Реализация инновационных механизмов всегда связано с достижением его конкурентоспособности, привлекательности для привлечения новых инвестиций. Конкурентоспособность и инновационно-инвестиционная привлекательность региона зависит от ряда качественных параметров, формирующих качество среды для ведения бизнеса. Наиболее актуальными параметрами для региона являются уровень образования населения, качество информационных и телекоммуникационных технологий. В последнее время все больше и больше затрагивается вопрос качества научных исследований, что также позволит обеспечить новый уровень инновационной привлекательности региона.

ЛИТЕРАТУРА

1. Адамов Б.И. Реструктуризация хозяйственного комплекса региона и её влияние на развитие городов Донецкой области / Б.И. Ададмов, В.А. Кавыршина // Менеджер. – 2002. – №6 (22). – С. 4 –10.

2. Голова И.М., Суховей А.Ф. Инновационно-технологическое развитие промышленных регионов в условиях социально-экономической нестабильности // Экономика региона. 2015. № 1. С. 131-144.

3. Симонов А.Б., Рогачев А.Ф. Особенности системы инновационной деятельности на уровне региона как объекта управления // Друкеровский вестник. 2020. № 4. С. 224-231.

СИСТЕМНЫЙ ПОДХОД К ПРАВОВОМУ ОБЕСПЕЧЕНИЮ БЕЗОПАСНОСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

В.Д. Гаврилова
г. Волгоград, Россия
Волгоградский государственный университет

Автор считает, что полная стандартизация создания, функционирования и отключения искусственного интеллекта обеспечит безопасность человека. Проанализированы соответствующие ГОСТы, предложены уточнения к требованиям безопасности, понятие допустимого риска, порядок обучения системы, разделение программного кода на закрытую и открытую часть, программирование системы искусственного интеллекта на основании типичных ситуаций, алгоритмы разрешения которых базируются на традиционных ценностях РФ.

Ключевые слова: *искусственный интеллект, безопасность, стандарт, авторское право, традиционные ценности.*

LEGAL PROVISION OF ARTIFICIAL INTELLIGENCE SECURITY

V.D. Gavrilova
Volgograd, Russia
Volgograd State University

The author thinks that the complete standardization of the creation, functioning and disabling of AI will ensure human security. The author analyzes GOSTs, proposes clarifications to security requirements, the concept of acceptable risk, the procedure for training the system, the division of the program code into closed and open parts, programming of the AI based on typical situations whose resolution algorithms are based on Russian traditional values.

Keywords: *artificial intelligence, security, standard, copyright, traditional values.*

Развитие цифровых технологий привело к созданию, испытанию и внедрению искусственного интеллекта (далее – AI). Для человечества система AI представляется

выгодным инструментом, который позволит не просто упростить и улучшить жизнь, а иметь власть над информацией и иными ресурсами, что, несомненно, порождает соперничество между государствами за лидерство в данной сфере. С учетом изложенного, для Российской Федерации развитие сферы AI становится приоритетным направлением, поскольку именно это позволит обеспечить технологический суверенитет страны, а, значит, и национальную безопасность государства.

В этой связи интеграция технологий AI – неизбежный и непредсказуемый процесс, который требует от органов публичной власти, научного сообщества и общественности оперативной и адекватной реакции. Так, необходимо сделать систему AI максимально безопасной и удобной для человека. Данное направление значимо для национальной безопасности РФ, его реализация требует как технологическое обеспечение, так и правовое: 1) тщательное правовое регулирование общественных отношений, возникающих в связи с применением AI; 2) полная стандартизация процессов, непосредственно связанных с созданием, функционированием и отключением системы AI.

В современных реалиях технологии AI представлены как передовые и перспективные, об их развитии написано в национальных стратегиях, появляются соответствующие ГОСТы и ПНСТ, которые служат общим ориентиром для создания, функционирования и отключения AI. На данном этапе в Российской Федерации интеграция AI происходит на уровне действующих экспериментальных правовых режимов, поэтому регулирование данной области во многом связано с общими положениями и дорожными картами. При интеграции цифровых технологий в российскую правовую систему необходимы гарантии безопасности человека, усиленное внимание к ценностному компоненту.

