UDC: 34(042)(575.1)
ORCID: 0000-0003-0123-7811

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: ВОЗМОЖНА ЛИ ОТВЕТСТВЕННОСТЬ РОБОТОВ?

Бозаров Сардор Сохибжонович,
доктор философии по юридическим наукам (PhD),
заведующий кафедрой "Международное частное право"
Ташкентского государственного
юридического университета,
e-mail: sardor_bazarov1980@mail.ru

Аннотация. *В научной статье автором рассмотрены существующие научно-теоретические подходы к определению юридической ответственности искусственного интеллекта, в частности роботов, являющихся на сегодняшний день наиболее яркими представителями сферы искусственного интеллекта. Как автор отмечает, правила и механизмы установления правовой ответственности роботов со временем будут изменяться и совершенствоваться, однако необходима четкая регламентация совокупности социальных принципов – правил "жизни" роботов. В статье автором приведен перечень подобных принципов, среди которых он отмечает такие, как требования идентификации роботов, предупреждения, независимости искусственного интеллекта от инженерной инфраструктуры и человеческого фактора, полезности роботизации, наказуемости (удаления), корректируемости, проблема "черного ящика" и "выключения" роботов. В качестве основных проблем правовой регламентации искусственного интеллекта указаны автономность и независимость создания объектов искусственного интеллекта, которые сегодня производятся в промышленных масштабах во многих странах. При этом отсутствие единых принципов их создания и работы серьезным образом затрудняет как унификацию требований к ним, так и определение мер юридической ответственности. В заключении отмечается, что при дальнейшем увеличении использования роботов следует найти баланс между полезностью роботизации и безопасностью общества.*

Ключевые слова: *роботизация, искусственный интеллект, право, этика, мораль, юридическая ответственность, регламент, объекты интеллектуальной собственности.*

СУНЪИЙ ИНТЕЛЛЕКТ – РОБОТЛАРНИ ЖАВОБГАР ҚИЛИШ МУМКИНМИ?

Бозаров Сардор Сохибжонович,
Тошкент давлат юридик университети
Халқаро хусусий ҳуқуқ кафедраси мудири,
юридик фанлар бўйича фалсафа доктори (PhD)

Аннотация. *Муаллиф ўз илмий мақоласида сунъий интеллектнинг, хусусан, ҳозирги вақтда сунъий интеллект соҳасининг энг кўзга кўринган вакиллари сифатида шаклланиб бораётган роботларнинг ҳуқуқий жавобгарлигини аниқлашда мавжуд илмий ва назарий ёндашувларни ўрганиб чиқади. Муаллиф таъкидлаганидек, роботларнинг қонуний жавобгарлигини белгилаш қоидалари ва механизмлари вақт ўтиши билан ўзгариб, такомиллашиб боради, шу билан бирга, ижтимоий тамойиллар – роботларнинг "ҳаёт" қоидаларини аниқ тартибга солиш талаб этилади. Ушбу мақолада муаллиф қатор тамойилларни таҳлил этган, улар орасида роботларни идентификациялаш талаби, огоҳлантириш талаби, "қора қути муаммоси", сунъий интеллектни муҳандислик инфратузилмаси ва инсон омилидан мустақиллиги талаби, роботлаштиришнинг фойдалилиги талаби, роботларни жазолаш (йўқ қилиш) ҳамда уларни тузатиш каби қоидалар ва роботларни "ўчириб қўйиш" каби муаммолар кўриб чиқилган. Шунингдек, сунъий интеллектни ҳуқуқий тартибга солишнинг асосий муаммолари сифатида муаллиф ҳозирда кўпгина мамлакатларда саноат миқёсида ишлаб чиқарилаётган сунъий интеллект объектларини яратишнинг автономлиги ва мустақиллигини кўрсатиб ўтади. Бундан ташқари, сунъий интеллектни яратиш ва улар фаолиятининг ягона та-*

мойиллари мавжуд эмаслиги уларга қўйиладиган талабларни бирлаштиришни ҳам, ҳуқуқий жавобгарлик чораларини белгилашни ҳам жиддий равишда мураккаблаштиради. Роботлардан фойдаланиш ҳажмини янада ошириш билан робототехниканинг фойдали жиҳатлари ва жамият хавфсизлиги ўртасидаги мувозанатни қарор топтириш муҳим аҳамият касб этади.

