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

Содержание

Введение

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 (Hirschberg & Manning, 2015; Truyens & Van Eecke, 2014). Голосовые помощники, сервисы проверки орфографии, перевода и озвучки текстов, голосовая биометрия, системы интерактивного ответа – все это примеры продуктов с использованием данной технологии.

Использование языка и речи в информационных технологиях открывает новые перспективы в развитии искусственного интеллекта, создает возможности для разработки инновационных цифровых продуктов, способствующих цифровой

трансформации общества, например, ChatGPT и его аналогов, определяет развитие LegalTech индустрии. Высокий экономический потенциал и социальная значимость технологии, обусловленная значением языка и его местом в процессе национальной и культурной идентичности (Hobsbawn, 1996), определяют заинтересованность общества, бизнеса и государства в ее развитии.

В свою очередь, несмотря на инновационный характер и общественно-экономическую пользу технологии, анализ существующего правопорядка указывает на наличие юридических препятствий для ее развития.

С технической точки зрения одной из главных задач для развития технологии обработки естественного языка является создание и последующее распространение лингвистических корпусов (электронных речевых ресурсов). В широком смысле под лингвистическими корпусами можно понимать базы данных, содержащие в себе множество текстов (книг, текстовых транскрипций, переводов и т. д.) и аудиофайлов (аудиокниг, записей трансляций, подкастов, другого аудиоконтента), которые впоследствии используются в алгоритмах машинного обучения (Kelli et al., 2012; Ilin, 2019). Создание лингвистических корпусов предполагает последовательное прохождение двух этапов: оцифровки языка – сбора и перевода данных в машиночитаемый формат и их последующего интеллектуального анализа (англ. Text and data mining, TDM) (Jents & Kelli, 2014; Ilin I., 2019). Без наличия массивных лингвистических корпусов, а также без возможности доступа к ним заинтересованных лиц технология не будет развиваться и работать.

Лингвистический корпус сам по себе и его отдельные элементы с юридической точки зрения могут включать в себя информацию, относящуюся к персональным или другим данным, обладающим специальным правовым режимом, например, данным, охраняемым тайной связью. Соответственно, на отношения по поводу создания и распространения лингвистических корпусов будут одновременно влиять законодательство в области защиты персональных данных, а также специальное отраслевое законодательство.

Цель настоящей статьи – концептуализировать с точки зрения законодательства в области защиты персональных данных процесс развития технологии обработки естественного языка, определить возможные правовые преграды для такого развития, а также предложить направления для дальнейшего исследования проблемы. Результат работы дополняет научную дискуссию о правовом регулировании обработки персональных данных системами искусственного интеллекта (Свиридова, 2021; Конев, 2020; Егорова и др., 2021) анализом проблемы в контексте работы обозначенной технологии, а также он может использоваться на практике для выявления и оценки правовых рисков, возникающих при обработке персональных данных разработчиками цифровых продуктов на базе технологии обработки естественного языка.

1. Персональные данные в составе лингвистического корпуса

1.1. Определение и правовой режим

Как было ранее обозначено, ключевым процессом для развития технологии обработки естественного языка будут создание и распространение среди заинтересованных лиц лингвистических корпусов. Элементы лингвистического корпуса могут быть квалифицированы в качестве персональных данных. Современный подход

к определению понятия «персональные данные», основанный на п. 1 ст. 3 ФЗ «О персональных данных»[1], позволяет широко толковать данный термин и квалифицировать в качестве персональных данных практически любые данные, которые прямо или косвенно позволяют определить физическое лицо. Это приводит к неизбежности появления персональных данных в составе лингвистического корпуса.

Персональные данные требуют особого режима правовой и технической защиты. В общем смысле особый правовой режим персональных данных, с одной стороны, направлен на обеспечение защиты прав, принадлежащих субъекту данных, с другой – накладывает ряд ограничений на использование таких данных в создании лингвистического корпуса (Ilin, 2020). В связи с этим для создания лингвистического корпуса принципиальной будет задача отличить используемые персональные данные от других видов данных. С практической точки зрения это не всегда удается сделать: граница между персональными и другими данными далеко не всегда четкая.