Достаточно четко описывает функции AI Ю. А. Гаврилова, рассмотревшая данный аспект с точки зрения информационной безопасности РФ: ученый отмечает, что технологии AI направлены на автоматизацию рутинной или опасной деятельности человека, консультационную поддержку принятия решений и помощь в коммуникации людей [1, с.99].

Однако наряду с консультационной поддержкой AI необходимо принять во внимание непредсказуемость механизма на различных жизненных циклах. Как справедливо отмечают P. Cerka, J. Grigiene и G. Sirbikyte, AI может принимать решения независимо от воли своего разработчика, поскольку он обладает способностью к самообучению и накоплению опыта. Ученые приводят в пример робота Gaak, который в рамках эксперимента в центре Magna убегал от хищника и, стремясь выжить, сбежал из центра, хотя не был запрограммирован на это, выбежал на дорогу и спровоцировал ДТП, причинив вред другому человеку [2, с.382].

Данный пример демонстрирует не только способность роботов к самообучению, но и их стремление к самосохранению. Полагаем, что оно существует ввиду особенностей программного кода AI: так, систему программируют на прагматических объектах – польза, вред, риск, а потому в представлении AI он функционирует ради целесообразной деятельности. В этой связи робот будет стремиться к самосохранению, поскольку его уничтожение и отключение приравниваются к пресечению целесообразной деятельности и существования AI.

На наш взгляд, важно отметить, что технологии AI работают на основании принципа вознаграждения: например, если робот выполняет заданные программой действия, то к нему поступает больше энергии. Таким образом, при отключении не будет не только «полезности», но и вознаграждения, и механизм будет противиться отключению: система AI может создать собственные клоны – усовершенствованные версии AI, изменить свою физическую структуру или программное обеспечение (далее – ПО), обмануть оператора и тому подобное. Представляется логичным, что AI изначально программируют с элементами развития и совершенствования, поскольку это влияет на потенциал механизма и соответственно на выгоду от его использования. Кроме того, ранее было указано, что

система AI способна к самообучению. В этой связи уточним, что противодействие со стороны AI отключению и иные реакции системы на внешние «угрозы» порождают в ней стремление к еще большему развитию и самосовершенствованию.

Думается, что подобная характеристика данной цифровой технологии делает ее потенциально опасной. Так, AI может функционировать при диспетчеризации и контроле оператора, но при наличии клонов система усилит свои способности к саморегулированию, станет более самостоятельной или вовсе неуправляемой. Предполагаемые последствия видятся негативными. Считаем, AI должен использоваться исключительно как инструмент, система не может и не должна быть наравне с человеком или выше его.

Однако как этого достичь? При внедрении и дальнейшей реализации AI должна быть обеспечена индивидуальная и национальная безопасность, создание AI требует тщательной стандартизации. Затрагивая стандартизацию, следует рассмотреть в одной плоскости с ней вопрос потенциальной опасности AI, связанный со стремлениями системы к еще большему совершенствованию и развитию. Отметим, что мы не могли бы просто заблокировать аппаратное обеспечение системы и не указывать ей, как получить доступ к ее собственному машинному коду, поскольку способный к самообучению AI смог бы пройти данные препятствия (например, изменить среду выполнения) для достижения более высокой, «полезной» цели.

Кроме того, с точки зрения технологического обеспечения безопасности AI, как верно отмечает Z. Chen, проблемой является отсутствие баланса между размером статистической обучающей выборки и затратами времени на машинное обучение [3, с.735]. Уточним, что имеется в виду закономерность. Представляется очевидным, что для выгодной реализации AI нужно, чтобы обучение прошло быстро и качественно. Качество обучения связано со степенью освоения механизмом информации, чем больше исходных данных получает система AI, тем дольше проходит ее обучение. Кроме того, чем больше информации поступает на датчики к AI, тем больше он стремится к самостоятельному принятию решений: это закономерно, механизм способен к самообучению и стремится к самосовершенствованию.