Калит сўзлар: *робототехника, сунъий интеллект, ҳуқуқ, ахлоқ, ҳуқуқий жавобгарлик, регламент, интеллектуал мулк объектлари.*

ARTIFICIAL INTELLIGENCE – WHETHER RESPONSIBILITY OF ROBOTS POSSIBLE?

Bozarov Sardor Sohibjonovich,
The head of International Private Law Department of
Tashkent State University of Law,
doctor philosophy in law (PhD)

Abstract. *In this scientific article, the author examines the existing scientific and theoretical approaches to determine the legal responsibility of artificial intelligence, in particular, robots, which are currently the most prominent representatives of the field of artificial intelligence. As the author notes, the rules and mechanisms for establishing the legal responsibility of robots will change and improve over time, however, a clear regulation of the set of social principles - the rules of the "life" of robots, is required. In the article, the author provides a list of similar principles such as the requirement for identification of robots, the requirement for a warning, the "black box problem", the requirement for the independence of artificial intelligence from the engineering infrastructure and human factor, the requirement for the usefulness of robotization, the requirement for punishment (removal), the requirement of correct ability and the problem of "turning off" robots. As the main problems of the legal regulation of artificial intelligence, the author points out the autonomy and independence of the creation of artificial intelligence objects, which are now being produced on an industrial scale in many countries. At the same time, the lack of single principles for their creation and operation seriously complicates both the unification of requirements for them and the determination of measures of legal responsibility. In conclusion, it is noted that with a further increase in the use of robots, a balance should be found between the usefulness of robotization and the safety of society.*

Keywords: *robotization, artificial intelligence, law, ethics, morality, legal responsibility, regulations, intellectual property.*

Введение

Из поколения в поколение человечество передает ценности цивилизации. Мы делаем это в надежде, что следующее поколение будет придерживаться этих принципов, а в дальнейшем, когда они будут достаточно зрелы, будут развивать свои собственные ценности и, в свою очередь, передавать их своим детям. Эти основные нормы есть базовая мораль человеческого общежития.

Однако сегодня, возможно, впервые в истории человечества, мы сталкиваемся с искусственными сущностями – роботами и иными автоматами, способными принимать сложные решения и следовать расширенным правилам. Каким ценностям мы должны их научить? [1]. Чтобы ответить на этот вопрос, нам нужно уточнить еще два: один связан с моральными принципами: "Какие нормы выбрать для передачи?", второй – технический: "Каким образом мы можем передать эти нормы-ценности искусственному интеллекту (ИИ)?".

Материалы и методы

Методологической базой исследования является диалектический метод с его законами и основными принципами: принципом системности, принципом развития, принципом единства полярностей, принципом детерминации и другими. Это обусловлено тем, что искусственный интеллект как сложное явление культуры очень часто не располагает к формулированию однозначных определений и трактовок, а само явление богаче этих субъективных интерпретаций.

Результаты исследования

Данное исследование направлено на выработку и предложение некоторых положений, которые могут сформировать в будущем

минимум конструкций – принципов деятельности ИИ. Тем не менее предлагаемые ниже положения носят ориентировочный, а не финальный характер.

Несомненно, правила и механизмы их достижения со временем будут изменяться и совершенствоваться. Но так же, как и совокупность социальных принципов, из которых состоит человеческая мораль, должна быть зарождена совокупность правил "жизни" роботов.

Требование идентификации

Этот принцип подразумевает, что организация должна сообщать, есть ли у нее возможности использования искусственного интеллекта. Тоби Уолш предлагает следующее правило [2]: автономная система должна быть спроектирована так, чтобы в ней было маловероятно ошибиться в ее идентификации как искусственного интеллекта, то есть она сама как автономная система должна идентифицировать себя в начале любого взаимодействия с другим агентом [3]. По мнению Уолша, это положение может быть аналогично требованиям, предъявляемым к игрушечным пистолетам, которые могут быть идентифицированы по яркому колпачку на конце, чтобы понять, что это не настоящее оружие [4].