Во-первых, возникает проблема в определении самого понятия «данные». Пункт 1 ст. 3 ФЗ «О персональных данных»[2] определяет данные как «любую информацию…». В свою очередь ст. 2 ФЗ «Об информации, информационных технологиях и о защите информации»[3] определяет информацию как «сведения (сообщения, данные)…». Иными словами, из соотношения названных определений можно сделать вывод о том, что данные – это данные, т. е. понятие определяется через само себя. Такое положение дел может создавать трудности при попытках определить форму, в которой персональные данные могут быть выражены, и, следовательно, квалифицировать ту или иную информацию в качестве персональных данных.

Во-вторых, одним из наиболее проблемных аспектов в квалификации данных в качестве персональных является то, что текущее законодательство опирается на бинарный подход в определении понятия персональных данных. Согласно этому подходу, данные могут быть либо персональными, либо нет. По мнению автора, бинарный подход к определению персональных данных в недостаточной степени учитывает современное состояние цифровизации общества, уровень технологического развития, а также социально-экономические изменения, произошедшие за последние несколько лет. Например, такое определение не учитывает, что с точки зрения информатики и компьютерных наук выделяют разные уровни возможной идентифицируемости и относят к каждому из уровней определенный набор рисков (Kolain et al., 2022). Кроме того, не принимается во внимание и тот факт, что данные могут быть идентифицируемыми для одного субъекта, например, в сочетании с другими наборами данных, но не для других (Oostveen, 2016).

В-третьих, статус данных в процессе обработки также может находиться в динамике и не быть статичным (Purtova, 2018). Иными словами, в процессе обработки данные могут становиться персональными и, наоборот, терять этот статус. Например, в процессе создания лингвистического корпуса и языковой модели на его базе используемые персональные данные теряют маркеры идентификации и, таким

---

1 О персональных данных. № 152-ФЗ от 27.07.2006. СПС «КонсультантПлюс». https://clck.ru/gLnFq

2 Там же.

3 Об информации, информационных технологиях и о защите информации. № 149-ФЗ от 27.07.2006. СПС «КонсультантПлюс». https://clck.ru/3967pC

образом, теряют статус персональных (Kelli et al., 2021). Это означает, что в практическом смысле данные в качестве персональных можно квалифицировать только на определенный момент времени или на этапе обработки, что может затруднить соблюдение законности всего процесса обработки.

Динамический характер данных влияет и имеет принципиальное значение, во-первых, для разграничения персональных данных от других видов данных (например, для обработки данных не относящихся к персональным, предъявляется меньше требований). Во-вторых, он может влиять на выбор категории обрабатываемых персональных данных.

Законодательство в области защиты персональных данных выделяет четыре основные категории персональных данных: «общие», «биометрические», «специальные» или «чувствительные», а также «персональные данные, разрешенные для распространения». Биометрические данные относятся к биологическим и физиологическим характеристикам человека, которые можно использовать для идентификации[4]. Специальная или чувствительная категория данных – это данные, которые указывают на «политические взгляды, расовое или этническое происхождение, философские или религиозные убеждения, состояние здоровья или сексуальную ориентацию»[5]. «Персональные данные, разрешенные для распространения», – это данные, которые с согласия субъекта могут быть распространены и стать доступны неограниченному кругу лиц[6]. Указанное разделение персональных данных на категории является исходной предпосылкой обработки данных. Например, биометрические данные могут быть обработаны только после получения явного согласия субъекта данных. Обработка специальной категории персональных данных, по общему правилу, запрещена. Для разных категорий данных установлены различные требования к обеспечению уровня как технической, так и правовой защиты (Кривогин, 2017). Таким образом, проблема квалификации данных в качестве персональных ставит задачу не только разграничить персональные данные от других видов данных, но и определить принадлежность персональных данных к определенной категории.

1.2. Вопросы квалификации голоса в качестве персональных данных в разных юрисдикциях

В контексте создания лингвистического корпуса использование речевых данных – образцов голоса человека – хорошо иллюстрирует сложность вышеописанной задачи. Технология обработки естественного языка в значительной степени полагается на обработку названных данных. Например, они используются, чтобы обеспечить работу голосовых помощников, систем расшифровки текстов, перевода разговоров в режиме реального времени и т. д. Возможность отнесения голоса к персональным данным и выбор соответствующей категории на практике могут вызвать ряд затруднений.

---

4 О персональных данных. № 152-ФЗ от 27.07.2006. Ст. 11. СПС «КонсультантПлюс». https://clck.ru/gLnFq

5 Там же. Ст. 10.

