Научная статья
УДК 346.1:006.44:004.8
EDN: https://elibrary.ru/dxnwhv
DOI: https://doi.org/10.21202/jdtl.2023.14

Правовые средства обеспечения принципа прозрачности искусственного интеллекта

Юлия Сергеевна Харитонова

Московский государственный университет имени М. В. Ломоносова
г. Москва, Российская Федерация

Ключевые слова

Автоматизированная обработка данных,
автономность,
алгоритм,
искусственный интеллект,
право,
принятие решений,
прозрачность,
цифровая экономика,
цифровые технологии,
этика

Аннотация

Цель: анализ действующих технологических и юридических теорий для определения содержания принципа прозрачности работы искусственного интеллекта с позиции правового регулирования, выбора применимых средств правового регулирования и установление объективных границ юридического вмешательства в технологическую сферу с помощью регулирующего воздействия.

Методы: методологическую основу исследования составляет совокупность общенаучных (анализ, синтез, индукция, дедукция) и специально-юридических (историко-правовой, формально-юридический, сравнительно-правовой) методов научного познания.

Результаты: подвергнуты критическому анализу нормы и предложения для нормативного оформления принципа прозрачности искусственного интеллекта с точки зрения невозможности получения полной технологической прозрачности искусственного интеллекта. Выдвинуто предложение обсудить варианты политики управления алгоритмической прозрачностью и подотчетностью на основе анализа социальных, технических и регулятивных проблем, создаваемых алгоритмическими системами искусственного интеллекта. Обосновано, что прозрачность является необходимым условием для признания искусственного интеллекта заслуживающим доверия. Обосновано, что прозрачность и объяснимость технологии искусственного интеллекта важна не только для защиты персональных данных, но и в иных ситуациях автоматизированной обработки данных, когда для принятия решений недостающие из входящей информации технологические данные восполняются из открытых источников, в том числе не имеющих значения хранилищ персональных данных. Предложено законодательно закрепить обязательный аудит и ввести стандарт, закрепляющий

компромисс между возможностями и преимуществами технологии, точностью и объяснимостью результата ее работы и правами участников общественных отношений. Введение сертификации моделей искусственного интеллекта, обязательных к применению, позволит решить вопросы ответственности обязанных применять такие системы субъектов. В контексте вопроса о профессиональной ответственности профессиональных субъектов, таких как врачи, военные, органы корпоративного управления юридического лица, требуется ограничить обязательное применение искусственного интеллекта в случаях, если не обеспечена его достаточная прозрачность.

Научная новизна: междисциплинарный характер исследования позволил выявить невозможность и необоснованность требований полного открытия исходного кода или архитектуры моделей искусственного интеллекта. Принцип прозрачности искусственного интеллекта может быть обеспечен за счет проработки и обеспечения права субъекта данных и субъекта, которому адресовано решение, принятое в результате автоматизированной обработки данных, на отказ от применения автоматизированной обработки данных для принятия решений и права на возражения против принятых таким способом решений.

Практическая значимость: обусловлена отсутствием в настоящее время достаточного регулирования принципа прозрачности искусственного интеллекта и результатов его работы, а также содержания и особенностей реализации права на объяснение и права на возражение субъекта решения. Наиболее плодотворный путь для установления доверия к искусственному интеллекту заключается в том, чтобы признать данную технологию частью сложной социотехнической системы, которая опосредует доверие, и повышать надежность этих систем. Основные положения и выводы исследования могут быть использованы для совершенствования правового механизма обеспечения прозрачности моделей искусственного интеллекта, применяемых в государственном управлении и бизнесе.

Для цитирования

Харитонова, Ю. С. (2023). Правовые средства обеспечения принципа прозрачности искусственного интеллекта. *Journal of Digital Technologies and Law*, *1*(2), 337–358. https://doi.org/10.21202/jdtl.2023.14

Содержание

Введение

В российском праве сформулированы принципы Национальной стратегии развития искусственного интеллекта на период до 2030 г., которые включают в том числе прозрачность как объяснимость работы искусственного интеллекта и процесса достижения им результатов, недискриминационный доступ пользователей продуктов, которые созданы с использованием технологий искусственного интеллекта, к информации о применяемых в этих продуктах алгоритмах работы искусственного интеллекта (п. 19 Стратегии, утв. Указом Президента РФ «О развитии искусственного интеллекта в Российской Федерации» № 490 от 10.10.2019). Понятия «объяснимость» и «недискриминационность» работы искусственного интеллекта выделены как составляющие принципа прозрачности.

Раскрытие информации также предусматривается в международных актах и национальном законодательстве разных стран. Эти правила в первую очередь наиболее подробно затрагивают вопросы защиты прав и свобод человека, как, например, в Общем регламенте по защите данных Европейского союза (General Data Protection Regulation, GDPR; Постановление (Европейский Союз) 2016/679[1]).

В России защита прав и свобод человека и безопасность работы искусственного интеллекта обозначены в Стратегии 2030 как самостоятельные принципы, хотя названные принципы тесно соприкасаются с прозрачностью. Как представляется, недискриминационность ведет к обеспечению защиты гарантированных российским и международным законодательством прав и свобод человека. Принцип безопасности работы искусственного интеллекта определен как недопустимость использования искусственного интеллекта в целях умышленного причинения вреда гражданам и юридическим лицам, а также как предупреждение и минимизация рисков возникновения негативных последствий использования технологий искусственного интеллекта. Представляется, что принцип прозрачности также позволяет добиться безопасности применения искусственного интеллекта.

В отсутствие ясного видения с точки зрения права содержания принципа прозрачности работы искусственного интеллекта нам представляется важным определить понятие прозрачности, а также исследовать допустимые границы юридического вмешательства в технологическую сферу с помощью регулирующего воздействия.

Прозрачность в контексте работы искусственного интеллекта может быть рассмотрена с позиции технологии, этики и права. Междисциплинарный подход позволяет критически взглянуть на нормы и предложения для нормативного оформления с точки зрения невозможности получения полной технологической прозрачности искусственного интеллекта. Требуется обсудить варианты политики управления алгоритмической прозрачностью и подотчетностью на основе анализа социальных, технических и регулятивных проблем, создаваемых алгоритмическими системами искусственного интеллекта.

---

[1] *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)*. http://eur-lex.europa.eu/

1. Понятие «черный ящик» и его значение для юридического оформления применения технологии искусственного интеллекта для принятия решений

В течение десятилетий проекты искусственного интеллекта опирались на человеческий опыт, накопленный инженерами по знаниям, и были как явно разработанными, так и легко понимаемыми людьми. Значительный прогресс в области искусственного интеллекта был достигнут с использованием контролируемых систем обучения, которые предназначены повторять решения людей (Hastie et al., 2009; LeCun et al., 2015; Pak & Kim, 2017). Например, экспертные системы, основанные на деревьях решений, являются совершенными моделями человеческого принятия решений и поэтому естественным образом понятны как разработчикам, так и конечным пользователям (Lawrence & Wright, 2001; Cho et al., 2002). То же можно сказать и о таблицах данных (Cragun & Steudel, 1987). Однако со сменой парадигмы в ведущих методологиях искусственного интеллекта на системы машинного обучения, основанные на глубоких нейронных сетях (DNN), появились новшества (Samek et al., 2021).

