Research article
UDC 34:004:34.096:004.8
EDN: https://elibrary.ru/acxhto
DOI: https://doi.org/10.21202/jdtl.2024.8

Personal Data in Artificial Intelligence Systems: Natural Language Processing Technology

Ilya G. Ilin

Saint Petersburg State University, Saint Petersburg, Russia

Keywords

artificial intelligence,
biometric data,
digital technologies,
law,
lawfulness,
legal risk,
LegalTech,
legislation,
natural language processing technology,
personal data

Abstract

Objective: to conceptualize, from the viewpoint of personal data protection legislation, the development of natural language processing technology, identifying possible legal barriers to such development and directions for further research of the issue.

Methods: the research is based on general scientific methods of cognition, along with which formal-legal and comparative-legal methods were applied, as well as the method of theoretical modeling.

Results: it was found that the observance of personal data regime in the development of natural language processing technology leads to a conflict between private-legal and public-legal interests, which, in turn, creates obstacles for further development of this technology. The shortcomings of the existing legal order are shown, namely, the insufficient correspondence to the technical features of technology development. This may lead to the risks of excessive regulation, or, on the contrary, to the risks of neglecting critical areas that require protection. Problems in qualifying the data involved in the technology development are outlined. An attempt is made to define the limits of ensuring the lawfulness of personal data processing within the natural language processing technology. The material, temporal and territorial effect of the legal regulation in this field is identified as the limits of ensuring the legality. The author touches upon the possibility of using personal data as a consideration, which is important for the development of natural language processing technology and for the improvement of the information and communication technology industry.

Scientific novelty: the paper supplements the scientific discussion on the legal regulation of personal data processing by artificial intelligence systems with an analysis of natural language processing technology. The latter is insufficiently studied, making it relevant to research information law, namely, the legal relations arising around artificial intelligence systems, and to assess the impact of a personal data regime on the development of natural language processing technology.

Practical relevance: the applied aspects of the problems researched and the results obtained can be used to improve the legal regulation of public relations in the field of creation and development of artificial intelligence, as well as to identify and assess the legal risks arising in the personal data processing by developers of digital products based on natural language processing technology.

For citation

Ilin, I. G. (2024). Personal Data in Artificial Intelligence Systems: Natural Language Processing Technology. *Journal of Digital Technologies and Law, 2*(1), 123–140. https://doi.org/10.21202/jdtl.2024.7

Contents

Introduction

Natural language processing (NLP) technology is actively used in digital goods and services (digital products) to build human-computer communication (Hirschberg & Manning, 2015; Truyens & Van Eecke, 2014). Voice assistants, spell-checking services, text translation and voicing, voice biometrics, and interactive voice response systems are all examples of products using this technology.

The use of language and speech in information technologies opens up new prospects in artificial intelligence development, creates opportunities for the development of innovative digital products that contribute to the digital transformation of society, such as ChatGPT and its analogs, and determines the LegalTech industry development. The high economic potential and social significance of the technology, due to the importance of language for national and cultural identity (Hobsbawn, 1996), determine the interest of society, business and government in its advancement.

In turn, despite its innovative nature and socio-economic benefits of this technology, the analysis of the current legal order indicates the existence of legal obstacles to its development.

From a technical point of view, one of the main challenges for the development of natural language processing technology is the creation and subsequent dissemination of linguistic corpora (electronic speech resources). In a broad sense, linguistic corpora can be understood as databases containing a variety of texts (books, text transcriptions, translations, etc.) and audio files (audiobooks, broadcast recordings, podcasts, other audio content), which are subsequently used in machine learning algorithms (Kelli et al., 2012; Ilin, 2019). The creation of linguistic corpora involves sequentially passing two stages: language digitization – the collection and translation of data into a machine-readable format and their subsequent intellectual analysis (text and data mining, TDM) (Jents & Kelli, 2014; Ilin, 2019). Without massive linguistic corpora available, as well as without the stakeholders' ability to access them, technology will not develop and work.

From a legal point of view, a linguistic corpus as it is and its individual elements may include personal or other data with a special legal regime, such as data protected by secret communication. Accordingly, the relations involving the creation and dissemination of linguistic corpora will be simultaneously influenced by the legislation on personal data protection and by special sectoral legislation.

The article objective is to conceptualize, from the viewpoint of personal data protection legislation, the development of natural language processing technology, to identify possible legal obstacles to such development, and to propose directions for further elaboration of the problem. The research result complements the academic discussion on the legal regulation of personal data processing by artificial intelligence systems (Sviridova, 2021; Konev, 2020; Egorova et al., 2021) by analyzing the problem in the context of this technology functioning. It can be used in practice to identify and assess the legal risks arising from the personal data processing by developers of digital products based on natural language processing technology.

1. Personal data within the linguistic corpus

1.1. Definition and legal regime

As was outlined above, a key process for the development of natural language processing technology will be the creation of linguistic corpora and their dissemination among stakeholders. Elements of a linguistic corpus can be qualified as personal data. The modern approach to the definition of the "personal data" concept, based on para. 1, Art. 3 of the Federal Law "On Personal Data"[1], provides for a broad interpretation of the term and for qualifying as personal data virtually any data that directly or indirectly allow identifying a natural person. This leads to the inevitable emergence of personal data as part of a linguistic corpus.