6 Там же. Ст. 10.1.

Предложенный широкий подход к определению понятия персональных данных позволяет предположить, что голос содержит в себе достаточно идентификаторов личности, что относит его к персональных данным. Такой подход также можно встретить в практике Европейского суда по правам человека (далее – ЕСПЧ)[7], юрисдикция которого до недавнего времени распространялась на Россию. Вместе с тем действующее законодательство РФ так же, как и существующая судебная практика, не дает однозначного ответа, всегда ли голос будет относиться к категории персональных данных. На практике можно встретить одиночные случаи, когда записанный голос не будет квалифицирован в качестве персональных данных (Arkhipov & Naumov, 2016). Действительно, можно предположить, что голос не всегда будет относиться к персональным данным, например, ввиду искаженности, краткости образца, отсутствия связанных идентификаторов.

Вместе с тем представляется, что вероятность квалификации голоса в качестве персональных данных достаточно велика, что поднимает вопрос об его отнесении к соответствующей категории. Особый интерес при этом вызывает деление между общей категорией и категориями, предусматривающими повышенный уровень охраны. С точки зрения последних, голос может рассматриваться как биометрические данные или данные о состоянии здоровья. Например, голос можно использовать в качестве биомаркера для выявления ранней стадии болезни Паркинсона (Tracy et al., 2020) или для предоставления информации о психическом состоянии человека, уровне стресса, эмоциональном состоянии, недостатке сна и других данных, связанных со здоровьем (Chang et al., 2011). Голос как уникальная характеристика человека (биометрические данные) часто используется для установления личности человека (голосовая биометрия) (Jain et al., 2004). Голос также может содержать в себе персональные данные, относящиеся к общей категории. Например, такие данные могут появиться в речевом содержании или связанных с образцом метаданных (возраст, пол и т. д.).

Соответствующие разъяснения Роскомнадзора[8], действие которых в настоящее время отменено[9], содержали рекомендацию, согласно которой отличить биометрические от других категорий данных можно было исходя из целей обработки. Если цель обработки была связана с идентификацией личности, то такие данные необходимо было квалифицировать и обрабатывать в качестве биометрических данных, в других случаях такая необходимость отсутствовала. Предложенный подход представляется логичным, в противном случае, например, трансляция любой теле- или радиопередачи означала бы обработку биометрических данных и требовала бы соблюдения соответствующего правового режима. Разумно предположить, что такой же подход может быть применен к данным о состоянии здоровья. Очевидно, что в большинстве случаев для создания лингвистического корпуса и работы самой технологии обработки естественного языка извлечение данных о здоровье не требуется

---

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 Роскомнадзор. (2013, 2 сентября). Разъяснения по вопросам отнесения фото- и видеоизображений, дактилоскопических данных и иной информации к биометрическим персональным данным и особенностей их обработки. https://clck.ru/3968ky

9 Письмо Роскомнадзора от 19.11.2021 № 09–78548 «О неактуальности разъяснений Роскомнадзора». СПС «КонсультантПлюс». https://clck.ru/3968n6

и, следовательно, пока такие данные не извлекаются, обработка голоса не будет требовать соблюдения специального правового режима.

Вместе с тем нельзя не отметить, что независимо от того, с какой целью производится обработка голоса, в нем все равно будут содержаться биометрические данные и данные о здоровье, а следовательно, всегда будет риск, что эти данные могут быть извлечены из голоса. Это поднимает общий и более концептуальный вопрос: достаточно ли наличия самого по себе такого риска, чтобы отнести персональные данные к категориям, предусматривающим повышенный уровень охраны, например, «специальной» или «биометрической» категории? Насколько автору известно, в России по данному вопросу судебная практика отсутствует. Если обратиться к другим юрисдикциям, можно найти схожее дело в практике Суда Европейского союза[10]. В рассматриваемом деле суд использовал широкий подход к толкованию понятия «специальная» категория данных и включил в ее состав данные, которые прямо к ней не относились, однако могли быть потенциально извлечены. По мнению автора, такой подход как минимум является дискуссионным. Во-первых, не совсем понятно, как определить, что используемые данные содержат в себе скрытые данные, которые в случае извлечения могут изменить их квалификацию. Наличие таких данных не всегда очевидно. Во-вторых, отнесение данных к «специальной» категории означает невозможность их обработки, за исключением конкретных случаев, перечисленных в ст. 10 закона о персональных данных или ст. 9 Общего регламента о защите персональных данных[11]. Обеспечить соблюдение обработки данных требованиям закона в такой ситуации довольно сложно, что, в свою очередь, может привести к невозможности использовать данные в необходимом объеме и, следовательно, невозможности достичь тех целей, для которых эти данные должны были быть обработаны.