Если AI с учетом огромного массива усвоенной информации начнет самостоятельно принимать решения, то при низком уровне безопасности и отсутствии надлежащей защиты данных создаются высокие риски утечки конфиденциальной информации и причинения вреда людям. Видится перспективным регламентировать разумные количественные ограничения применительно к исходным данным, чтобы система усваивала информацию последовательно и адаптировалась к ней при полном контроле со стороны человека.

В контексте изложенного представляется интересным ГОСТ Р 59276-2020 [4], раскрывающий общие вопросы доверия к системам AI. Примечательно, что понятие доверия к системе AI раскрывается через указание на способности данного механизма выполнять задачи с требуемым качеством (3.3). На наш взгляд, безопасность человека должна входить в понятие «требуемого качества» выполнения задач.

Рассматривая действующие применительно к AI ГОСТы и ПНСТ, видим, что аспект «безопасность» наиболее полно раскрывается в стандартах, связанных с применением AI в клинической медицине. Общие положения описываются в ГОСТе Р 59921.0-2022: в п.3.1.2. цитируется общее понимание безопасности как отсутствия недопустимого риска, в примечаниях дается конкретизация безопасности функционирования AI [5].

Отметим, что из п.3.1.2. следует, что уровни допустимого риска не установлены. Считаем, что выработка «универсального» уровня допустимых рисков затруднена в связи с конкретными клиническими ситуациями и разными информационными составляющими. Вместе с тем, считаем, допустимый риск возможно представить как ошибку при функционировании AI, последствия которой исправимы в разумные сроки. Уровень таких рисков динамичен до возникновения крайностей – непосредственной угрозы жизни, что определяется по показаниям здоровья пациента.

Анализируя примечания к п.3.1.2, выделим следующие требования к безопасности AI: 1) функционирование согласно определению изготовителя; 2) использование по назначению; 3) реализация с учетом уровня технических знаний пользователей, физического состояния потенциальных пациентов; 4) соблюдение конфиденциальности информации; 5) прозрачность алгоритмов; 6) бесперебойность; 7) отсутствие ошибок; 8) выполнение требований качества.

Думается, положение об отсутствии ошибок при работе AI не в полной мере согласуется с представленным понятием безопасности. Если последствия ошибок возможно исправить в разумные сроки, то это не помешает AI выполнить свою функцию и оставаться безопасным для человека. В этой связи следует конкретизировать риск, как сделано в определении безопасности: отсутствие недопустимых ошибок.

Кроме того, видятся дискуссионными требования о прозрачности алгоритмов и соблюдения конфиденциальности информации. На наш взгляд, написанный для AI программный код должен получать правовую охрану по аналогии с программой для ЭВМ, поскольку создатель такого кода творчески представил данные и команды в целях получения определенного результата. По общему правилу, именно создатель как автор имеет право на неприкосновенность и обнародование произведения (ст.1255 ГК РФ), может получить имущественную выгоду от заключения лицензионных договоров (ст.1236 ГК РФ). Если данный код написан как служебное произведение, исключительное право, по общему правилу, будет принадлежать работодателю. Так, прозрачность алгоритмов требует, чтобы обществу было известно, на основании чего AI принимает решения: соответствующий программный код должен быть доступен, что не согласуется с авторскими правами в случае, если автор против обнародования. Для разрешения описанного противоречия представляется возможным разделить программный код на две части: закрытую, охраняемую авторским правом, и открытую, подлежащую обнародованию по установленным законом правилам.

Потенциальная польза системы AI должна быть больше, чем потенциальный вред. Данный вопрос специфичен для телесного AI – робота, который способен моделировать эмоциональную составляющую: улыбаться, хмуриться, удивляться. Думается, наличие у роботов возможности имитировать такие психические состояния обусловлено тем, что данные объекты дают определенную оценку информации из окружающей среды, поступающей на вмонтированные в них датчики. В дальнейшем механизм принимает решения, руководствуясь данной информацией и произведенной оценкой, которая базируется на соответствующем программном коде.

