Japan and Singapore) (January 1, 2019). (2019) 1 European Human Rights Law Review 23, UNSW Law Research Paper No. 19-2.

12. Yudiana, T. C., Rosadi, S. D., & Priowirjanto, E. S. The Urgency of Doxing on Social Media Regulation and the Implementation of Right to Be Forgotten on Related Content for the Optimization of Data Privacy Protection in Indonesia // Padjadjaran Jurnal Ilmu Hukum (Journal OF Law). 2022. Vol. 9(1). Pp 24–45.

И. Г. Ильин,
аспирант,
Санкт-Петербургский государственный университет

ПЕРСОНАЛЬНЫЕ ДАННЫЕ В СИСТЕМАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ТЕХНОЛОГИЯ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Аннотация. Статья посвящена концептуализации с точки зрения закона о защите персональных данных процесса развития технологии обработки естественного языка. В результате исследования было выявлено, что существующий правопорядок не в полной мере отвечает техническим особенностям развития данной технологии, что может привести или к излишнему регулированию, или же, напротив, оставить без внимания критические области, требующие защиты. В статье представлены основные проблемы и обозначены направления исследований.

Ключевые слова: право, цифровые технологии, персональные данные, искусственный интеллект, интеллектуальный анализ данных, технология обработки естественного языка, биометрические данные

PERSONAL DATA IN ARTIFICIAL INTELLIGENCE SYSTEMS: NATURAL LANGUAGE PROCESSING TECHNOLOGY

Abstract. The report focuses on the research results aimed at conceptualizing the development of natural language processing (NLP) from the perspective of data protection law. As a result of the research, it was identified that the existing legal regime does not fully meet the technical features of the development of NLP, which can lead to excessive regulation or, on the contrary, leave critical areas that require protection unattended. The following lecture notes aim to briefly describe the problems identified during the research and indicate the directions for further analysis.

Keywords: law, digital technologies, personal data, artificial intelligence, data mining, natural language processing technology, biometric data

Технология обработки естественного языка (англ. Natural language processing, NLP) активно используется в цифровых товарах и услугах (цифровых продуктах) для построения коммуникации между человеком и компьютером [2]. Голосовые помощники, сервисы перевода и озвучки текстов, системы интерактивного

ответа – все это примеры продуктов данной отрасли. В основе рассматриваемой технологии находятся генеративные нейросети, для обучения которых используются электронные лингвистические корпуса – базы данных, содержащие в себе множество текстов (книг, текстовых транскрипций, переводов и т. д.) и аудиофайлов (аудиокниг, записей трансляций, подкастов, другого аудиоконтента) [3, 8]. Создание лингвистических корпусов предполагает последовательное прохождение нескольких этапов: оцифровка языка – сбор, обработка и перевод данных в машиночитаемый формат, разметка корпуса и его последующий интеллектуальный анализ (англ. Text and data mining, TDM) [3–6].

В контексте создания лингвистических корпусов и развития технологии отдельную и важную роль приобретает вопрос использования данных, требующих особого режима правовой и технической защиты – персональных данных. В связи с этим принципиальными для разрешения становятся проблемы разграничения и категоризации персональных данных, пределов, до которых режим персональных данных будет влиять на процесс создания и развития названной технологии, а также частные случаи использования персональных данных, применительно к последующему ее распространению (прим. оплата цифровых услуг персональными данными).

Проблемы, связанные с разграничением и категоризацией персональных данных в общем смысле, объясняются необходимостью сначала выделить из объема всех используемых данных – персональные, а затем соотнести их с соответствующей категорией. Вместе с тем на практике это не всегда удается сделать: граница между персональными и другими данными не всегда четкая.

Во-первых, возникает проблема в определении самого понятия «данные». Из соотношения существующих норм можно сделать вывод о том, что данные – это данные, т. е. понятие определяется через само себя. Это создает трудности при попытках определить форму, в которой персональные данные могут быть выражены.

Во-вторых, действующее законодательство исходит из бинарного подхода к определению понятия персональных данных: данные могут быть либо персональными, либо нет. По мнению автора, такой подход не в полной мере учитывает современное состояние цифровизации общества, уровень технологического развития, а также последние социально-экономические изменения. Например, с точки зрения информатики и компьютерных наук выделяют разные уровни возможной идентифицируемости и относят к каждому из уровней определенный набор рисков [8, 9]. Кроме того, такое определение не учитывает, что данные могут быть идентифицируемыми для одного субъекта, например, в сочетании с другими наборами данных, но не для других [10].

В-третьих, статус данных в процессе обработки также может находиться в динамике и не быть статичным [11]. Иными словами, в процессе обработки данные могут становиться персональными и, наоборот, терять этот статус. Например, в процессе создания языковой модели на базе лингвистического корпуса, задействованные персональные данные теряют маркеры идентификации и, следовательно, теряют статус персональных [7].