Кроме того, необходимо также отметить, что уровень риска потенциального извлечения дополнительной информации, способной изменить квалификацию данных, будет разным в зависимости от конкретного случая обработки. Например, речевые данные в составе лингвистического корпуса не являются общедоступными, круг лиц, которые могут получить доступ к этим данным и использовать их для извлечения дополнительной информации, ограничен. Причем для этого необходимы специальные знания и технические средства. Другими словами, обычная обработка данных в составе лингвистического корпуса имеет невысокий риск извлечения «специальных» или «биометрических» данных из голоса. Таким образом, представляется, что, с одной стороны, широкий подход к определению категорий данных повышает уровень защиты прав, принадлежащих субъекту данных, с другой – в значительной степени ограничивает возможности их использования для развития технологии. В данной ситуации для соблюдения баланса между интересами субъекта данных и лицами, осуществляющими обработку, важно не только выявить риск потенциального извлечения информации, но и провести техническую и правовую оценку данного риска.

---

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

11 Статья 10 Федерального закона «О персональных данных», ст. 9. 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) (2016April 27). https://clck.ru/3969GL

2. Пределы обеспечения законности обработки персональных данных в составе технологии обработки естественного языка

Использование персональных данных для создания лингвистического корпуса и развития технологии обработки естественного языка создает необходимость обеспечить обработку в строгом соответствии с законом. Российское законодательство широко толкует понятие обработки персональных данных[12], в связи с чем не возникает особых сомнений, что действия, совершаемые над данными в процессе их использования в составе технологии обработки естественного языка, будут квалифицироваться в качестве обработки персональных данных (Ilin & Kelli, 2020). В свою очередь возникает вопрос предела, до которого обработка данных должна соответствовать требованиям законодательства. Например, если языковая модель была создана с использованием персональных данных, означает ли это, что дальнейшее использование продуктов, созданных с использованием данной модели, также попадает под действие Закона о защите персональных данных?

Пределы в обеспечении законности обработки персональных данных в составе технологии обработки естественного языка могут быть определены через материальное, временное и территориальное действие правового регулирования в данной области (Ilin, 2020).

В контексте технологии обработки естественного языка материальное действие можно определить через различные уровни использования персональных данных в разработке соответствующих цифровых продуктов. К таким уровням будут относиться сбор материала для создания массива данных, составление массива данных, аннотация, создание языковой модели, разработка конечного цифрового продукта (Kelli et al., 2021). Например, языковая модель состоит из лингвистических правил языка, и извлечение каких-либо персональных данных из нее на сегодняшний день представляет собой практически невыполнимую задачу, при этом в процессе создания языковой модели идентифицируемость субъектов данных теряется. Поэтому режим данных, использованных для создания модели, в общем случае не будет соответствовать правовому режиму языковой модели, построенной на этих данных (Kelli et al., 2021).

Временные пределы обеспечения законности обработки данных можно определить через срок, в течение которого будет действовать право субъекта на защиту данных о нем. В то же время российское законодательство не устанавливает срок действия такого права, указывая на необходимость соблюдать регулирование в области персональных данных, в том числе в отношении данных умерших людей[13] и без какого-либо указания на сроки. Можно согласиться с мнением некоторых исследователей, что данный пробел необходимо устранить, для чего было бы целесообразно установить срок защиты персональных данных равным сроку охраны частной жизни лица (Важорова, 2012).

Пределы обеспечения законности обработки данных с точки зрения территориального действия будут определяться с учетом национальных юрисдикций стран, в которых создаются или распространяются соответствующие цифровые продукты.

---

[12] О персональных данных. № 152-ФЗ от 27.07.2006. П. 3 ст. 3. СПС «КонсультантПлюс». https://clck.ru/gLnFq

[13] Там же. П. 7 ст. 9.