Простота понимания была принесена в жертву скорости принятия решения и получила название черного ящика – непрозрачного для человеческого понимания, но чрезвычайно способного как в отношении результатов, так и в отношении обучения новым областям. Модели, которые «открывают черный ящик», делая нелинейный и сложный процесс принятия решений понятным для людей-наблюдателей, являются многообещающими решениями проблемы черного ящика в ИИ, но ограничены, по крайней мере, в их нынешнем состоянии, в их способности сделать эти процессы менее непрозрачными для большинства наблюдателей. Искусственный интеллект использует глубокое обучение (DL), алгоритмическую систему глубоких нейронных сетей (algorithmic system of deep neural networks), которые в целом остаются непрозрачными или скрытыми от человеческого понимания.

В чем проявляется непрозрачность и каковы ее причины? Основной целью машинного обучения (ML) является обучение точных систем принятия решений – предикторов (predictors), способных помочь автоматизировать задачи, которые в противном случае должны были бы выполнять люди. Машинное обучение располагает множеством алгоритмов, которые продемонстрировали значительные успехи в науке и промышленности. Наиболее популярными рабочими лошадками ML считаются методы ядра (kernel methods) (Hofmann et al., 2008) и, особенно в последнее десятилетие, методы глубокого обучения (deep learning methods) (Vargas et al., 2018).

Поскольку ML все чаще используется в реальных программах и приложениях, сложилось общее мнение, что высокая точность решения или предсказания на практике может быть недостаточной (Gunning, 2017).

Первая трудность связана с многомасштабной и распределенной природой представлений нейронных сетей. Некоторые нейроны активируются только для нескольких точек данных, в то время как другие действуют более глобально. Таким образом, предсказание представляет собой сумму локальных и глобальных эффектов, что затрудняет (или делает невозможным) поиск корневой точки x, которая линейно расширяется до предсказания для интересующей точки данных. Переход от глобального к локальному эффекту вносит нелинейность, которую невозможно уловить (Samek et al., 2021).

Второй источник нестабильности возникает из-за большой глубины современных нейронных сетей, где наблюдается эффект «разбитого градиента» (shattered gradient problem) (Balduzzi et al., 2017). Градиентом в нейронных сетях называется вектор частных производных функции потерь по весам нейронной сети. Он используется в оптимизаторе весов для улучшения качества модели. Градиент показывает изменение ошибок на разных наборах данных.

Наконец, выделяется проблема объяснимости технологии искусственного интеллекта с необходимостью поиска корневой точки x, на которой будет основано объяснение и которая одновременно близка к данным и не является неблагоприятным примером (the problem of adversarial examples) (Goodfellow et al., 2014). Проблема неблагоприятных примеров объясняется градиентным шумом, который заставляет модель давать «чрезмерную реакцию» (overreact) на определенные пиксельные возмущения, а также высокой размерностью данных (dimensionality of the data) (Najafabadi et al., 2015), где множество мелких пиксельных эффектов суммируются в большой эффект на результат модели (Samek et al., 2021).

Эти особенности работы искусственного интеллекта как класса приводят к тому, что, пока доступны большие данные и огромные вычисления, для достижения сверхчеловеческой производительности требуется «нулевое знание человека» (zero human knowledge) (Silver et al., 2017).

Исследователи предлагают признать, что искусственный интеллект находится внутри социотехнической системы, которая опосредует доверие, а повышая надежность этих систем, стараясь сделать эти процессы менее непрозрачными для большинства наблюдателей, мы тем самым повышаем доверие к искусственному интеллекту (von Eschenbach, 2021). В данном контексте исключение человека из процесса принятия решения добавляет ему доверия, так как исключает фактор субъективности в полученном результате.

В то же время вопрос доверия зависит не только от возможности человека вмешиваться в процесс принятия искусственным интеллектом решения. На современном этапе растет спрос на объяснимый искусственный интеллект (explainable artificial intelligence (XAI). Роман Ямпольский указывает, что «если все, что у нас есть, – это “черный ящик”, то невозможно понять причины сбоев и повысить безопасность системы. Кроме того, если мы привыкнем принимать ответы искусственного интеллекта без объяснения причин, мы не сможем определять, если он начнет давать неправильные или манипулятивные ответы» (Yampolskiy, 2019). Ученый в красках расписывает опасности непрозрачного искусственного интеллекта, предлагая представить, что в ближайшем будущем искусственный интеллект может ошибаться в диагностике заболеваний в 5 % случаев, что приведет к массовым операциям здоровых людей. Отсутствие механизма проверки модели искусственного интеллекта на отклонения и предотвращения подобных сбоев может привести к непоправимым последствиям. Таким образом, прозрачность и подотчетность являются инструментами, способствующими принятию справедливых алгоритмических решений, обеспечивая основу для получения возможности обратиться к значимому объяснению, исправлению или способам установления недостатков, которые могут привести к компенсационным процессам (Koene, 2019).

2. Правовые и этические риски применения непрозрачной технологии

Проблема прозрачности определяется в российской Концепции развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 г. (далее – Концепция 2024) как «использование для принятия решений системами искусственного интеллекта вероятностных оценок и невозможность в ряде случаев полного объяснения принятого ими решения (проблема алгоритмической прозрачности систем искусственного интеллекта)»[2].

В Концепции 2024 прозрачность названа среди таких проблемных направлений регулирования искусственного интеллекта, как соблюдение баланса между требованиями по защите персональных данных и необходимостью их использования для обучения систем искусственного интеллекта; определение предмета и границ регулирования сферы использования систем искусственного интеллекта и робототехники; правовое «делегирование» решений системам искусственного интеллекта и робототехники; ответственность за причинение вреда с использованием систем искусственного интеллекта и робототехники. То есть вопросы правового обеспечения прозрачности искусственного интеллекта играют концептуально важную роль для выработки правовых подходов.

Выше было показано, что разработчик представляет данные, но не может контролировать, на основании каких именно критериев искусственный интеллект сформулировал результат или прогноз. Складывается видение, что иногда осмысленную нейронную сеть создать не представляется возможным. Это связано с трудностями определения входных данных и их фактической недостаточностью. По сути, потеря контроля над искусственным интеллектом основывается на неопределенности данных, с которыми взаимодействует модель (Kharitonova et al., 2021).

Способны ли разработчики и юристы аргументированно вмешаться в работу системы и поспорить с ее выводами, если они не понимают принципов принятия этих решений? Разработчики могут указать критерии принятия решений, но искусственный интеллект самостоятельно может дополнить условно недостающие данные для формирования решений на выходе. Например, машина анализирует точки или пиксели, но не знает, что это цвет кожи или глаз. Манипулирует «кубиками» пикселей, а не общей картиной.