Для обеспечения индивидуальной и национальной безопасности значим человеко-ориентированный подход (безопасность человека), который обеспечивается внедрением ценностного компонента в функционирование AI. Традиционные ценности РФ сложно представить как программный код и внести как исходные данные. Однако представляется возможным программирование реакций робота на основании типичных ситуаций, алгоритмы разрешения которых базируются на традиционных ценностях, что согласуется с национальной безопасностью России. Видится перспективным многоступенчатое и последовательное обучение телесного AI: необходимо начинать с таких представлений, как дружба, добро, жизнь, постепенно переходить к труду, приоритету духовного над материальным, высшим нравственным идеалам.

Таким образом, технологии AI работают на основании принципа вознаграждения: механизм осуществляет полезную, целесообразную деятельность и получает за это энергию. В этой связи способный к самообучению AI в любом случае будет стремиться к самосохранению и самосовершенствованию. Однако чем больше информации поступает на датчики AI, тем больше он стремится принимать решения самостоятельно и противодействовать собственному отключению: AI может создать собственные усовершенствованные версии, изменить свою физическую структуру или ПО, обмануть

оператора и тому подобное. Тем самым система усилит свои способности к саморегулированию, станет более самостоятельной или вовсе неуправляемой, что создает риски для индивидуальной и национальной безопасности.

Видится необходимым внести изменения в действующие стандарты, которые способствовали бы обеспечению безопасности AI.

Во-первых, следует регламентировать разумные количественные ограничения применительно к исходным данным, чтобы система AI усваивала информацию последовательно и адаптировалась к ней при полном контроле со стороны человека, что снизит риски неуправляемости AI.

Во-вторых, безопасность человека должна входить в понятие требуемого качества выполнения искусственным интеллектом задач, система AI может вызывать доверие только при максимальном удобстве и безопасности для человека.

В-третьих, допустимый риск возможно представить как ошибку при функционировании AI, последствия которой исправимы в разумные сроки. Важно уточнить требования безопасности AI: не отсутствие ошибок в целом (допустимый риск не помешает AI выполнить нужную функцию и не принесет вред человеку), а отсутствие именно недопустимых ошибок.

В-четвертых, следует разрешить выявленное автором противоречие между прозрачностью алгоритмов и соблюдением конфиденциальности информации. Представляется возможным разделить программный код на две части: закрытую, охраняемую авторским правом, и открытую, подлежащую при необходимости обнародованию по установленным законом правилам.

Необходимо правовое обеспечение безопасности человека при функционировании AI, которое должно сопровождаться человеко-ориентированным подходом, созданием AI на основании традиционных российских ценностей. К сожалению, их сложно представить как программный код, однако представляется возможным программирование реакций робота на основании типичных ситуаций, алгоритмы разрешения которых базируются на традиционных ценностях. Перспективным является многоступенчатое и последовательное обучение робота, начиная с таких представлений, как дружба, добро и жизнь, постепенно переходя к высшим нравственным идеалам.

ЛИТЕРАТУРА

1. Гаврилова Ю.А. Конституционализация информационной безопасности в российском праве: проблема совершенствования теоретической модели // Вестник Российского университета дружбы народов. Серия: Юридические науки. –2023. – Т. 27, № 1. – С. 97–116.

2. Cerka P., Grigiene J., Sirbikyte G. Liability for Damages Caused by Artificial Intelligence. Computer Law & Security Review. –2015. – № 31(3). – pp. 376–389.

3. Chen Z., Jia X., Zhang L., Yin G. Intelligent Security Image Classification on Small Sample Learning. In: Sun X., Zhang X., Xia Z., Bertino E. (eds). AI and Security. ICAIS. Lecture Notes in Computer Science. – 2021. – Vol. 12736. – Springer, Cham. pp. 726–737.

4. Об утверждении национального стандарта Российской Федерации: Приказ Росстандарта от 23.12.2020 N 1371-ст – Режим доступа: https://rst.gov.ru:8443/file-service/file/load/1682520287213 (дата обращения: 8.10.2023).

5. Об утверждении национального стандарта Российской Федерации: Приказ Росстандарта от 18.10.2022 N 1141-ст – Режим доступа: https://rst.gov.ru:8443/file-service/file/load/1682517235318 (дата обращения: 8.10.2023).