Орен Эциони, генеральный директор исследовательского института AI, предложил несколько иную формулировку: "... система ИИ должна ясно показывать себя, что она не является человеком" [5]. Проблема с предложением Эциони в том, что не весь ИИ напоминает людей или имитирует их, скорее, большинство из них вообще не похожи на людей.

Требование идентификации полезно по нескольким причинам: во-первых, они играют важную роль в обеспечении или поддержке функций, уникальных для ИИ.

Во-вторых, учитывая, что ИИ в определенных условиях действует иначе, чем люди, имея некоторое представление о том, является ли объект человеком или ИИ, можно будет более предсказуемо определять его поведение, повышая как эффективность, так и безопасность ИИ [6]. К примеру, если человек выбегает перед автомобилем, движущимся со скоростью 70 миль в час, средний человек не сможет отреагировать достаточно быстро, чтобы уклониться от аварии, в то время как система ИИ вполне может это сделать [7]. Иными словами, в ситуациях, которые требуют "здравого смысла", ИИ (по крайней мере, на данном этапе развития), вероятно, будет значительно уступать человеческому мышлению. ИИ-водитель или машины-роботы могут ездить по автомагистрали, но оценивать сложные или необычные элементы, такие как неожиданные дорожные работы или марш протеста на улицах, им будет сложно.

В-третьих, идентификация ИИ может потребоваться для управления определенными видами деятельности, могущими повлиять на справедливость и честность. Игрок в покер хочет знать, что он играет против обычного человека, когда ставит 5 000 долларов, в отличие от потенциально непревзойденной системы игрока – ИИ [8].

В-четвертых, идентификация позволит людям узнать источник виртуального общения. Так, в отчете о злонамеренном использовании ИИ за 2018 год указана серьезная проблема: "... существует возможность использования ИИ для автоматизации задач, связанных с ... убеждением (например, создание целевой пропаганды) и обманом (например, манипулирование видео), которые могут усилить угрозы, связанные с вторжением в частную и социальную жизнь" [9].

Анонимность социальных сетей может позволить небольшому количеству людей оказывать гораздо большее влияние, чем если бы они действовали лично, особенно, если они контролируют сеть ботов, распространяющих их контент и/или взаимодействие с пользователями-людьми. Хотя требование идентификации ИИ не может запретить злонамеренное использование, это может усложнить использование социальных сетей, сводя к минимуму возможности для гнусных инсинуаций и использования ИИ.

Требование предупреждения

Из-за присущих им опасностей некоторые продукты и услуги могут предлагаться на законных основаниях лишь при наличии соответствующих предупреждений. Пользователи тяжелых машин, как правило, строго предупреждаются о необходимости отстранения от

работ, если водитель находится под воздействием алкоголя или наркотиков. Часто можно увидеть продукты с маркировкой “Внимание! Может содержать орехи”. В один прекрасный день мы можем увидеть на продукции знак: “Внимание, может содержать ИИ!”.

Учитывая разнообразие систем и типов ИИ, маловероятно наличие единого технологического решения для реализации требования об идентификации. Следовательно, положение об идентификации для ИИ должно быть сформулировано в общих чертах, оставляя это на усмотрение отдельных конструкторов и изготовителей ИИ. Например, Комитет парламента Нового Южного Уэльса для проекта “Беспилотные автомобили и безопасность дорожного движения” указал: “публичная идентификация автоматизированных транспортных средств требует, чтобы они визуально отличались от других пользователей дорог, особенно на этапе испытаний и тестирования” [10].

Периодические проверки и тестирование могут использоваться для определения сущности и природы ИИ. Однако тестирование и режимы проверки безопасности могут использоваться также при транспортировке и поставке многих товаров и услуг. Подобные мероприятия в настоящее время применяются правоохранительными органами для отслеживания вредоносного ПО и взломов (и, без сомнения, будут разработаны новые), они также могут быть использованы для мониторинга правильной маркировки ИИ.