В то же время решения, принимаемые людьми, юристами и даже судьями, чья деятельность в этом аспекте подробно регулируется, несвободны от осознанного и/или бессознательного предубеждения. Исследователи человеческих предрассудков доказали, что люди когнитивно предрасположены к пристрастиям и стереотипам (Plous, 2003; Quillian, 2006), хотя современные формы предрассудков часто трудно обнаружить и они могут быть даже неизвестны самим носителям предрассудков. Практика обоснования решений может быть просто недостаточной для противодействия влиянию различных факторов, а причины, предлагаемые для человека, принимающего решения, могут скрывать мотивы, едва ли известные тем, кто принимает решения (McEwen et al., 2018).

---

[2] *Об утверждении Концепции развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 года*: распоряжение Правительства РФ от 19.08.2020 № 2129-р.

То есть зачастую алгоритмическая и человеческая предвзятость и необъяснимость принятого решения существуют в латентном виде, не осознаваемые ее носителем и не выявленные третьими лицами. Это позволяет предполагать, что бездушный, безэмоциональный алгоритм все еще может выступать более объективным мерилом принятия решения, поскольку избавлен от личных субъективных предвзятостей.

Риск использования непрозрачного искусственного интеллекта приобретает критическое значение, если такая технология обязательно должна быть применена субъектом деятельности. Так, в беспилотном автомобиле в ходе движения решение принимает система искусственного интеллекта, в то время как ответственность за источник повышенной опасности все еще возлагается на водителя (Payre & Cestac, 2014). Или иной пример ближайшей перспективы. Сегодня в медицине все шире используются роботы на основе искусственного интеллекта для ассистирования хирургам (Kalis et al., 2014). В ходе оказания медицинской помощи некоторые процедуры становятся обязательными и, следовательно, врач может оказаться в ситуации, когда его решения повлекли ответственность, хотя фактически причиной вреда стали проблемы программного кода искусственного интеллекта.

Исследуя вопросы правового вмешательства в распространение дипфейков, В. О. Калятин приходит к выводу, что «разрабатывать соответствующее законодательство следует не в отношении дипфейков как таковых, а в отношении использования ИИ в целом» (Калятин, 2022). Перед юристами стоит выбор: оставаться в действующей правовой традиции либо создавать новую. Полагаем, попытки создать правовой режим применения предпринимателями искусственного интеллекта не могут увенчаться успехом в отсутствие понимания технологических особенностей искусственного интеллекта. Однако прозрачность как объяснимость технологии не может быть понята в буквальном смысле. Требуется создание критериев проверки результатов работы искусственного интеллекта для целей соблюдения прав и свобод граждан, защиты государственных и общественных интересов.

3. Автоматизированная система обработки данных и их качество

Если транспарентность сама по себе не свойственна природе алгоритмов (Kalpokas, 2019) при условии, что на входе для запуска приложений искусственного интеллекта предоставляется информация, возникает вопрос о возможности расстановки акцентов о правилах анализа данных алгоритмом искусственного интеллекта.

В литературе выделены несколько аспектов алгоритмической прозрачности и подотчетности, которые включают повышение осведомленности, подотчетность при использовании алгоритмических решений, и прежде всего в государственном секторе, а также нормативный надзор и юридическую ответственность, которые приведут к глобальной координации алгоритмического управления (Koene et al., 2019).

Осведомленность, рассматриваемая многими исследователями как решение проблемы прозрачности, может пониматься по-разному. Прежде всего при обеспечении прозрачности искусственного интеллекта во главу угла ставится работа с данными и осведомленность об их использовании в определенном ключе.

Обратим внимание, во многих юрисдикциях анализ данных и его пределы установлены в отношении персональных данных. В России действуют положения ст. 16

Закона о персональных данных[3], согласно которой запрещается принятие на основании исключительно автоматизированной обработки персональных данных решений, порождающих юридические последствия в отношении субъекта персональных данных или иным образом затрагивающих его права и законные интересы, за исключением случаев, предусмотренных законом. К таким случаям отнесены ситуации, когда решение, порождающее юридические последствия в отношении субъекта персональных данных или иным образом затрагивающее его права и законные интересы, принято на основании исключительно автоматизированной обработки его персональных данных при наличии согласия в письменной форме субъекта персональных данных (п. 2 ст. 16 Закона о персональных данных).

Приведенное положение российского законодательства сопоставимо со ст. 15 ныне не применяемой Директивы № 95/46/ЕС Европейского парламента и Совета ЕС «О защите физических лиц при обработке персональных данных и о свободном обращении таких данных»[4]. В действующем GDPR содержатся аналогичные правила. Статья 22(3) Общего регламента по защите данных гласит, что в некоторых случаях автоматизированной обработки «контролер данных должен принять соответствующие меры для защиты прав, свобод и законных интересов субъекта данных, по крайней мере, право на вмешательство человека со стороны контролера выражать свою точку зрения и оспаривать решение»[5].

В то же время для обеспечения осведомленности важно и раскрытие информации о данных, на которых базируется принятое решение. Сюда относятся вопросы достоверности и нейтральности, репрезентативности данных, непредвзятости методов их обработки и анализа, а также информация о самообучении искусственного интеллекта.

Прозрачность как объяснимость и недискриминационность технологии зависит от качества данных, с которыми работает система искусственного интеллекта. Исследователи установили (Buolamwini & Gebru, 2018), что все популярные системы распознавания лиц точнее всего определяют мужчин со светлой кожей (2,4 % ошибок). Но чаще всего ошибаются при обнаружении темнокожих женщин (61 % ошибок). По сути, тем самым было доказано, что «фотографий женщин с темной кожей в базах данных меньше всего; разработчики таких систем сами преимущественно являются белыми мужчинами; датчики камер хуже распознают детали в темных цветах»[6].

Приведенный пример демонстрирует, что недостаточно поставить под сомнение достоверность данных, доступных искусственному интеллекту. Проблема качества данных в том, что доступные данные не обладали нейтральностью даже при условии, что они могли считаться репрезентативными. Системы распознавания лиц

---

3 О персональных данных: Федеральный закон от 27.07.2006 № 152-ФЗ. *СПС КонсультантПлюс*. https://www.consultant.ru/document/cons_doc_LAW_61801/

4 Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. (1995, November 23). *Official Journal of the European Communities, L 281*, 31–50.

5 *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)*. http://eur-lex.europa.eu/

6 Ученые все чаще не могут объяснить, как работает ИИ. Теория «черного» и «белого» ящика. (2022, 23 ноября). *Хабр*. https://habr.com/ru/company/getmatch/blog/700736/

используются во многих странах. В том числе в работе правоохранительных органов. Доказано, что если вы одно из расовых меньшинств в такой стране, эта система будет чаще определять в вас преступника[7]. Высказано мнение, что если обучать модели искусственного интеллекта с помощью больших данных, встроенные в них расовые и другие предубеждения будут неизбежностью (Bender et al., 2021), ведь у некоторых групп людей просто меньше доступа к Интернету, а данные о них мало представлены на разного рода ресурсах (жители отдаленных населенных пунктов, например, в сравнении с программистами).