Personal data require a special regime of legal and technical protection. In a general sense, the special legal regime of personal data, on the one hand, is aimed at ensuring the protection of rights belonging to the data subject; on the other hand, it imposes a number of restrictions on the use of such data in the creation of a linguistic corpus (Ilin, 2020). In this regard, the task of distinguishing the personal data used from other types of data will be fundamental for the creation of a linguistic corpus. In practice, this is not always possible: the boundary between personal and other data is not always clear.

First of all, there is a problem in defining the very concept of "data". Paragraph 1 of Art. 3 of the Federal Law "On Personal Data"[2] defines data as "any information...". In turn, Art. 2 of the Federal Law "On Information, Information Technologies and Information Protection"[3] defines information as "data (messages, data)...". In other words, from the correlation of these definitions we may conclude that data is data, i.e. the concept is defined through itself. This may create difficulties when trying to determine the form in which personal data can be expressed and, consequently, to qualify a piece of information as personal data.

Second, one of the most problematic aspects in qualifying data as personal data is that the current legislation relies on a binary approach in defining the concept of personal data. According to this approach, data can be either personal or not. In the author's opinion, the binary approach to the definition of personal data does not sufficiently take into account the current state of digitalization of society, the level of technological development, and the socio-economic changes that have occurred over the past few years. For example, such a definition does not take into account that data science and computer science distinguish different levels of possible identifiability and attribute a certain set of risks to each level (Kolain et al., 2022). In addition, it does not take into account the fact that data may be identifiable for one subject, for example, in combination with other datasets, but not for others (Oostveen, 2016).

Third, the status of data during processing can also be dynamical and not static (Purtova, 2018). In other words, during processing, data can become personalized or, conversely, lose this status. For example, during creating a linguistic corpus and a language

[1] On Personal Data. No. 152-FZ of 27.07.2006 (ed. of 06.02.2023). SPS KonsultantPlyus. https://clck.ru/gLnFq

[2] Ibid.

[3] On information, information technologies and information protection. No. 149-FZ of 27.07.2006. SPS KonsultantPlyus. https://clck.ru/3967pC

model based on it, the personal data used may lose identity markers and thus lose the status of personal (Kelli et al., 2021). This means that, in a practical sense, data can only be qualified as personal data at a certain point in time or stage of processing, which can make it difficult to respect the legality of the entire processing.

The dynamic nature of data is of fundamental importance to and affects, firstly, distinguishing personal data from other types of data (for example, there are fewer requirements for the processing of non-personal data); secondly, it influences the choice of the processed personal data category.

Legislation in the sphere of personal data protection identifies four main categories of personal data: “general”, “biometric”, “special” or “sensitive”, and “personal data authorized for dissemination”. Biometric data refer to the biological and physiological characteristics of an individual that can be used for identification[4]. The special or sensitive category of data is the data that indicate “political opinions, racial or ethnic origin, philosophical or religious views, health status, or sexual orientation”[5]. Personal data authorized for dissemination are the data that, with the subject’s consent, may be disseminated and made available to an unlimited number of persons[6]. This categorization of personal data is an initial prerequisite for data processing. For example, biometric data can be processed only after obtaining the explicit consent of the data subject. Processing of the special category of personal data is, as a general rule, prohibited. Different categories of data have different requirements to ensure the level of protection, both technical and legal (Krivogin, 2017). Hence, qualifying data as personal data poses the task of not only distinguishing personal data from other types of data, but also determining the belonging of personal data to a certain category.

1.2. Qualification of voice as personal data in various jurisdictions

In the context of creating a linguistic corpus, the use of speech data – human voice samples – illustrates the complexity of the task described above. Natural language processing technology relies heavily on the processing of said data. For example, they are used in voice assistants, text transcription systems, real-time translation of conversations, etc. The possibility of categorizing voice as personal data and choosing the appropriate category may cause a number of difficulties in practice.

[4] On Personal Data. No. 152-FZ of 27.07.2006 (ed. of 06.02.2023). Art. 11. SPS KonsultantPlyus. https://clck.ru/gLnFq

[5] Idid. Art. 10.

[6] Idid. Art. 10.1.

The proposed broad approach to the definition of personal data allows assuming that voice contains enough personal identifiers to be classified as personal data. This approach can also be found in the practice of the European Court of Human Rights (ECHR)[7], whose jurisdiction until recently extended to Russia. At the same time, the current Russian legislation, as well as the existing judicial practice, does not provide an unambiguous answer as to whether voice shall always be categorized as personal data. In practice, one may find cases when a recorded voice is not qualified as personal data (Arkhipov & Naumov, 2016). Indeed, it can be assumed that the voice will not always qualify as personal data, for example, due to distortion, brevity of the sample, or lack of associated identifiers.

