificial Intelligence Act. *Law, Innovation and Technology*, vol. 13, no. 2, pp. 257–301.

8. Hirschberg J., Manning C.D. (2015) Advances in natural language processing. *Science*, vol. 349, no. 6245, pp. 261–266.

9. Kashanin A.V. (2010) Development of ideas on the form and content of works in the copyright doctrine. The problem of protectability of research works. *Vestnik grazhdanskogo prava*=Bulletin of Civil Law, vol. 10, no. 2, pp. 68–138 (in Russ.)

10. Kelli A., Vider K., Lindén K. (2016) The regulatory and contractual framework as an integral part of the CLARIN infrastructure. CLARIN Annual Conference. Linköping University Electronic Press, pp. 13-24. Available at: https://helda.helsinki.fi/server/api/core/bitstreams/1f7b8a3c-790c-4e66-9677-f5f9aca785d6/content (accessed: 04.07.2024)

11. Khyani D. et al. (2021) An interpretation of lemmatization and stemming in natural language processing. *Journal of Shanghai University for Science and Technology,* vol. 22, no. 10, pp. 350–357.

12. Kolain M., Grafenauer C., Ebers M. (2021) Anonymity assessment-a universal tool for measuring anonymity of data sets under the GDPR with a special focus on smart robotics. *Rutgers Computer & Technology Law Journal,* vol. 48, p. 174.

13. Kolzdorf M.A. (2021) Free use of the items subject to copyright and related rights in Big Data processing. *Zakon*=Law, no. 5, pp. 142–164 (in Russ.)

14. Li T.C. (2022) Algorithmic destruction. *Southern Methodist University Law Review,* vol. 75, pp. 480-505. DOI: https://doi.org/10.25172/smulr.75.3.2

15. Lythreatis S. et al. (2022) The digital divide: a review and future research agenda. *Technological Forecasting and Social Change,* vol. 175, pp. 1–11.

16. Mushakov V.E. (2022) Constitutional human rights in the context of addressing the digital divide. *Vestnik Sankt-Petersburgskogo universiteta* MVD=Bulletin of Saint Petersburg University of Interior Ministry, no. 1, pp. 69–73 (in Russ.)

17. Oostveen M. (2016) Identifiability and the applicability of data protection to big data. *International Data Privacy Law,* vol. 6, no. 4, pp. 299–309.

18. Rahman A. (2020) Algorithms of oppression: how search engines reinforce racism. *New Media & Society*, vol. 22, no. 3, pp. 575–577. DOI: https://doi.org/10.1177/1461444819876115.

19. Rogers S. E. (2016) Bridging the 21st century digital divide. *TechTrends*, vol. 60, no. 3, pp. 197–199.

20. Russo A., Proutiere A. (2021) Poisoning attacks against data-driven control methods. 2021 American Control Conference (ACC). IEEE, pp. 3234–3241. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9482992 (accessed: 04.07.2024). DOI: 10.23919/ACC50511.2021.9482992.

21. Schneier B. (2015) *Data and Goliath: the hidden battles to collect your data and control your world.* N.Y.: Norton, 448 p.

22. Truyens M., Van Eecke P. (2014) Legal aspects of text mining. *Computer Law & Security Review,* vol. 30, no. 2, pp. 153–170.

23. Zhou M. et al. (2020) Progress in neural NLP: modeling, learning, and reasoning. *Engineering*, vol. 6, no. 3, pp. 275–290.

Information about the author:

I.G. Ilyin — Postgraduate Student.

The article was submitted to editorial office 21.06.2024; approved after reviewing 29.06.2024; accepted for publication 29.06.2024.