В целом, мы полагаем, уязвимостью подхода, сосредоточенного только на персональных данных, является несоответствие его действительности. В дополнение к данным, которые поступают непосредственно от пользователей в алгоритм, бизнес должен пополнять эти категории пользователей дополнительными аналитическими данными (Camilleri, 2018), более подробно описывающими различные группы и делающими обоснование классификации еще более ясным.

Следовательно, необходимо устанавливать правила для определения качественного (достоверного и нейтрального набора данных) в ситуации, когда ограничить такой набор данных не представляется возможным. Риск необъяснимых предвзятых решений искусственного интеллекта придется исключать путем обучения с подкреплением и аудита полученного результата.

Это приводит к мысли, что объяснения понятного человеку алгоритма при использовании метода черного ящика ждать не стоит, но и юридического значения раскрытие алгоритма здесь иметь не будет. Научить систему искусственного интеллекта понимать этические ценности невозможно, юристы могут лишь указать критерии проверки решения искусственного интеллекта на объективность. Не во всех случаях можно поставить человека на выходе для проверки результата. Следовательно, право может узаконить лишь необходимость контроля со стороны программы, созданной независимыми разработчиками.

В связи с этим труднодостижимым представляется предложение достичь прозрачности искусственного интеллекта не в отношении к системе в целом, а через объяснение логики отдельных индивидуальных решений (Кутейников и др., 2018). В качестве таких мер авторами предлагается анализ входных данных, статистическое объяснение, проверка архитектуры/кода и статистический анализ, определение чувствительности отдельных данных (какие именно переменные предопределяют результат) (Кутейников и др., 2018).

Напротив, более выполнимым представляется требовать открытости алгоритма с указанием общей логики принятия решений в целом. Эта система принята на вооружение в штате Калифорния, США. В феврале 2020 г. там был принят Закон о подотчетности автоматизированных систем принятия решений, в котором предписано проводить систематический контроль и выявлять ошибки функционирования автоматизированных систем, а также направлять полученные отчеты в Департамент регулирования бизнеса (Department of Business Oversight) начиная с 1 января 2022 г. и размещать их в открытом доступе в сети Интернет[8].

---

[7] Ученые все чаще не могут объяснить, как работает ИИ. Теория «черного» и «белого» ящика. (2022, 23 ноября). *Хабр*. https://habr.com/ru/company/getmatch/blog/700736/

[8] USA. State of California. Automated Decision Systems Accountability Act of 2020. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB2269

На это же обращают внимание исследователи GDPR. Как пишут А. Селбст и Дж. Паулс, «проблема в том, что для проверки точности защиты и потенциального оспаривания ее правильности необходимы конкретные объяснения решения, включая веса и факторы, используемые для его достижения» (Selbst & Powles, 2018), что составляет техническую, не всегда существующую в рамках понимания человека информацию. С этих позиций закрепленное в ст. 22 (3) Общего регламента по защите данных «право» недостаточно подробно раскрыто в законодательстве и подвергается обоснованной критике в связи с невыполнимостью. Большинство ученых сходятся во мнении, что право на разъяснение индивидуальных решений, которое может включать глобальные или локальные разъяснения, не следует из ст. 22(3) GDPR (Wachter, 2017; Goodman & Flaxman, 2017). Статья 15(1)(h) GDPR предусматривает, что в случае автоматизированной обработки в смысле ст. 22(1) GDPR контролер должен предоставить «значимую информацию о задействованной логике». Некоторые исследователи убеждены, что это относится только к общей структуре и архитектуре модели обработки, но нет необходимости в объяснениях отдельных решений или конкретных весов и характеристик модели (Wachter, 2017; Malgieri & Comandé, 2017).

В отсутствие стандартизированного подхода к обоснованию индивидуальных и общих решений без ответа остаются не только вопросы о том, кому вообще должна раскрываться логика принятия решения, только ли пользователям или субъектам данных либо всем заинтересованным лицам, в каком объеме и др.

В данном контексте следует обратить внимание на значимость применения искусственного интеллекта в отношениях с участием государства. Согласимся, что необходимо закрепление норм об обязательной общедоступности результатов использования государственными органами технологий искусственного интеллекта и больших данных. Как убедительно доказывает вслед за европейскими коллегами В. В. Силкин, если государство осуществляет свои функции с применением искусственного интеллекта, прозрачность технологии необходима. При этом автор предлагает возложить на государственные органы обязанность обосновывать и раскрывать цели использования технологий автоматизированной обработки данных. Возможности больших данных и искусственного интеллекта достаточно обширны, однако использование их государством должно быть обусловлено необходимостью достижения именно общественно значимых целей (Силкин, 2021). Однако мы полагаем, при общем распространении данной технологии, обоснование применения искусственного интеллекта в отдельных видах государственной деятельности решит эту задачу в целом, но не обеспечит прозрачность решений.

Однако хотелось бы подчеркнуть, что принцип прозрачности работы искусственного интеллекта не тождественен принципу прозрачности деятельности органов власти или иных операторов данных по применению автоматизированных систем. В. В. Силкин предлагает «при реализации принципа транспарентности в деятельности государственных органов по применению автоматизированных систем обработки данных исходить из открытости информации о целях, способах и результатах их использования» (Силкин, 2021). При этом автор верно утверждает, что «в то же время в комплексных автоматизированных системах обработки алгоритмы формируются и усложняются самостоятельно, что исключает возможность предугадать или заранее однозначно понять все возможности таких систем» (Силкин, 2021).

Полагаем, в данном случае легко может возникнуть подмена понятий: прозрачность принятия решения искусственным интеллектом не связана исключительно

с прозрачностью целей, установленных для применения алгоритма. Нужно разделять прозрачность алгоритма как принцип регулирования искусственного интеллекта и прозрачность деятельности (госуправления, гражданского оборота и т. п.), которая достигается отчасти с применением технологии искусственного интеллекта. Фактически второй аспект автоматической обработки данных наравне с качеством данных составляет прозрачность целей и методик применения искусственного интеллекта. В этом принцип прозрачности работы искусственного интеллекта совпадает, но не тождествен понятию прозрачности деятельности.

Возможно, цели использования искусственного интеллекта можно было бы определить как цели принятия решений с объяснением процессов искусственного интеллекта по стандартизированной форме, которая требует регулярного обновления каждый раз, когда бизнес меняет свои методы автоматизированной обработки (Wulf & Seizov, 2022).

Представляется, что закрепление права на осведомленность о применении технологии искусственного интеллекта при автоматизированной обработке данных для принятия решений не исчерпывает возможные правовые средства для достижения прозрачности интеллектуальных систем. Более того, защита прав граждан не достигается, входя в противоречие с методами работы искусственного интеллекта, и прежде всего черного ящика. Раскрытие задач и приоритетов алгоритма для целей принятия решения может строиться на стандартах, но с учетом необходимости обращения к новым данным и восполнения пробелов в полученных от пользователя данных, достичь прозрачности и в этом направлении не представляется возможным.