At the same time, it seems that the probability of qualifying the voice as personal data is quite high, which raises the question of its categorization. Of particular interest is the division between the general category and the categories providing for a higher level of protection. From the viewpoint of the latter, voice can be considered as biometric or health data. For example, voice can be used as a biomarker to detect an early stage of Parkinson's disease (Tracy et al., 2020) or to provide information about an individual's mental state, stress level, emotional state, sleep deprivation, and other health-related data (Chang et al., 2011). Voice as a unique characteristic of an individual (biometric data) is often used to establish a person's identity (voice biometrics) (Jain et al., 2004). Voice may also contain personal data that fall under a general category. For example, such data may appear in speech content or sample-related metadata (age, gender, etc.).

The relevant clarifications of Roskomnadzor[8], now repealed[9], recommended that biometric data could be distinguished from other categories of data based on the purpose of processing. If the purpose of processing is related to personal identification, such data are to be qualified and processed as biometric data, otherwise they are not. This approach seems logical; otherwise, for example, the broadcasting of any TV or radio program would mean biometric data processing and would require compliance with the relevant legal regime. It is reasonable to assume that the same approach could be applied to health data. Obviously, in most cases, health data are not required to create a linguistic corpus and operate the natural language processing technology; therefore, as long as such data are not extracted, voice processing would not require a special legal regime.

[7] Case of S. and Marper v. the United Kingdom (App. No 30562/04, 30566/04). 04.12.2008. ECtHR. §84. https://clck.ru/3968V5

[8] Roskomnadzor. (2013, 2 September). Clarifications on referring photo- and video-images, dactyloscopic data and other information to biometric personal data and features of their processing. https://clck.ru/3968ky

[9] Letter of Roskomnadzor of 19.11.2021 No. 09–78548 "On irrelevance of clarifications by Roskomnadzor". SPS KonsultantPlyus. https://clck.ru/3968n6

However, it should be noted that, regardless of the purpose for which the voice is processed, it still contains biometric and health data; therefore, there is always a risk that this data could be extracted from the voice. This raises a general and more conceptual question: is the presence of such a risk sufficient to place personal data in categories with a higher level of protection, such as the "special" or "biometric" category? As far as the author is aware, there is no judicial practice on this issue in Russia. If we turn to other jurisdictions, we can find a similar case in the practice of the Court of Justice of the European Union[10]. In the case in question, the court used a broad approach to the interpretation of the "special" data category and included data that do not directly belong to it, but could potentially be extracted. In our opinion, this approach is at least debatable. Firstly, it is not quite clear how to determine that the data used contain hidden data, which in case of extraction may change its qualification. Their presence is not always obvious. Secondly, categorizing data as "special" means that they cannot be processed, except in specific cases listed in Art. 10 of the Law on personal data or Art. 9 of the General Data Protection Regulation[11]. In such a situation, it is difficult to ensure that data processing complies with legal requirements; this, in turn, may make it impossible to use the data to the extent necessary and, consequently, to achieve the purposes for which the data were processed.

In addition, it is important that the risk level of potentially extracting additional information that could change the data qualification varies depending on the particular case of processing. For example, speech data in a linguistic corpus are not publicly available, and the range of persons who can access and use it to extract additional information is limited. Moreover, this requires specialized knowledge and technical means. In other words, the ordinary processing of data within a linguistic corpus has a low risk of extracting "special" or "biometric" data from the voice. Hence, it seems that, on the one hand, the broad definition of data categories increases the level of protection of the data subject rights; on the other hand, it largely limits the possibilities to use them for technological development. In this situation, in order to balance the interests of the data subject and those who process data, it is important not only to identify the risk of potential information extraction but also to technically and legally assess that risk.

[10] Case C-184/20 OT v Vyriausioji tarnybinės etikos komisija. EU:C:2022:601. https://clck.ru/3969Ee

[11] Article 10 of Federal Law "On Personal Data", Art. 9 of 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, and repealing Directive 95/46/EC (General Data Protection Regulation) (2016, April 27). https://clck.ru/3969GL

2. Limits of ensuring the lawfulness of personal data processing as part of natural language processing technology

The use of personal data to create a linguistic corpus and develop natural language processing technology necessitates their processing in strict compliance with the law. The Russian legislation broadly interprets the concept of personal data processing[12], so there is little doubt that actions performed with data while using them as part of natural language processing technology will qualify as personal data processing (Ilin & Kelli, 2020). In turn, this poses the question of the limit to which data processing must comply with legal requirements. For example, if a language model was created using personal data, does this mean that further use of products created using this model is also subject to the personal data protection law?

The limits in ensuring the lawfulness of personal data processing as part of natural language processing technology can be defined through the material, temporal and territorial effects of the legal regulation in this field (Ilin, 2020).

In the context of natural language processing technology, the material effect can be defined through the various levels of personal data use in developing the relevant digital products. These levels include gathering material to create a dataset, compiling the dataset, annotating, creating a language model, and developing the final digital product (Kelli et al., 2021). For example, a language model consists of linguistic rules and it is currently almost impossible to extract any personal data from it, as the identifiability of data subjects is lost in the process of language model creation. Therefore, the treatment of the data used to create the model will generally not correspond to the legal treatment of the language model built on that data (Kelli et al., 2021).