Требование идентификации не было бы таким полезным и нужным, если бы субъекты, не относящиеся к ИИ, не хотели бы маскироваться под личиной ИИ. Ложные срабатывания снизят доверие к системам идентификации, будет подрываться ее полезность как сигнального механизма. По этой причине любое требование идентификации должно действовать в обоих направлениях: запрещать организациям, не относящимся к производителям ИИ, отмечать продукцию как содержащие ИИ, аналогично производитель продуктов питания может столкнуться с штрафами, если он описывает продукт как “подходящий для вегетарианцев”, в то время как он содержит продукты животного происхождения.

Требование четкого разъяснения

Данное требование подразумевает, чтобы возможности ИИ были понятны людям. Это может включать требование о предоставлении информации об общем процессе принятия решений ИИ (прозрачность) и/или о том, что конкретные решения рационализируются после того, как они были приняты (индивидуализированное разъяснение).

Для разъяснения ИИ обычно предлагается два основных пути: инструментальное и внутреннее (ценностное). Инструментальное фокусируется на разъяснении ИИ, как инструменте, улучшающим жизнь и исправляющем ошибки людей. Внутренний (ценностный) подход фокусируется на правах всех нуждающихся в ИИ людей. Эндрю Селбст и Джулия Паулз объяснили, что “внутренняя ценность разъяснений определяет потребность человека в свободе воли и контроле машин” [11].

Агентство перспективных исследовательских проектов Министерства обороны США (DARPA) [12] имеет одну из самых передовых и известных программ в этой сфере – XAI [13]. DARPA дает как инструментальные, так и внутренние (ценностные) разъяснения своего проекта: “Эффективность систем [ИИ] будет ограничена способностью машины объяснять свои мысли и действия пользователям. Понятный ИИ будет иметь важное значение, если пользователи будут понимать, доверять и эффективно управлять этим новым поколением партнеров с искусственным интеллектом [14].

Проблема черного ящика

Основная трудность с применением требования разъяснения состоит в том, что многие системы ИИ работают как “черные ящики”: они могут быть искусными в выполнении задач, но даже их собственные конструкторы могут быть не в состоянии объяснить, каким образом внутренний процесс привел к определенному результату [15].

Как отмечают Брайс Гудман и Сет Флаксман, многие обучаемые модели машин не созданы с учетом возможности интерпретации человеком во время беспокойства, и сомнительно, что весь их спектр эффектов может быть достигнут, если прозрачность будет встроена в процесс. Конечно, существует компромисс между репрезентативной способностью мо-

дели и ее интерпретируемостью, начиная от линейных моделей (которые могут только представлять простые отношения, их легко интерпретировать) до непараметрических моделей ИИ (которые могут представлять богатый набор функций, но их будет трудно интерпретировать). Нейронные сети, особенно с развитием глубокого обучения, станут, пожалуй, самой большой проблемой – как объяснить функции, происходящие в многослойной нейронной сети ИИ со сложной архитектурой?

Исследователь Дженна Баррелл, работающая в информационной школе Калифорнийского университета в Беркли пишет, что в машинном обучении существует "непрозрачность, связанная с несоответствием между математической оптимизацией больших объемов машинного обучения и требованиями человеческого масштаба смысловой интерпретации" [16]. Трудность усугубляется тем, что системы машинного обучения обновляются по мере работы в процессе переоценки их внутренних узлов, чтобы каждый раз улучшать результаты. В результате процесс, который привел к одному результату, может не совпадать с полученным позднее.

Один из методов разъяснения, обеспечивающий понимание процесса индивидуализированных решения, – это научить систему ИИ семантическим ассоциациям в процессе изготовления. ИИ можно научить выполнять главную задачу, например определить, отображает ли видео сцену свадьбы, а также уточнить второстепенные задачи – связать события в видео с определенной ситуацией [17]. Так, У. Эхсан, Б. Харрисон, Л. Чан и М. Ридл разработали технику, которую они описывают как "рационализация ИИ, подход к генерации объяснений поведения автономной системы с точки зрения поведения человека" [18]. Система просит людей объяснять их определенные действия, которые они предпринимают. Ассоциации между методами, принятыми ИИ, и человеком записываются таким образом, чтобы создать набор помеченных функций (действий).