4. Открытость алгоритмов и результатов их работы

Раскрытие сведений о разработке программного обеспечения, кода и порядка его работы также стоит в приоритете обеспечения прозрачности искусственного интеллекта.

В ст. 16 Закона о персональных данных, по справедливому мнению А. И. Савельева, условиями использования инструментов автоматизированной обработки персональных данных для целей принятия юридически значимых решений в отношении субъекта предусмотрены «дополнительные информационные обязанности оператора, выражающиеся в его обязанности предоставить субъекту разъяснения относительно порядка принятия такого решения и возможных юридических последствий такого решения» (Савельев, 2021).

Проект Закона об искусственном интеллекте[9], представленный Европейской комиссией 21 апреля 2021 г., предлагает для обеспечения прозрачности принимаемых искусственным интеллектом решений, например, раскрывать информацию о характеристиках, возможностях и ограничениях системы искусственного интеллекта, предназначении системы, а также информацию, необходимую для обслуживания систем искусственного интеллекта.

[9] European Commission. (2021, April 21). *Proposal for a Regulation laying down harmonised rules on artificial intelligence*. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

В США с 2022 г. обсуждается проект закона об алгоритмической подотчетности[10], согласно которому от компаний, использующих системы искусственного интеллекта, потребуется проводить оценку своих автоматизированных систем в соответствии с правилами Федеральной торговой комиссии на предмет обеспечения недискриминации пользователей и конфиденциальности информации.

Введение на законодательном уровне требования о раскрытии алгоритма дискутируется и в России, хотя пока только в отношении рекомендательных алгоритмов социальных сетей[11]. Аналогичные требования содержатся также в праве Китая – в Положении об управлении алгоритмическими рекомендациями информационных услуг в сети Интернет[12].

В то же время требование об абсолютном раскрытии исходного или архитектуры применимых моделей искусственного интеллекта не представляется обоснованным. Во всяком случае такое раскрытие не может быть доскональным по ряду причин. Технологически такое раскрытие является весьма затратным и требует огромных ресурсов. Публикация исходного кода также может приводить к нарушениям права интеллектуальной собственности или нарушению коммерческой тайны. Такой подход к прозрачности вступает в противоречие с имеющимися правовыми режимами, которым подчинены те или иные составляющие единой технологии искусственного интеллекта.

При этом убедительной кажется позиция, согласно которой системы искусственного интеллекта в первую очередь охраняются как коммерческая тайна (торговый секрет), поскольку попытки защитить системы искусственного интеллекта в соответствии с законами об авторском праве и патентах наталкиваются на трудности (Foss-Solbrekk, 2021). Проблема предоставления авторско-правовой защиты самому алгоритму связана с постоянным его изменением и дописыванием в ходе самообучения и автономной работы. Спорным является и вопрос о самом творческом характере происхождения искусственного интеллекта как объекта правовой охраны. Получение патентов на системы искусственного интеллекта также затруднено. Эта ситуация наблюдается в разных правопорядках. Например, в российском законодательстве об авторском праве алгоритмы по существу получили самостоятельное регулирование в рамках ст. 1261 ГК РФ, в ЕС алгоритмы исключены из-под защиты копирайта в соответствии с Директивой ЕС по программному обеспечению[13]. Во всяком случае идущие бурные дискуссии об отнесении искусственного интеллекта к тому или иному виду объектов права интеллектуальной собственности далеки от завершения.

Защита моделей искусственного интеллекта с помощью правового режима коммерческой тайны приводит к столкновению требований прозрачности

---

10 Metcalf, J., Smith, B., & Moss, E. (2022, February 9). A New Proposed Law Could Actually Hold Big Tech Accountable for Its Algorithms. *Slate*. https://slate.com/technology/2022/02/algorithmic-accountability-act-wyden.html

11 СМИ: в Госдуму планируют внести проект о регулировании рекомендательных сервисов в соцсетях. (2021, 15 октября). *Парламентская газета*. https://www.pnp.ru/politics/smi-v-gosdumu-planiruyut-vnesti-proekt-o-regulirovanii-rekomendatelnykh-servisov-v-socsetyakh.html

12 互联网信息服务算法推荐管理规定. (2021, 31 декабря). http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm

13 *О правовой охране компьютерных программ (кодифицированная версия)*: Директива № 2009/24/ЕС Европейского парламента и Совета Европейского союза. https://base.garant.ru/71657620/

и подотчетности. В частности, исследователи приходят к выводу, что Европейская директива 2016/943 оставляет мало места для гипотез об алгоритмической прозрачности (Maggiolino, 2018).

В определении понятия «коммерческая тайна», приведенном в тексте Директивы 2016/943, речь идет о коммерческой ценности без указания на ее фактический или потенциальный характер. Согласно ст. 2 (1), коммерческая тайна означает информацию, которая отвечает всем следующим требованиям:

«(a) она является секретной, то есть не является общеизвестной или легкодоступной для лиц в кругах, которые обычно имеют дело с рассматриваемой информацией;

(b) она имеет коммерческую ценность, потому что является секретной;

(c) при данных обстоятельствах лицо, на законном основании контролирующее информацию, приняло разумные меры для сохранения ее в тайне»[14].

При таком подходе требование о раскрытии кода или архитектуры модели искусственного интеллекта будет означать потерю конкурентного преимущества на рынке. Поскольку защита коммерческой тайны существует до тех пор, пока информация остается конфиденциальной, и требует, чтобы субъекты предпринимали шаги для обеспечения конфиденциальности, защита коммерческой тайны способствует алгоритмической непрозрачности.

Из этого следует, что отказ от секретности алгоритмов искусственного интеллекта должен сопровождаться усилением защиты интересов раскрывшей сведения стороны перед третьими лицами, как сейчас это происходит в патентном праве.

5. Применимые правовые средства для предотвращения проблем непрозрачности правовых решений: возразить или отказаться

Осведомленность о применении искусственного интеллекта для принятия решений также приводит к необходимости обсуждения права на отказ от применения технологии искусственного интеллекта в конкретном случае, а также права на возражение против принятого решения.

В силу п. 3 ст. 16 российского Закона о персональных данных оператор обязан предоставить субъекту персональных данных возможность заявить возражение против решения, принятого в результате автоматизированной обработки данных, а также разъяснить порядок защиты субъектом персональных данных своих прав и законных интересов. А. И. Савельев поясняет, что речь «идет о тех правах, которые касаются именно принятия юридически значимых решений по результатам исключительно автоматизированной обработки персональных данных. В частности, данный порядок защиты предполагает уведомление субъекта о его праве требовать вмешательства человека в процесс принятия решения, которое является неотъемлемой частью права на предоставление возражений» (Савельев, 2021).

[14] *О защите конфиденциальных ноу-хау и деловой информации (коммерческой тайны) от незаконного приобретения, использования и раскрытия*: Директива № 2016/943 Европейского парламента и Совета Европейского союза. https://base.garant.ru/71615160/

В России также обсуждается законопроект, в котором предполагается предоставить пользователям возможность полностью или частично отказаться от применения рекомендательных алгоритмов.