The time limits for ensuring the lawfulness of data processing can be defined through the period during which the data subject's right to their data protection is in force. However, the Russian legislation does not set a time limit for the validity of such a right, indicating the need to comply with personal data regulation, including in relation to the data of deceased persons[13], and without any indication of time limits. One may agree with the opinion of some researchers that this gap should be bridged, for which it would be advisable to set the term of personal data protection equal to that of a person's privacy protection (Vazhorova, 2012).

The limits of ensuring the lawfulness of data processing in terms of territorial validity will be determined taking into account the national jurisdictions in which the relevant digital products are created or distributed. The problem is that digital products built using natural

[12] On Personal Data. No. 152-FZ of 27.07.2006 (ed. of 06.02.2023). Para. 3 Art. 3. SPS KonsultantPlyus. https://clck.ru/gLnFq

[13] Ibid. Para. 7 Art. 9.

language processing technology usually do not focus on one country, but seek to cover markets in different countries. For example, Alice the voice assistant by Yandex supports the Turkish language; the speech-to-text system by Google supports over 120 languages and can be integrated with other digital products. Therefore, the question arises about the need to comply not only with national legislation on personal data protection, but also with the relevant regulation of other countries – for example, Turkish law – to collect data and form a linguistic corpus in Turkish.

Russian legislation in the field of personal data protection as a general rule has no extraterritorial effect. Consequently, it does not apply to non-residents processing personal data of Russian citizens abroad. This rule has two exceptions. The first relates to the requirement of data localization, and the second relates to cases of state security.

The rule of localization of processing of the personal data of Russian citizens obliges data processors to store, collect and use personal data of Russian citizens only in databases located in the territory of the Russian Federation. This rule applies: if the information being processed contains personal data; if these data were collected (i.e., received from third parties) and processed; and if these data are related to Russian citizens (Savelyev, 2016). The latter condition raises the problem of determining citizenship within the framework of the use of information technologies.

This problem was partially resolved after Roskomnadzor provided clarifications according to which citizenship should be understood as the territory where processing takes place; if there are doubts about the data subject's citizenship, all information processed and collected in Russia should be localized in databases located in Russia. At the same time, the issue of identification and processing of personal data of Russian citizens that are collected outside the Russian jurisdiction is still open (Ilin, 2020).

An example of cases where state security measures have an extraterritorial effect on personal data protection legislation is the requirement for organizers of information dissemination and providers of telecommunication services to store Internet traffic (voice and text messages, photos, videos, sounds, file metadata), as well as to provide encryption keys to decrypt Internet traffic if the required data are stored or processed in encrypted form[14]. It is important that this requirement is not limited to any territory. It should be noted, however, that compliance with these rules may be difficult for companies, as they

[14] On amendments to the Federal Law "On Combating Terrorism" and certain legislative acts of the Russian Federation with regard to the introduction of additional measures to combat terrorism and ensure public security. No. 374-FZ of 06.07.2016. (2016). SPS KonsultantPlyus. https://clck.ru/3969WT; On amendments to the Criminal Code of the Russian Federation and the Criminal-Procedural Code of the Russian Federation with regard to the establishment of additional measures to counter terrorism and ensure public security. No. 375-FZ of 06.07.2016. (2016). SPS KonsultantPlyus. https://clck.ru/3969aF

will be forced to violate their national data protection regulation. For example, some provisions of the General Regulation on personal data protection[15], which regulates legal relations in the field of personal data protection in the EU, contradict the above requirement (Ilin, 2020).

3. Using personal data to pay for digital products based on natural language processing technology and legal qualification issues

In today's digital economy, products using natural language processing technology are widely marketed under a business model that does not involve a monetary reward from the product user. Instead, the vendor benefits from the use of data that have either been intentionally provided by the user or automatically collected or created by the vendor. The application of this business model and the payment for goods and services with data has opened a debate about the extent to which data per se can be a means of payment. First and foremost, doubts arise regarding the economic and legal characteristics of data. While one of the main concerns about the economic characteristics of data is the problem of measuring their economic potential and value, which cannot be determined either by the market or by contract (Lohsse et al., 2020), the legal discussion arises around the definition of the personal data legal regime, the relationship governing the legal rules, and the content of legal relations in this area (Helberger et al., 2017; Metzger, 2017; Svantesson, 2018).

The use of natural language processing products involves an intensive data exchange between the user and the provider (Goldberg, 2017). The data can be intentionally transmitted by the user to the provider (e.g., by voice command), collected by the provider (e.g., recording a voice sample, location data), or created by the service (e.g., the result of a text translation). In most cases, the provider uses the collected data not only to provide the user with a digital product, but also to further develop and improve it, as well as for commercial purposes not directly related to the marketed product. For example, a supplier can use the user's voice to analyze the emotional reaction to the advertising content and use the analysis result to sell other goods and services (Sartor, 2020). The possibility of such use of data raises the question of its qualification, the applicable legal regime, and the possibility of defining data as an object of property right (Talapina, 2020). In addition, the use of personal data as a consideration raises

[15] 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, and repealing Directive 95/46/EC (General Data Protection Regulation) (2016 April 27). https://clck.ru/3969dk

the question of the legal concept of personal data as one of the fundamental human rights – the right to respect for private and family life. It seems that the current legislation in the field of personal data protection does not in itself prohibit such use of data, but imposes a number of limitations, in a general sense related to the need to ensure the legality of such processing (Saveliev, 2021).