Идя другим путем, специалист по анализу данных Д. Уайтхоук выделяет три общие возможности, необходимые для прозрачности ИИ: происхождение данных (знание источника всех данных); воспроизводимость (возможность воссоздать запрограммированный результат); управление версиями данных (сохранение копий ИИ в определенных государствах с целью записи, какой ИИ действует неверно). Он предлагает сделать эти три "стандарта в области науки о данных" обязательными, в качестве инструментов интеграции этих характеристик в рабочие процессы.

Требование независимости ИИ от инженерной инфраструктуры и человеческого фактора

Независимость от инфраструктуры означает, что инструменты ИИ должны быть развернуты и работать в существующей инфраструктуре – локально, в облаке или виртуально. Было бы совершенно непрактично вносить каждый раз изменения в рабочий процесс ИИ, если это не может быть масштабировано до производственных требований.

Следующий вопрос очень актуален. Как известно, юридические правила работают таким образом, что концентрируются в первую очередь на действиях, а не на мыслях. В целом, теоретически ИИ должен обеспечивать полную беспристрастность, свободную от человеческого фактора, ошибок и предрассудков. Однако во многих случаях этого не происходит.

Например, научные статьи изобилуют примерами очевидных случаев предвзятости на конкурсах красоты, проводимых с использованием ИИ, где побеждает только европеоид, можно указать программное обеспечение правоохранительных органов, которое использует расу для прогноза совершения людьми преступлений в будущем. То есть ИИ, похоже, разделяют многие из тех же проблем, что и люди.

Возникают три вопроса: что такое предвзятость ИИ, почему она возникает и что с этим делать?

Предвзятость часто связана с решениями, которые считаются "несправедливыми" или "справедливыми" по отношению к отдельным лицам или группам людей. Проблема с решением таких моральных понятий в определении предвзятости заключается в том, что они тоже неопределенны и расплывчаты. Понятие результата или процесса "несправедливости" очень субъективно. Некоторые люди считают положительную дискриминацию неспра-

ведливой, в то время как другие считают это справедливым ответом на общественные дисбалансы. Если и должно быть правило, касающееся предвзятости ИИ, то предпочтительнее использовать механизм, сводящий к минимуму роль личного мнения.

Непосредственным источником предвзятости ИИ часто являются данные, вводимые в систему. Машинное обучение – в настоящее время доминирующая форма ИИ – признает шаблоны в данных, а затем принимает решения на основе распознавания уже заложенных шаблонов. Если входные данные каким-то образом искажены, то вероятность заключается в том, что созданные шаблоны также будут ошибочными. Предвзятость, возникающую из-за таких данных, можно резюмировать фразой: "ты – то, что у тебя внутри".

Искаженные наборы данных возникают, когда информации полно, они доступны для представления достаточной картины соответствующей среды, но люди-операторы совершают нерепрезентативную выборку. Этот феномен не уникален для ИИ. В области статистики "системная ошибка выборки" означает ошибки в оценке, которые возникают, когда некоторые элементы набора данных будут отобраны с большей вероятностью, чем другие данные.

Иногда ошибки данных возникают не из-за выбора конкретных данных, установленных людьми, а, скорее, потому, что вся совокупность доступных данных ошибочна. Это связано с тем, что некоторые предубеждения могут быть глубоко укоренившимися в обществе и для исправления данных необходимы тщательная сортировка и даже изменение модели ИИ. Интернет – это не единственный источник массового набора данных, которые могут быть подвержены аналогичным проблемам врожденной предвзятости. Возможно, что такие библиотеки программного обеспечения для машинного обучения, как Google Tensor Flow, а также Amazon и Microsoft Gluon, могут иметь похожие скрытые дефекты.

Иногда вся совокупность данных доступных в машиночитаемом формате недостаточно детальна для получения объективных результатов. Например, ИИ могут попросить определить, какие кандидаты лучше всего подходят для работы в качестве разнорабочих на стройплощадке на основе данных. Если единственные данные, доступные ИИ, – возраст и пол, то наиболее вероятно, что ИИ выберет более молодых мужчин для работы. Однако пол или возраст заявителей не относятся к их способностям. Скорее, важны ключевые навыки, а не только сила и ловкость. Это может быть связано с возрастом и полом (особенно в том, что касается физической силы). Но важно не путать корреляцию с причинно-следственной связью: обе эти точки данных – просто шифры. Если ИИ обучался на основе данных многих способностей, то это приведет к выбору, который все еще может быть в пользу молодых мужчин, но, по крайней мере, это будет сделано таким образом, чтобы свести к минимуму предвзятость.