Проблема в том, что цифровые продукты, построенные с использованием технологии обработки естественного языка, как правило, не сосредоточены на одной стране, а стремятся охватить рынки разных стран. Например, голосовой помощник «Алиса», разработанный компанией «Яндекс», поддерживает турецкий язык, система преобразования речи в текст, разработанная компанией Google, поддерживает более 120 языков и может быть интегрирована с другими цифровыми продуктами. Поэтому возникает вопрос о необходимости соблюдения не только национального законодательства в области защиты персональных данных, но и соответствующего регулирования других стран, скажем, турецкого законодательства – для сбора данных и формирования лингвистического корпуса на турецком языке.

Российское законодательство в области защиты персональных данных по общему правилу не имеет экстерриториального действия. Следовательно, оно не распространяется на нерезидентов, осуществляющих обработку персональных данных граждан России за рубежом. Это правило имеет два исключения. Первое касается требования локализации данных, а второе относится к случаям обеспечения государственной безопасности.

Правило локализации обработки персональных данных граждан Российской Федерации обязывает лиц, осуществляющих обработку данных, хранить, собирать и использовать персональные данные граждан России только в базах данных, расположенных на территории РФ. Данное правило применяется: если в составе обрабатываемой информации содержатся персональные данные, если эти данные были собраны (т. е. получены от третьих лиц) и обработаны, а также если эти данные связаны с гражданами РФ (Savelyev, 2016). Последнее условие поднимает проблему определения гражданства в рамках использования информационных технологий.

Данная проблема была частично решена после того, как Роскомнадзор предоставил разъяснения, согласно которым под гражданством следует понимать территорию, на которой происходит обработка, при наличии сомнений в отношении гражданства субъекта данных вся информация, обрабатываемая и собираемая на территории России, должна быть локализована в базах данных, расположенных в России. Вместе с тем до сих пор остается открытым вопрос идентификации и обработки персональных данных российских граждан, которые собираются за пределами российской юрисдикции (Ilin, 2020).

Примером случаев, когда меры по обеспечению государственной безопасности оказывают экстерриториальный эффект на законодательство в области защиты персональных данных, может служить требование для организаторов распространения информации и поставщиков телекоммуникационных услуг хранить интернет-трафик (голосовые и текстовые сообщения, фото, видео, звуки, метаданные файлов), а также предоставлять ключи шифрования для расшифровки интернет-трафика в случае, если требуемые данные хранятся или обрабатываются в зашифрованном виде[14].

[14] О внесении изменений в Федеральный закон «О противодействии терроризму» и отдельные законодательные акты Российской Федерации в части установления дополнительных мер по противодействию терроризму и обеспечению общественной безопасности. № 374-ФЗ от 06.07.2016. (2016). СПС «КонсультантПлюс». https://clck.ru/3969WT; О внесении изменений в Уголовный кодекс Российской Федерации и Уголовно-процессуальный кодекс Российской Федерации в части установления дополнительных мер по противодействию терроризму и обеспечению общественной безопасности. № 375-ФЗ от 06.07.2016. (2016). СПС «КонсультантПлюс».https://clck.ru/3969aF

Важно отметить, что такое требование не ограничивается какой-либо территорией. При этом надо отметить, что соблюдение этих правил может быть затруднено для компаний, поскольку они будут вынуждены нарушать свое национальное регулирование в области защиты данных. Например, некоторые положения Общего Регламента о защите персональных данных[15], регулирующего правоотношения в области защиты персональных данных в Европейском союзе, будут противоречить вышеуказанному требованию (Ilin, 2020).

3. Использование персональных данных для оплаты цифровых продуктов на базе технологии обработки естественного языка и вопросы юридической квалификации

В условиях современной цифровой экономики продукты, использующие технологию обработки естественного языка, широко реализуются с использованием бизнес-модели, которая не предполагает получения от пользователя денежного вознаграждения за продукт. Вместо этого поставщик извлекает выгоду от использования данных, которые либо были намеренно предоставлены пользователем, либо автоматически собраны или созданы поставщиком. Применение данной бизнес-модели и оплата товаров и услуг данными открыли дискуссию о том, насколько данные сами по себе могут являться средством платежа. В первую очередь сомнения возникают касательно экономических и юридических характеристик данных. В то время как одной из основных проблем экономических характеристик данных является проблема измерения их экономического потенциала и ценности, которые не могут быть определены ни рынком, ни договором (Lohsse et al., 2020), обсуждение в правовой плоскости возникает вокруг определения правового режима персональных данных, взаимоотношений, регулирующих правовые нормы, и содержания правоотношений в данной области (Helberger et al., 2017; Metzger, 2017; Svantesson, 2018).