At the same time, the use of data as a consideration raises another issue concerning the recoverability of civil law contracts used for the practical application of the business model in question and, consequently, the possibility of extending the law on consumer protection to such relations. On the one hand, para. 1 Art. 423 of the Civil Code of the Russian Federation[16] explicitly stipulates, as one of the characteristics of recoverability, the execution of "another consideration" for fulfillment of contractual obligations, while providing the user with additional guarantees and legal mechanisms for the consumer rights protection balances the user's relations with the supplier, in relation to whom the user has a weaker position.

On the other hand, the use of data as a new means of payment may cause problems in law enforcement practice. For example, in order to apply consumer protection measures that directly depend on the product or service value, it is necessary to determine the monetary equivalent of the transferred data, which, given the economic and legal nature of data, seems difficult to implement.

Conclusions

The article objective was to conceptualize, from the viewpoint of personal data protection legislation, the development of natural language processing technology. The analysis of the existing legal framework has shown that it does not fully correspond to the technical features of the technology development, which may lead either to over-regulation or, on the contrary, to the neglect of critical areas requiring protection. In particular, problems arise in qualifying the data, involved in the technology development, as personal data and in their subsequent categorization. This problem is well illustrated by the example of the use of speech data – samples of a person's voice. An attempt to define the limits of ensuring the legality of data processing also causes difficulties and requires further clarification both in terms of substantive, temporal and territorial validity of legal regulation in this area. Another issue that needs further research is the possibility of using personal data as a consideration. This problem appears relevant not only for the development of natural language processing technology, but also for the development of the information and communication technology industry in general.

[16] Civil Code of the Russian Federation (part 1) of 30.11.1994 No. 51-FZ, ed. of 24.07.2023. SPS Konsultant-Plyus. https://clck.ru/3969nv

Further research may be aimed at analyzing and developing solutions that will contribute, on the one hand, to overcoming the revealed legal obstacles to the technology development, and, on the other hand, to improving the legal protection of critical areas that require additional protection.

References

Arkhipov, V., & Naumov, V. (2016). The legal definition of personal data in the regulatory environment of the Russian Federation: Between formal certainty and technological development. *Computer Law & Security Review*, *32*(6), 868–887. https://doi.org/10.1016/j.clsr.2016.07.009

Chang, K. H., Fisher, D., & Canny, J. (2011). Ammon: A speech analysis library for analyzing affect, stress, and mental health on mobile phones. *Proceedings of PhoneSense*, 2011.

Egorova, M. A., Minbaleev, A. V., Kozhevina, O. V., & Dufolt, A. (2021). Main directions of legal regulation of the use of artificial intelligence in the context of a pandemic. *Vestnik of Saint Petersburg University. Law*, *12*(2), 250–262. (In Russ.). https://doi.org/10.21638/spbu14.2021.201

Goldberg, Y. (2017). *Neural network methods for natural language processing*. Springer Nature. https://doi.org/10.1007/978-3-031-02165-7

Helberger, N., Borgesius, F. Z., & Reyna, A. (2017). The perfect match? A closer look at the relationship between EU consumer law and data protection law. *Common Market Law Review*, 54(5), 1427–1465. https://doi.org/10.54648/cola2017118

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Hobsbawn, E. (1996). Language, culture, and national identity. *Social Research*, *63*(4), 1065–1080.

Ilin, I. (2019). Legal Regime of the Language Resources in the Context of the European Language Technology Development. In Z. Vetulani, P. Paroubek, & M. Kubis (Eds.), *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2019. Lecture Notes in Computer Science* (vol 13212, pp. 367–376). Springer, Cham. https://doi.org/10.1007/978-3-031-05328-3_24

Ilin, I. (2020). The Voice and Speech Processing within Language Technology Applications: Perspective of the Russian Data Protection Law. *Legal Issues in the digital Age*, *1*(1), 99–123. https://doi.org/10.17323/2713-2749.2020.1.99.123

Ilin, I., & Kelli, A. (2020). The use of human voice and speech for development of language technologies: the EU and Russian data-protection law perspectives. *Juridica Int'l*, *29*, 71–85. https://doi.org/10.12697/ji.2020.29.07

Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, *14*(1), 4–20. https://doi.org/10.1109/tcsvt.2003.818349

Jents, L., & Kelli, A. (2014). Legal aspects of processing personal data in development and use of digital language resources: the Estonian perspective. *Jurisprudence*, *21*(1), 164–184. https://doi.org/10.13165/jur-14-21-1-08

Kelli, A., Lindén, K., Kamocki, P., Vider, K., Labropoulou, P., Birštonas, R., Mantrov., V., Hannesschläger, V., Del Gratta, R., Värv, A., Tavits, G., & Vutt, A. (2021). The interplay of legal regimes of personal data, intellectual property and freedom of expression in language research. In M. Monachini & M. Eskevich (Eds.), *Proceedings CLARIN Annual Conference 2021* (pp. 154–159). https://doi.org/10.3384/ecp1898

Kelli, A., Tavast, A., & Pisuke, H. (2012). Copyright and constitutional aspects of digital language resources: The Estonian approach. *Juridica International*, *19*, 40.