Выбор данных – это не только наука, но и искусство. Многое зависит от набора образцов, используемых социологами при опросе населения. Точно так же мы должны быть осторожны при загрузке данных в системы ИИ, чтобы гарантировать, что используемые данные являются надлежащим образом репрезентативными. Косвенный способ обеспечить лучший выбор данных – это не просто просить программистов быть "более чувствительными", но и стремиться расширить демографию программистов, включить представителей меньшинств и женщин. Обеспечение разнообразия среди программистов – это не только вопрос пола и расы, это может также потребовать разнообразия в национальном происхождении, религии и др. Тем не менее было бы неправильно полагать, что только разнородная группа программистов может создавать ИИ, который дает объективные результаты, или такие разнообразные программисты всегда будут создавать непредвзятый ИИ. Разнообразие помогает свести к минимуму предвзятость, но этого недостаточно.

Требования полезности роботизации

Из всех применений ИИ его применение в качестве автономного оружия (или робота-убийцы), пожалуй, самый противоречивый. Как только стала ближе к реальности перспектива оружия, способного самостоятельно выбирать и вести огонь по целям, в

2013 году была начата международная кампания по запрету роботов-убийц [19]. В августе 2017 году 116 экспертов и основателей ИИ-компаний написали открытое письмо, выражая серьезную озабоченность по этому поводу [20].

Несмотря на справедливые опасения, есть веские аргументы, что полный запрет ИИ может быть контрпродуктивным – не только в отношении автономного оружия, но и в любой области. Утверждается, что существует принципиальное решение вопроса о том, когда и как использовать ИИ в спорных областях – это так называемый *телеологический принцип использования ИИ.*

Начнем с того, каких ценностей мы стремимся придерживаться в данной деятельности. Если (и только тогда) ИИ сможет последовательно поддерживать эти ценности, но при этом, очевидно, превосходит людей, тогда использование ИИ должно быть разрешено.

Одним из примеров того, как ИИ превосходит людей в важной задаче, является способность распознавать определенные виды рака. В то время как врачам могут потребоваться несколько минут на анализ сканирования каждой части тела, ИИ может сделать это за миллисекунды, в некоторых случаях с явно более высокой точностью, чем у экспертов [21]. Телеологический принцип нельзя использовать абстрактно; даже там, где он будет удовлетворен, директивным органам необходимо будет помнить об общественных взглядах на приемлемость использования ИИ - или вообще любой нечеловеческой технологии для выполнения соответствующей задачи. Вопрос о том, могут ли люди принять конкретную технологию, более широкий аспект социальной и политической легитимности, результаты которой могут различаться в разных обществах.

Возвращаясь к примеру с автономным оружием, в международном гуманитарном праве широко признаны два руководящих принципа: соразмерность – требование не причинять вреда, чем это необходимо для достижения законной цели, и различие – различать комбатантов и гражданских лиц [22].

Сторонники запрета автономного оружия часто указывают на то, что многие страны уже принимают какое-то оружие, возможности запрета которого сильно ограничены [23]. Примерами являются слепящие лазеры [24] и использование наземных мин [25]. Однако главное различие между автономным оружием и технологиями, которые до сих пор запрещены, в том, что запрещенные технологии обычно не соответствуют основным законам ведения войны. После того, как наземная мина будет установлена, она взорвется независимо от того, кто наступит на нее, будь то гражданское лицо или комбатант. Ядовитый газ не делает различий в отношении того, кого он отравляет. Ослепляющим лазерам нельзя сказать, чтобы они жалели глаза мирных жителей.