Использование продуктов с применением технологии обработки естественного языка предполагает интенсивный обмен данными между пользователем и поставщиком (Goldberg, 2017). Данные могут быть намеренно переданы пользователем поставщику (например, путем голосовой команды), собраны поставщиком самостоятельно (например, запись образца голоса, данных о местоположении) или же созданы самостоятельно сервисом (например, результат текстового перевода). В большинстве случаев поставщик использует полученные данные не только для предоставления пользователю цифрового продукта, но и для его дальнейшей разработки и улучшения, а также в коммерческих целях, напрямую не связанных с реализуемым продуктом. Например, поставщик может использовать голос пользователя для анализа эмоциональной реакции на предложенный рекламный контент и использовать результат анализа для продажи других товаров и услуг (Sartor, 2020). Возможность такого использования данных ставит вопрос об их квалификации, применимом правовом режиме, а также возможности определения данных как объекта права собственности (Талапина, 2020). Кроме того, использование в качестве встречного представления

[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) (2016April 27). https://clck.ru/3969dk

персональных данных ставит вопрос самой правовой концепции персональных данных как одного из основополагающих прав человека – права на уважение частной и семейной жизни. Представляется, что действующее законодательство в области защиты персональных данных само по себе не запрещает подобное использование данных, однако накладывает ряд ограничений, в общем смысле связанных с необходимостью обеспечения законности подобной обработки (Савельев, 2021).

Вместе с тем использование данных в качестве встречного предоставления поднимает другой вопрос, касающийся характеристики возмездности гражданско-правовых договоров, используемых для практического применения рассматриваемой бизнес-модели и, следовательно, возможности распространения Закона о защите прав потребителя на такие отношения. С одной стороны, п. 1 ст. 423 ГК РФ[16] прямо предусматривает в качестве одной из характеристик возмездности совершение «иного встречного предоставления» за исполнение своих обязанностей по договору, а предоставление пользователю дополнительных гарантий и правовых механизмов защиты своих прав, предусмотренные законодательством о защите прав потребителей, сбалансирует его отношения с поставщиком, по отношению к которому пользователь занимает более слабую позицию.

С другой стороны, использование данных в качестве нового средства платежа может вызвать проблемы в правоприменительной практике. Например, для применения мер защиты прав потребителя, непосредственно зависящих от стоимости самого товара или услуги, возникает необходимость определить денежный эквивалент переданных данных, что с учетом экономической и правовой природы данных представляется труднореализуемым.

Заключение

Целью настоящей статьи было концептуализировать с точки зрения законодательства в области защиты персональных данных процесс развития технологии обработки естественного языка. Анализ существующего правопорядка показал, что он не в полной мере отвечает техническим особенностям развития технологии, что может привести или к излишнему регулированию, или же, напротив, оставить без внимания критические области, требующие защиты. В частности, проблемы возникают при квалификации данных, задействованных в развитии рассматриваемой технологии, в качестве персональных и их последующее отнесение к определенной категории. Данная проблема хорошо иллюстрируется на примере использования речевых данных – образцов голоса человека. Попытка определить пределы обеспечения законности обработки данных также вызывает трудности и требует дальнейшего уточнения как в части материального, так и временного и территориального действия правового регулирования в данной области. Еще одним вопросом, нуждающимся в дальнейшем исследовании, является возможность использования персональных данных в качестве встречного представления. Данная проблема представляется актуальной не только для развития технологии обработки естественного языка, но и в целом для развития отрасли информационно-коммуникационных технологий.

[16] Гражданский кодекс Российской Федерации (часть первая) от 30.11.1994 № 51-ФЗ, ред. от 24.07.2023. СПС «КонсультантПлюс». https://clck.ru/3969nv

Представляется, что задачей для дальнейшего исследования будут анализ и разработка решений, с одной стороны, способствующих преодолению обнаруженных правовых преград для развития технологии, с другой –повышения правовой защиты критических областей, требующих дополнительной защиты.

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.

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

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