В Китае такой подход уже принят к реализации и проверяется на исполнимость. По мнению исследователей, раскрытие логики алгоритма должно устранить риски необоснованных отказов поставщиков услуг в случае алгоритмической рекомендации предоставить необходимую информацию, такую как область применения алгоритма, пользователь услуги, уровень риска алгоритма и другие, на том основании, что четкие законодательные установления в этом отношении отсутствуют (Xu Ke & Liu Chang, 2022). Однако иногда такое раскрытие ведет не к успеху, а является декоративным оформлением сайта.

Таким образом, добиваться прозрачности и объяснимости работы искусственного интеллекта в понятной человеку форме юристам стоит добиваться в том числе и для того, чтобы обеспечить возможность пользователю системы искусственного интеллекта возражать против принятого решения. Этот вопрос также связан и с определением субъекта ответственности за принимаемые решения, которые могут существенно ущемлять права человека.

Ученые заметили, что за текущей дискуссией о требованиях защиты данных в отношении объяснимости упускается из виду важность данной характеристики для оценки договорной и деликтной ответственности в отношении использования инструментов искусственного интеллекта (Hacker et al., 2020). В связи с этим дальнейшей конкретизации потребуют положения законодательства по использованию искусственного интеллекта в сфере усиления обязанности для разработчиков, производителей и поставщиков услуг искусственного интеллекта постоянно оценивать возможные неблагоприятные последствия применения искусственного интеллекта для прав человека и основных свобод и, учитывая эти последствия, принимать меры по предотвращению и смягчению рисков (Дьяконова и др., 2022).

Выводы

Прозрачность является необходимым условием для признания искусственного интеллекта заслуживающим доверия. Наиболее плодотворный путь для установления доверия к искусственному интеллекту заключается в том, чтобы признать данную технологию частью сложной социотехнической системы, которая опосредует доверие и повышает надежность этих систем.

Большая часть дебатов вокруг прозрачности искусственного интеллекта с юридической точки зрения сосредоточена на законах о защите данных. Полагаем, круг этих дискуссий следует расширять. Прозрачность и объяснимость технологии искусственного интеллекта важна не только для защиты персональных данных, но и в иных ситуациях автоматизированной обработки данных, когда для принятия решений недостающие из входящей информации технологические данные восполняются из открытых источников, в том числе не имеющих значения хранилищ персональных данных. Законодатель может добиваться лишь установления некоего стандарта, закрепляющего компромисс между возможностями и преимуществами технологии, точностью и объяснимостью результата ее работы и правами участников общественных отношений. Введение сертификации моделей искусственного интеллекта, обязательных к применению, позволит решить вопросы ответственности

обязанных применять такие системы субъектов. В контексте вопроса о профессиональной ответственности профессиональных субъектов, таких как врачи, военные, органы корпоративного управления юридического лица, требуется ограничить обязательное применение искусственного интеллекта в случаях, если не обеспечена его достаточная прозрачность.

Развитие правовой дискуссии должно двигаться по выработке предложений о содержании права на отказ от применения автоматизированной обработки данных для принятия решений и права на возражения против принятых таким способом решений.

Список литературы

Дьяконова, М. О., Ефремов, А. А., Зайцев, О. А. и др.; И. И. Кучерова, С. А. Синицына (ред.). (2022). Цифровая экономика: актуальные направления правового регулирования: научно-практическое пособие. Москва: ИЗиСП, НОРМА.

Калятин, В. О. (2022). Дипфейк как правовая проблема: новые угрозы или новые возможности? *Закон, 7*, 87–103. https://doi.org/10.37239/0869-4400-2022-19-7-87-103

Кутейников, Д. Л., Ижаев, О. А., Зенин, С. С., Лебедев, В. А. (2020). Алгоритмическая прозрачность и подотчетность: правовые подходы к разрешению проблемы «черного ящика». *Lex russica (Русский закон)*, *73*(6), 139–148. https://doi.org/10.17803/1729-5920.2020.163.6.139-148

Савельев, А. И. (2021). Научно-практический постатейный комментарий к Федеральному закону «О персональных данных» (2-е изд., перераб. и доп.). Москва: Статут.

Силкин, В. В. (2021). Транспарентность исполнительной власти в цифровую эпоху. *Российский юридический журнал*, *4*, 20–31. https://doi.org/10.34076/20713797_2021_4_20

Balduzzi, D., Frean, M., Leary, L., Lewis, J. P., Ma, K. W. D., & McWilliams, B. (2017). The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning* (pp. 342–350). PMLR.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91). PMLR.

Camilleri, M. A. (2018). Market segmentation, targeting and positioning. In *Travel marketing, tourism economics and the airline product* (pp. 69–83). New York: Springer.

Cho, Y. H., Kim, J. K., & Kim, S. H. (2002). A personalized recommender system based on web usage mining and decision tree induction. *Expert systems with Applications*, *23*(3), 329–342. https://doi.org/10.1016/s0957-4174(02)00052-0

Cragun, B. J., & Steudel, H. J. (1987). A decision-table-based processor for checking completeness and consistency in rule-based expert systems. *International Journal of Man-Machine studies*, *26*(5), 633–648. https://doi.org/10.1016/s0020-7373(87)80076-7

Foss-Solbrekk, K. (2021). Three routes to protecting AI systems and their algorithms under IP law: The good, the bad and the ugly. *Journal of Intellectual Property Law & Practice*, *16*(3), 247–258. https://doi.org/10.1093/jiplp/jpab033

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. arXiv preprint arXiv:1412.6572.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gunning, D. (2017). Explainable artificial intelligence (XAI). *Defense advanced research projects agency (DARPA)*. nd Web. *2*(2), 1.

Hacker, P., Krestel, R., Grundmann, S., & Naumann, F. (2020). Explainable AI under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law*, *28*(4), 415–439. https://doi.org/10.1007/s10506-020-09260-6

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). New York: Springer.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). *Kernel methods in machine learning*.

Kalis, B., Collier, M., & Fu, R. (2014). 10 promising AI applications in health care. *Harvard Business Review*.