Kolain, M., Grafenauer, C., & Ebers, M. (2022). Anonymity Assessment – A Universal Tool for Measuring Anonymity of Data Sets under the GDPR with a Special Focus on Smart Robotics. *Rutgers University Computer & Technology Law Journal*, *48*(2). https://ssrn.com/abstract=3971139

Konev, S. I. (2020). Administrative responsibility for violations of the legislation of the Russian Federation in the field of personal data: new challenges. *Voprosy rossiiskogo i mezhdunarodnogo prava*, *10*(10-1), 274–282. (In Russ.).

Krivogin, M. (2017). Peculiarities of legal regulating biometric personal data. *Law. Journal of the Higher School of Economics*, *2*, 80–89. (In Russ.). https://doi.org/10.17323/2072-8166.2017.2.80.89

Lohsse, S., Schulze, R., & Staudenmayer, D. (Eds.) (2020). *Data as counter-performance-Contract Law 2.0?* Baden-Baden: Nomos Verlagsgesellschaft mbH & Company KG. https://doi.org/10.5771/9783748908531

Metzger, A. (2017). Data as counter-performance: what rights and duties do parties have. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, *8*(1), 2.

Oostveen, M. (2016). Identifiability and the applicability of data protection to big data. *International Data Privacy Law*, *6*(4), 299–309. https://doi.org/10.1093/idpl/ipw012

Purtova, N. (2018). The law of everything. Broad concept of personal data and future of EU data protection law. *Law, Innovation and Technology*, *10*(1), 40–81. https://doi.org/10.1080/17579961.2018.1452176

Sartor, G. (2020). *New aspects and challenges in consumer protection. Digital services and artificial intelligence*. European Parliament.

Savelyev, A. (2016). Russia's new personal data localization regulations: A step forward or a self-imposed sanction? *Computer Law & Security Review*, *32*(1), 128–145. https://doi.org/10.1016/j.clsr.2015.12.003

Savelyev, A. I. (2021). Civil law aspects of commercialization of personal data. *Civil Law Review*, *21*(4), 104–129. (In Russ.). https://doi.org/10.24031/1992-2043-2021-21-4-104-129

Svantesson, D. J. B. (2018). Enter the quagmire – the complicated relationship between data protection law and consumer protection law. *Computer Law & Security Review*, *34*(1), 25–36. https://doi.org/10.1016/j.clsr.2017.08.003

Sviridova, E. A. (2021). Open and personal data in artificial intelligence systems: legal aspects. *Economic Problems and Legal Practice*, *17*(6), 61–68. (In Russ.).

Talapina, E. V. (2020). The Law on information during an era of Big Data. *Vestnik of Saint Petersburg University. Law*, *11*(1), 4–18. (In Russ.). https://doi.org/10.21638/spbu14.2020.101

Tracy, J. M., Özkanca, Y., Atkins, D. C., & Ghomi, R. H. (2020). Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. *Journal of Biomedical Informatics*, *104*, 103362. https://doi.org/10.1016/j.jbi.2019.103362

Truyens, M., & Van Eecke, P. (2014). Legal aspects of text mining. *Computer Law & Security Review*, *30*(2), 153–170. https://doi.org/10.1016/j.clsr.2014.01.009

Vazhorova, M. A. (2012). The ratio of concepts "the information on private life" and "personal data". *Saratov State Law Academy Bulletin*, *4*(87), 55–59. (In Russ.).

Author information

Ilya G. Ilin – Master of Arts in Information Technology Law, postgraduate student, Faculty of Law, Saint Petersburg State University
Address: 22nd line of Vasilievsky Island, 7199106 Saint Petersburg, Russian Federation
E-mail: i.g.ilin@spbu.ru
ORCID ID: https://orcid.org/0000-0003-1076-2765
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57765898000
WoS Researcher ID: https://www.webofscience.com/wos/author/record/FDF-0979-2022
Google Scholar ID: https://scholar.google.com/citations?user=YruuMK0AAAAJ

Conflict of interest

The author declares no conflict of interest.

Financial disclosure

The research had no sponsorship.

Thematic rubrics

OECD: 5.05 / Law
PASJC: 3308 / Law
WoS: OM / Law

Article history

Date of receipt – December 25, 2023
Date of approval – January 5, 2024
Date of acceptance – Март 15, 2024
Date of online placement – Март 20, 2024

Научная статья
УДК 34:004:34.096:004.8
EDN: https://elibrary.ru/acxhto
DOI: https://doi.org/10.21202/jdtl.2024.7

Персональные данные в системах искусственного интеллекта: технология обработки естественного языка

Илья Геннадьевич Ильин

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Ключевые слова

биометрические данные,
законность,
законодательство,
искусственный интеллект,
лигалтех,
технология обработки естественного языка,
персональные данные,
право,
правовой риск,
цифровые технологии

Аннотация

Цель: концептуализировать с точки зрения законодательства в области защиты персональных данных процесс развития технологии обработки естественного языка, выявив возможные правовые барьеры для такого развития и направления дальнейшего исследования проблемы.