Таким образом, в практическом смысле, данные в качестве персональных и можно квалифицировать только на определенный момент времени или этапе обработки, что может затруднить соблюдение законности всего процесса обработки.

Другая проблема, требующая решения – это проблема определения предела, до которого обработка данных должна соответствовать требованиям закона. Например, если языковая модель или корпус были созданы с использованием персональных данных, означает ли это, что дальнейшее использование продуктов, построенных на их базе, также попадает под действие закона о защите персональных данных?

Представляется, что пределы в обеспечении законности обработки персональных данных в рассматриваемом случае может быть определено через материальное, временное и территориальное действие правового регулирования в области защиты персональных данных [4]. Например, материальное действие можно определить через различные уровни использования персональных данных в создании соответствующих цифровых продуктов [7], временные пределы – через срок, в течение которого будет действовать право субъекта на защиту данных о нем, территориальные – через юрисдикции стран, в которых создаются или распространяются соответствующие цифровые продукты. Вместе с тем такой подход не может быть универсальным, а его применение влечет за собой ряд трудностей, таких как необходимость соблюдать регулирование в области защиты персональных данных, в том числе в отношении данных умерших людей и без какого-либо ограничения по сроками, необходимость одновременного соблюдения не только национального законодательства в области защиты персональных данных, но и законов других стран, так как цифровые продукты редко сосредоточены на одной стране, а реализуются на рынках разных стран и т. п.

Последней из обозначенных выше проблем является проблема использования персональных данных, применительно к процессу последующего распространения технология обработки естественного языка – оплате цифровых услуг персональными данными. Использование цифровых продуктов на базе описываемой технологии предполагает интенсивный обмен данными между пользователем и поставщиком [1]. Поставщик зачастую заинтересован в использовании этих данных не только для предоставления самого продукта, но и для его разработки, улучшения, а также в коммерческих целях. Например, голосовые данные могут быть использованы для анализа эмоциональной реакции на рекламный контент [12]. Однако возникает вопрос, насколько такое использование соответствует правовому режиму персональных данных? Представляется, что на сегодняшний день такое использование само по себе не запрещено, но должно осуществляться в строгом соответствии с действующим регулированием в области защиты персональных данных [13]. Вместе с тем остается открытым вопрос об определении данных как объекта права собственности [14], а также о характеристике возмездности соответствующих гражданско-правовых договоров.

Список литературы

1. Goldberg Y. Neural Network Methods for Natural Language Processing // Synthesis Lectures on Human Language Technologies. 2017. Vol. 10, № 1. Pp. 1–309.

2. Hirschberg J., Manning C. D. Advances in natural language processing // Science. 2015. Vol. 349, № 6245. Pp. 261–266.

3. Ilin I. Legal Regime of the Language Resources in the Context of the European Language Technology Development // Language and Technology Conference. Cham: Springer International Publishing, 2019. Pp. 367–376.

4. Ilin I. The Voice and Speech Processing within Language Technology Applications: Perspective of the Russian Data Protection Law // Legal Issues in the digital Age. 2020. № 1. Pp. 99–123.

5. Ilin I., Kelli A. The use of human voice and speech for development of language technologies: the EU and Russian data-protection law perspectives // Juridica Int'l. 2020. Vol. 29. Pp. 71–105.

6. Jents L., Kelli A. Legal aspects of processing personal data in development and use of digital language resources: the Estonian perspective // Jurisprudencija. – 2014. Vol. 21, № 1. Pp. C. 164–184.

7. Kelli A. et al. The interplay of legal regimes of personal data, intellectual property and freedom of expression in language research // Proceedings CLARIN annual conference. 2021. Vol. 2021. Pp. 154–159.

8. Kelli A., Tavast A., Pisuke H. Copyright and constitutional aspects of digital language resources // Juridica Int'l. 2012. Vol. 19. Pp. 40–64.

9. Kolain M., Grafenauer C., Ebers M. Anonymity Assessment-A Universal Tool for Measuring Anonymity of Data Sets under the GDPR with a Special Focus on Smart Robotics // Rutgers Computer & Tech. LJ. 2021. Vol. 48. Pp. 174–188.

10. Oostveen M. Identifiability and the applicability of data protection to big data // International Data Privacy Law. 2016. Vol. 6, № 4. Pp. 299–309.

11. Purtova N. The law of everything. Broad concept of personal data and future of EU data protection law // Law, Innovation and Technology. 2018. Vol. 10, № 1. Pp. 40–81.

12. Sartor G. et al. Study: New aspects and challenges in consumer protection. Digital services and artificial intelligence. European Parliament, 2020. Pp. 1–41.

13. Савельев А. И. Гражданско-правовые аспекты регулирования оборота персональных данных // Вестник гражданского права. 2021. Т. 21, № 4. Pp. 104–129.

14. Талапина Э. В. Закон об информации в эпоху больших данных // Вестник Санкт-Петербургского университета. Право. 2020. Т. 11, № 1. С. 4–18.