Kalpokas, I. (2019). *Algorithmic Governance. Politics and Law in the Post-Human Era*. Cham: Palgrave Pivot.
Kharitonova, Y. S., Savina, V. S., & Pagnini, F. (2021). Artificial Intelligence's Algorithmic Bias: Ethical and Legal Issues. *Perm U. Herald Jurid. Sci*, *3*(53), 488–515. https://doi.org/10.17072/1995-4190-2021-53-488-515
Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). *A governance framework for algorithmic accountability and transparency*. Panel for the Future of Science and Technology.
Lawrence, R. L., & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric engineering and remote sensing*, *67*(10), 1137–1142.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.
Maggiolino, M. (2018). *EU trade secrets law and algorithmic transparency*. SSRN 3363178.
Malgieri, G., & Comandé, G. (2017). Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*, *7*(4), 243–265. https://doi.org/10.1093/idpl/ipx019
McEwen, R., Eldridge, J., & Caruso, D. (2018). Differential or deferential to media? The effect of prejudicial publicity on judge or jury. *The International Journal of Evidence & Proof*, *22*(2), 124–143. https://doi.org/10.1177/1365712718765548
Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21. https://doi.org/10.1186/s40537-014-0007-7
Pak, M., & Kim, S. (2017). A review of deep learning in image recognition. In *2017 4th international conference on computer applications and information processing technology (CAIPT)*. https://doi.org/10.1109/caipt.2017.8320684
Payre, W., Cestac, J., & Delhomme, P. (2014). Intention to use a fully automated car: Attitudes and a priori acceptability. *Transportation research part F: traffic psychology and behavior*, *27*, 252–263. https://doi.org/10.1016/j.trf.2014.04.009
Plous, S. E. (2003). *Understanding prejudice and discrimination*. McGraw-Hill.
Quillian, L. (2006). New approaches to understanding racial prejudice and discrimination. *Annu. Rev. Sociol.*, *32*, 299–328. https://doi.org/10.1146/annurev.soc.32.061604.123132
Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), 247–278. https://doi.org/10.1109/jproc.2021.3060483
Selbst, A., & Powles, J. (2017). "Meaningful information" and the right to explanation. *International Data Privacy Law*, *7*(4), 233–242. https://doi.org/10.1093/idpl/ipx022
Silver, D. et al. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354.
Vargas, R., Mosavi, A., & Ruiz, R. (2018). *Deep learning: a review*. Preprints.org. https://doi.org/10.20944/preprints201810.0218.v1
von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, *34*(4), 1607–1622.
Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005
Wulf, A. J. & Seizov, O. (2022). "Please understand we cannot provide further information": evaluating content and transparency of GDPR-mandated AI disclosures. *AI & SOCIETY*, 1–22. https://doi.org/10.1007/s00146-022-01424-z
Yampolskiy, R. V. (2019). *Unexplainability and incomprehensibility of artificial intelligence*. arXiv preprint arXiv:1907.03869
许可、刘畅. 论算法备案制度 // 人工智能. 2022. № 1. P. 66. [Xu Ke, Liu Chang. (2022). On the Algorithm Filing System. *Artificial Intelligence*, *1*, 66.]

Сведения об авторе

Харитонова Юлия Сергеевна – доктор юридических наук, профессор, профессор кафедры предпринимательского права, руководитель Центра правовых исследований искусственного интеллекта и цифровой экономики, Московский государственный университет имени М. В. Ломоносова
Адрес: 119991, Российская Федерация, г. Москва, Ленинские горы, 1
E-mail: sovet2009@rambler.ru
ORCID ID: https://orcid.org/0000-0001-7622-6215
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57316440400
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/708572
Google Scholar ID: https://scholar.google.ru/citations?user=61mQtb4AAAAJ
РИНЦ Author ID: https://elibrary.ru/author_items.asp?authorid=465239

Конфликт интересов

Автор является членом редакционной коллегии журнала, статья прошла рецензирование на общих основаниях.

Финансирование

Исследование не имело спонсорской поддержки.

Тематические рубрики

Рубрика OECD: 5.05 / Law
Рубрика ASJC: 3308 / Law
Рубрика WoS: OM / Law
Рубрика ГРНТИ: 10.23.01 / Общие вопросы предпринимательского права
Специальность ВАК: 5.1.3 / Частно-правовые (цивилистические) науки

История статьи

Дата поступления – 6 марта 2023 г.
Дата одобрения после рецензирования – 13 апреля 2023 г.
Дата принятия к опубликованию – 16 июня 2023 г.
Дата онлайн-размещения – 20 июня 2023 г.

Research article
DOI: https://doi.org/10.21202/jdtl.2023.14

Legal Means of Providing the Principle of Transparency of the Artificial Intelligence

Yuliya S. Kharitonova

Lomonosov Moscow State University
Moscow, Russian Federation

Keywords

Algorithm,
artificial intelligence,
automated data processing,
autonomy,
decision-making,
digital economy,
digital technologies,
ethics,
law,
transparency

Abstract

Objective: to analyze the current technological and legal theories in order to define the content of the transparency principle of the artificial intelligence functioning from the viewpoint of legal regulation, choice of applicable means of legal regulation, and establishing objective limits to legal intervention into the technological sphere through regulatory impact.

Methods: the methodological basis of the research is the set of general scientific (analysis, synthesis, induction, deduction) and specific legal (historical-legal, formal-legal, comparative-legal) methods of scientific cognition.

Results: the author critically analyzed the norms and proposals for normative formalization of the artificial intelligence transparency principle from the viewpoint of impossibility to obtain the full technological transparency of artificial intelligence. It is proposed to discuss the variants of managing algorithmic transparency and accountability based on the analysis of social, technical and regulatory problems created by algorithmic systems of artificial intelligence. It is proved that transparency is an indispensible condition to recognize artificial intelligence as trustworthy. It is proved that transparency and explainability of the artificial intelligence technology is essential not only for personal data protection, but also in other situations of automated data processing, when, in order to make a decision, the technological data lacking in the input information are taken from open sources, including those not having the status of a personal data storage. It is proposed to legislatively stipulate the obligatory audit and to introduce a standard, stipulating a compromise between the technology abilities and advantages, accuracy and explainability of its result, and the rights of the participants of civil relations. Introduction

of certification of the artificial intelligence models, obligatory for application, will solve the issues of liability of the subjects obliged to apply such systems. In the context of professional liability of professional subjects, such as doctors, militants, or corporate executives of a juridical person, it is necessary to restrict the obligatory application of artificial intelligence if sufficient transparency is not provided.

Scientific novelty: the interdisciplinary character of the research allowed revealing the impossibility and groundlessness of the requirements to completely disclose the source code or architecture of the artificial intelligence models. The principle of artificial intelligence transparency may be satisfied through elaboration and provision of the right of the data subject and the subject, to whom the decision made as a result of automated data processing is addressed, to reject using automated data processing in decision-making, and the right to object to the decisions made in such a way.

Practical significance: is due to the actual absence of sufficient regulation of the principle of transparency of artificial intelligence and results of its functioning, as well as the content and features of the implementation of the right to explanation the right to objection of the decision subject. The most fruitful way to establish trust towards artificial intelligence is to recognize this technology as a part of a complex sociotechnical system, which mediates trust, and to improve the reliability of these systems. The main provisions and conclusions of the research can be used to improve the legal mechanism of providing transparency of the artificial intelligence models applied in state governance and business.

For citation

Kharitonova, Yu. S. (2023). Legal Means of Providing the Principle of Transparency of the Artificial Intelligence. *Journal of Digital Technologies and Law, 1*(2), 337–358. https://doi.org/10.21202/jdtl.2023.14

References

Balduzzi, D., Frean, M., Leary, L., Lewis, J. P., Ma, K. W. D., & McWilliams, B. (2017). The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning* (pp. 342–350). PMLR.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91). PMLR.