Напротив, если его развитие должным образом регулируется, ИИ вполне может превосходить людей в способности различать мирных жителей и комбатантов при проведении сложных операций. Некоторые скептически относятся к тому, что это когда-либо произойдет, но история показывает, что пессимизм неуместен. ИИ уже работает лучше, чем люди в некоторых тестах на распознавание лиц [26], что является ключевым навыком в выборе целевой аудитории. Более того, системы ИИ не устают, они не злы или мстительны, какими могут быть солдаты-люди. Роботы не насилуют, не грабят и не убивают.

Тем не менее есть один важный аргумент против автономного оружия – оно может быть взломано или неисправно [27]. Это правда, но те же аргументы могут быть применены к любой из десятков тысяч единиц техники, используемой в современной войне: глобальной системе позиционирования, используемой военными бомбардировщиками для точного обнаружения целей, рулевому управлению атомных подводных лодок.

Этот пункт выходит за рамки военной сферы и включает в себя контроль над коммунальными объектами, такими как плотины, атомные электростанции и транспортные сети, многие из которых сильно зависят от технологий. Всякий раз, когда потенциально опасные действия выполняются, важно убедиться, насколько это возможно, что задействованные компьютерные системы безопасны и защищены от внешних атак или неисправности.

Требования наказуемости (удаления)

Как известно, в системе человеческого правосудия высшей мерой наказания является смертная казнь. Эквивалент ИИ – это кнопка выключения или аварийный выключатель: механизм выключения ИИ либо по решению человека, либо автоматически по заданному алгоритму. Иногда это называют "большой красной кнопкой", ссылаясь на специальные выключатели, часто встречающиеся на мощных машинах.

Хотя ИИ могут действовать иначе, чем человеческая психология, их возможные мотивации остаются актуальными. Важно отметить, что широко признано, что справедливая система может признавать права человека, но сохранять систему наказаний, включая ограничение таких прав. Некоторые права, такие как свобода от пыток, рассматриваются как абсолютные (по крайней мере, во многих странах). Однако другие права, такие как свобода, должны быть сбалансированы относительно общественных целей: лишение свободы преступников не умаляет общее мнение о том, что все граждане должны иметь право вести свою жизнь без вмешательства.

Возмездие – это психологический феномен, который применим ко всем человеческим обществам [28]. Функции воздаяния на двух уровнях: самому преступнику и по отношению к остальному населению. Эта двойная роль возмездия описывается как "решительное осуждение обществом преступления" [29]. Пожалуй, самый известный пример – это ветхозаветный список наказаний: "око за око, зуб за зуб, рука за руку, ступня за ступню". Философы права описывают разницу между человеческими ожиданиями, что кто-то будет нести ответственность за вред от ИИ, и текущей неспособностью наказания ИИ как "брешь в возмездии". Они утверждают:

1. Если агент несет причинную ответственность за морально вредный результат, люди будут пытаться возложить вину на этого агента (или на какого-либо другого агента, который считается ответственным за этого агента).

2. Повышенная роботизация означает, что роботы-агенты могут причинно нести ответственность за все более и более вредное с точки зрения морали и права последствия. Таким образом, повышенная роботизация означает, что люди будут стремиться возложить ответственность на роботов (или на других связанных агентов, которые, как считается, могут нести ответственность за этих роботов, например, на производителя/программиста).

3. Но ни роботы, ни связанные с ними агенты (производители/программисты) не будут подходящими объектами возмездия.

4. Если нет подходящих субъектов карательного воздействия, люди все же будут искать подобные объекты или предметы, тогда будет брешь в возмездии. Следовательно, усиление роботизации приведет к бреши в возмездии [30].

Возможно, однажды будет создан ИИ, который почувствует моральную вину так же, как человек. Но чтобы оправдать наказание, для возмездия это не обязательно. Поскольку возмездие преследует двойную цель, оно может быть эффективно, даже если преступник сам не испытывает моральной вины. В качестве примера – внешняя роль возмездия сохраняется: если есть общественное требование о наказании кого-либо или чего-либо, нет человека, которого можно было бы назвать ответственным, прекращая (выключая) ИИ может заполнить пробел, тем самым поддерживая доверие к системе правосудия в целом.

В этом свете использование аварийного выключателя в качестве механизма возмездия исполняет основное желание – "справедливость свершилась" [31].