Методы: в основе исследования находятся общенаучные методы познания, наряду с которыми применялись формально-юридический, сравнительно-правовой методы, а также метод теоретического моделирования.

Результаты: установлено, что соблюдение режима персональных данных в процессе разработки технологии обработки естественного язык приводит к возникновению конфликта между частными и публично-правовыми интересами, что, в свою очередь, создает препятствия для дальнейшего развития обозначенной технологии. Показаны недостатки существующего правопорядка, который не в полной мере отвечает техническим особенностям развития технологии, что может привести к рискам излишнего регулирования, или же, напротив, к рискам оставления без внимания критических областей, требующих защиты. Обозначены проблемы при квалификации данных, задействованных в развитии рассматриваемой технологии. Предпринята попытка определить пределы обеспечения законности обработки персональных данных в составе технологии обработки естественного языка. Выделено в качестве пределов обеспечения законности материальное, временное и территориальное действие правового регулирования в данной области. Затрагивается проблема возможности использования персональных данных в качестве встречного представления, что является важным для развития технологии обработки естественного языка и для совершенствования отрасли информационно-коммуникационных технологий.

Научная новизна: данная работа дополняет научную дискуссию о правовом регулировании обработки персональных данных системами искусственного интеллекта аналитикой, выполненной в контексте технологии обработки естественного языка. Невысокая степень изученности последней обуславливает необходимость исследования области информационного права в части правоотношений по созданию систем искусственного интеллекта и оценки влияния режима персональных данных на развитие технологии обработки естественного языка.

Практическая значимость: затрагиваемые в статье прикладные аспекты исследуемой проблематики и полученные результаты могут быть использованы для совершенствования правового регулирования общественных отношений в области создания и развития искусственного интеллекта, а также для выявления и оценки правовых рисков, возникающих при обработке персональных данных разработчиками цифровых продуктов на базе технологии обработки естественного языка.

Для цитирования

Ильин, И. Г. (2024). Персональные данные в системах искусственного интеллекта: технология обработки естественного языка. *Journal of Digital Technologies and Law*, *2*(1), 123–140. https://doi.org/10.21202/jdtl.2024.7

Список литературы

Важорова, М. А. (2012). Соотношение понятий «Информация о частной жизни» и «Персональные данные». *Вестник Саратовской государственной юридической академии*, *4*(87), 55–59.

Егорова, М. А., Минбалеев, А. В., Кожевина, О. В., Ален, Д. (2021). Основные направления правового регулирования использования искусственного интеллекта в условиях пандемии. *Вестник Санкт-Петербургского университета. Право*, *12*(2), 250–262. https://doi.org/10.21638/spbu14.2021.201

Конев, С. И. (2020). Административная ответственность за нарушения законодательства Российской Федерации в области персональных данных: новые вызовы. *Вопросы российского и международного права*, *10*(10-1), 274–282. https://elibrary.ru/fasdwm

Кривогин, М. С. (2017). Особенности правового регулирования биометрических персональных данных. *Право. Журнал высшей школы экономики*, *2*, 80–89. EDN: https://elibrary.ru/zfcizt. DOI: https://doi.org/10.17323/2072-8166.2017.2.80.89

Савельев, А. И. (2021). Гражданско-правовые аспекты регулирования оборота персональных данных. *Вестник гражданского права*, *21*(4), 104–129. EDN: https://elibrary.ru/vgyyau. DOI: https://doi.org/10.24031/1992-2043-2021-21-4-104-129

Свиридова, Е. А. (2021). Открытые и персональные данные в системах искусственного интеллекта: правовые аспекты. *Проблемы экономики и юридической практики*, *17*(6), 61–68. https://elibrary.ru/xokzik

Талапина, Э. В. (2020). Закон об информации в эпоху больших данных. *Вестник Санкт-Петербургского университета. Право*, *11*(1), 4–18. EDN: https://elibrary.ru/grjtqp. DOI: https://doi.org/10.21638/spbu14.2020.101

Arkhipov, V., & Naumov, V. (2016). The legal definition of personal data in the regulatory environment of the Russian Federation: Between formal certainty and technological development. *Computer Law & Security Review*, 32(*6*), 868–887. https://doi.org/10.1016/j.clsr.2016.07.009

Chang, K. H., Fisher, D., & Canny, J. (2011). Ammon: A speech analysis library for analyzing affect, stress, and mental health on mobile phones. *Proceedings of PhoneSense*, 2011.

Goldberg, Y. (2017). *Neural network methods for natural language processing*. Springer Nature. https://doi.org/10.1007/978-3-031-02165-7

Helberger, N., Borgesius, F. Z., & Reyna, A. (2017). The perfect match? A closer look at the relationship between EU consumer law and data protection law. *Common Market Law Review*, *54*(5), 1427–1465. https://doi.org/10.54648/cola2017118

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Hobsbawn, E. (1996). Language, culture, and national identity. *Social Research*, *63*(4), 1065–1080.