Camilleri, M. A. (2018). Market segmentation, targeting and positioning. In *Travel marketing, tourism economics and the airline product* (pp. 69–83). New York: Springer.

Cho, Y. H., Kim, J. K., & Kim, S. H. (2002). A personalized recommender system based on web usage mining and decision tree induction. *Expert systems with Applications*, *23*(3), 329–342. https://doi.org/10.1016/s0957-4174(02)00052-0

Cragun, B. J., & Steudel, H. J. (1987). A decision-table-based processor for checking completeness and consistency in rule-based expert systems. *International Journal of Man-Machine studies*, *26*(5), 633–648. https://doi.org/10.1016/s0020-7373(87)80076-7

Dyakonova, M. O., Efremov, A. A., Zaitsev, O. A., et al.; I. I. Kucherova, S. A. Sinitsyna (Eds.). (2022). *Digital economy: topical areas of legal regulation: scientific-practical tutorial*. Moscow: IZISP, NORMA. (In Russ.).

Foss-Solbrekk, K. (2021). Three routes to protecting AI systems and their algorithms under IP law: The good, the bad and the ugly. *Journal of Intellectual Property Law & Practice*, *16*(3), 247–258. https://doi.org/10.1093/jiplp/jpab033

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. arXiv preprint arXiv:1412.6572.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gunning, D. (2017). Explainable artificial intelligence (XAI). *Defense advanced research projects agency (DARPA)*. nd Web. *2*(2), 1.

Hacker, P., Krestel, R., Grundmann, S., & Naumann, F. (2020). Explainable AI under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law*, *28*(4), 415–439. https://doi.org/10.1007/s10506-020-09260-6

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). New York: Springer.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). *Kernel methods in machine learning*.

Kalis, B., Collier, M., & Fu, R. (2014). 10 promising AI applications in health care. *Harvard Business Review*.

Kalpokas, I. (2019). *Algorithmic Governance. Politics and Law in the Post-Human Era*. Cham: Palgrave Pivot.

Kalyatin, V. O. (2022). Deepfake as a legal problem: new threats or new opportunities? *Zakon*, *7*, 87–103. (In Russ.). https://doi.org/10.37239/0869-4400-2022-19-7-87-103

Kharitonova, Y. S., Savina, V. S., & Pagnini, F. (2021). Artificial Intelligence's Algorithmic Bias: Ethical and Legal Issues. *Perm U. Herald Jurid. Sci*, *3*(53), 488–515. https://doi.org/10.17072/1995-4190-2021-53-488-515

Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). *A governance framework for algorithmic accountability and transparency*. Panel for the Future of Science and Technology.

Kuteynikov, D. L., Izhaev, O. A., Zenin, S. S., & Lebedev, V. A. (2020). Algorithmic Transparency and Accountability: Legal Approaches to Solving the "Black Box" Problem. *Lex Russica*, *73*(6), 139–148. (In Russ.). https://doi.org/10.17803/1729-5920.2020.163.6.139-148

Lawrence, R. L., & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric engineering and remote sensing*, *67*(10), 1137–1142.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Maggiolino, M. (2018). *EU trade secrets law and algorithmic transparency*. Available at SSRN 3363178.

Malgieri, G., & Comandé, G. (2017). Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*, *7*(4), 243–265. https://doi.org/10.1093/idpl/ipx019

McEwen, R., Eldridge, J., & Caruso, D. (2018). Differential or deferential to media? The effect of prejudicial publicity on judge or jury. *The International Journal of Evidence & Proof*, *22*(2), 124–143. https://doi.org/10.1177/1365712718765548

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21. https://doi.org/10.1186/s40537-014-0007-7

Pak, M., & Kim, S. (2017). A review of deep learning in image recognition. In *2017 4th international conference on computer applications and information processing technology (CAIPT)*. https://doi.org/10.1109/caipt.2017.8320684

Payre, W., Cestac, J., & Delhomme, P. (2014). Intention to use a fully automated car: Attitudes and a priori acceptability. *Transportation research part F: traffic psychology and behavior*, *27*, 252–263. https://doi.org/10.1016/j.trf.2014.04.009

Plous, S. E. (2003). *Understanding prejudice and discrimination*. McGraw-Hill.

Quillian, L. (2006). New approaches to understanding racial prejudice and discrimination. *Annu. Rev. Sociol.*, *32*, 299–328. https://doi.org/10.1146/annurev.soc.32.061604.123132

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), 247–278. https://doi.org/10.1109/jproc.2021.3060483

Savelyev, A. I. (2021). *Scientific-practical article-by-article commentary to Federal Law "On personal data"* (2nd ed., amended and abridged). Moscow: Statut. (In Russ.).

Selbst, A., & Powles, J. (2017). "Meaningful information" and the right to explanation. *International Data Privacy Law*, *7*(4), 233–242. https://doi.org/10.1093/idpl/ipx022

Silkin, V. V. (2021). Transparency of executive power in digital epoch. *Russian Juridical Journal*, *4*, 20–31. (In Russ.). https://doi.org/10.34076/20713797_2021_4_20

Silver, D. et al. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354.

Vargas, R., Mosavi, A., & Ruiz, R. (2018). *Deep learning: a review*. Preprints.org. https://doi.org/10.20944/preprints201810.0218.v1

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, *34*(4), 1607–1622.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wulf, A. J. & Seizov, O. (2022). "Please understand we cannot provide further information": evaluating content and transparency of GDPR-mandated AI disclosures. *AI & SOCIETY*, 1–22. https://doi.org/10.1007/s00146-022-01424-z

Yampolskiy, R. V. (2019). *Unexplainability and incomprehensibility of artificial intelligence*. arXiv preprint arXiv:1907.03869

许可、刘畅. 论算法备案制度 // 人工智能. 2022. № 1. P. 66. [Xu Ke, Liu Chang. (2022). On the Algorithm Filing System. *Artificial Intelligence*, *1*, 66.]

Author information

Yuliya S. Kharitonova – Doctor of Law, Professor, Professor of the Department of Entrepreneurial Law, Head of the Center for legal research of artificial intelligence and digital economy, Lomonosov Moscow State University
Address: 1 Leninskiye gory, 119991 Moscow, Russian Federation
E-mail: sovet2009@rambler.ru
ORCID ID: https://orcid.org/0000-0001-7622-6215
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57316440400
Web of Science Researcher ID:
https://www.webofscience.com/wos/author/record/708572
Google Scholar ID: https://scholar.google.ru/citations?user=61mQtb4AAAAJ
RSCI Author ID: https://elibrary.ru/author_items.asp?authorid=465239

Conflict of interests

The author is a member of the Editorial Board of the Journal; the article has been reviewed on general terms.

Funding

The research was not sponsored.

Thematic rubrics

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

Article history

Date of receipt – March 6, 2023
Date of approval – April 13, 2023
Date of acceptance – June 16, 2023
Date of online placement – June 20, 2023