Ilin, I. (2019). Legal Regime of the Language Resources in the Context of the European Language Technology Development. In Z. Vetulani, P. Paroubek, & M. Kubis (Eds.), *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2019. Lecture Notes in Computer Science* (vol. 13212, pp. 367–376). Springer, Cham. https://doi.org/10.1007/978-3-031-05328-3_24

Ilin, I. (2020). The Voice and Speech Processing within Language Technology Applications: Perspective of the Russian Data Protection Law. *Legal Issues in the digital Age*, *1*(1), 99–123. https://doi.org/10.17323/2713-2749.2020.1.99.123

Ilin, I., & Kelli, A. (2020). The use of human voice and speech for development of language technologies: the EU and Russian data-protection law perspectives. *Juridica Int'l*, *29*, 71–85. https://doi.org/10.12697/ji.2020.29.07

Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, *14*(1), 4–20. https://doi.org/10.1109/tcsvt.2003.818349

Jents, L., & Kelli, A. (2014). Legal aspects of processing personal data in development and use of digital language resources: the Estonian perspective. *Jurisprudence*, *21*(1), 164–184. https://doi.org/10.13165/jur-14-21-1-08

Kelli, A., Lindén, K., Kamocki, P., Vider, K., Labropoulou, P., Birštonas, R., Mantrov., V., Hannesschläger, V., Del Gratta, R., Värv, A., Tavits, G., & Vutt, A. (2021). The interplay of legal regimes of personal data, intellectual property and freedom of expression in language research. In M. Monachini & M. Eskevich (Eds.), *Proceedings CLARIN Annual Conference 2021* (pp. 154–159). https://doi.org/10.3384/ecp1898

Kelli, A., Tavast, A., & Pisuke, H. (2012). Copyright and constitutional aspects of digital language resources: The Estonian approach. *Juridica International*, *19*, 40.

Kolain, M., Grafenauer, C., & Ebers, M. (2022). Anonymity Assessment – A Universal Tool for Measuring Anonymity of Data Sets under the GDPR with a Special Focus on Smart Robotics. *Rutgers University Computer & Technology Law Journal*, *48*(2). https://ssrn.com/abstract=3971139

Lohsse, S., Schulze, R., & Staudenmayer, D. (Eds.). (2020). *Data as counter-performance-Contract Law 2.0?* Baden-Baden: Nomos Verlagsgesellschaft mbH & Company KG. https://doi.org/10.5771/9783748908531

Metzger, A. (2017). Data as counter-performance: what rights and duties do parties have. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, *8*(1), 2.

Oostveen, M. (2016). Identifiability and the applicability of data protection to big data. *International Data Privacy Law*, *6*(4), 299–309. https://doi.org/10.1093/idpl/ipw012

Purtova, N. (2018). The law of everything. Broad concept of personal data and future of EU data protection law. *Law, Innovation and Technology*, *10*(1), 40–81. https://doi.org/10.1080/17579961.2018.1452176

Sartor, G. (2020). *New aspects and challenges in consumer protection. Digital services and artificial intelligence*. European Parliament.

Savelyev, A. (2016). Russia's new personal data localization regulations: A step forward or a self-imposed sanction? *Computer Law & Security Review*, *32*(1), 128–145. https://doi.org/10.1016/j.clsr.2015.12.003

Svantesson, D. J. B. (2018). Enter the quagmire – the complicated relationship between data protection law and consumer protection law. *Computer Law & Security Review*, *34*(1), 25–36. https://doi.org/10.1016/j.clsr.2017.08.003

Tracy, J. M., Özkanca, Y., Atkins, D. C., & Ghomi, R. H. (2020). Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. *Journal of Biomedical Informatics*, *104*, 103362. https://doi.org/10.1016/j.jbi.2019.103362

Truyens, M., & Van Eecke, P. (2014). Legal aspects of text mining. *Computer Law & Security Review*, *30*(2), 153–170. https://doi.org/10.1016/j.clsr.2014.01.009

Сведения об авторе

Ильин Илья Геннадьевич – магистр права в области информационных технологий, аспирант юридического факультета, Санкт-Петербургский государственный университет
Адрес: 199106, Россия, г. Санкт-Петербург, 22-я линия В.О., 7
E-mail: i.g.ilin@spbu.ru
ORCID ID: https://orcid.org/0000-0003-1076-2765
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57765898000
WoS Researcher ID: https://www.webofscience.com/wos/author/record/FDF-0979-2022
Google Scholar ID: https://scholar.google.com/citations?user=YruuMK0AAAAJ

Конфликт интересов

Автор сообщает об отсутствии конфликта интересов.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.19.25 / Правовой режим информации, информационных систем и сетей
Специальность ВАК: 5.1.2 / Публично-правовые (государственно-правовые) науки

История статьи

Дата поступления – 25 декабря 2023 г.
Дата одобрения после рецензирования – 5 января 2024 г.
Дата принятия к опубликованию – 15 марта 2024 г.
Дата онлайн-размещения – 20 марта 2024 г.