Хотя "аварийный выключатель" может показаться драматичным, эта фраза обычно используется для описания механизмов временного отключения ИИ, а не его полного уничтожения. В качестве прагматичного ответа на ошибки ИИ, которые вызывают конкретные случаи вредоносного поведения, временное отключение полезно тем, что позволяет третьим лицам (будь то люди или другой ИИ) проверить неисправность, диагностировать и устранить причину проблемы. Это соответствует одной из целей наказания в системах правосудия человека – исправление личности. Многие системы правосудия, такие как тюрьмы, предназначены, по крайней мере, частично в качестве возможности предотвращения

рецидивизма путем оснащения преступников новыми навыками, улучшенным моральным компасом.

Люди – не единственные существа, поддающиеся карательному сдерживанию, – животные могут быть обучены действовать определенным образом, если их отклонение от этого действия наказывается.

Некоторые формы ИИ уже полагаются на тип обучения, который чем-то напоминает то, как мы учим животных или маленьких детей. В обучении с подкреплением используется функция вознаграждения для поощрения "хорошего" поведения со стороны ИИ, а также может включаться формы наказания, чтобы воспрепятствовать "плохому" поведению [32]. Почему наличие аварийного выключателя удерживает "плохое" поведение ИИ? Причины, по которым ИИ делает это, просты. Если у ИИ есть конкретная задача или цель – от получения прибыли на фондовом рынке до наведения порядка в комнате – ИИ не сможет достичь этой цели, если он будет отключен или удален. Следовательно, при прочих равных условиях ИИ будет иметь мотивацию избегать поведения, которое, как он осознает, приведет к его выключению [33].

Наконец, наличие аварийного выключателя выполняет ту же роль, что и человеческое наказание, которое сдерживает преступника на практическом уровне от нанесения такого же вреда более широкому обществу. Приговоры к лишению свободы ограничивают доступ преступника к обществу. Смертный приговор в странах, где он существует, идет еще дальше, прекращая жизнь рассматриваемого преступника. Те же мотивы применимы и к ИИ. Самые надежные аварийные выключатели будет сочетать предупредительный подход автоматического отключения, если происходят предопределенные события с произвольным отключением человеком, чтобы обеспечить гибкость в случае непредвиденного события или непредсказуемого поведения, делающего дальнейшую работу ИИ вредной.

Корректируемость и проблема "выключения"

В отличие от других технологий, упомянутых выше, аварийный выключатель для ИИ – это не просто вставить автоматический выключатель или добавить большую красную кнопку.

Почему ИИ может сопротивляться аварийному выключению? Исследователи Нейт Соарес и Фалленштейн из Исследовательского института машинного интеллекта объясняют: чтобы исправить современную систему ИИ, нужно просто закрыть систему, изменив его исходный код. Изменение системы умнее человека может оказаться более трудным: система, обладающая суперинтеллектом, может приобретать новое оборудование, изменять его программное обеспечение, создавать субагентов и принимать другие действия, которые оставят первоначальным программистам лишь небольшую возможность контроля над этим ИИ. Это особенно верно, если у агента есть стимулы сопротивляться модификации или выключению [34].

Иногда это называют "проблемой коррелируемости" [35]. Так, аналогично приговоренному к смертной казни, не принявшему такой исход добровольно, ИИ могут обладать инстинктом самосохранения, который заставляет его сопротивляться таким мерам.

Ник Бостром утверждает необходимость контрмер [36]. Такие контрмеры, возможно, потребуются, чтобы избежать экстремального риска, связанного с суперинтеллектом ИИ, но они также важны задолго до того, как ИИ станет всемогущим.

Сложности возникают, если существует несоответствие между полезностью, которую ИИ ожидает достичь от выполнения данной задачи, и полезности, которую ИИ рассчитывает получить от отключения. Предполагая, что ИИ в этом вопросе – рациональный агент, который пытается максимизировать ожидаемую выгоду в соответствии с некоторой функцией полезности, и если ИИ ставится задача более высокой степени полезности, чем при выключении, то при прочих равных условиях ИИ будет стремиться избежать отключения, возможно, даже отключив его надзирателей-людей. Однако, если для аварийного выключателя задано такое же или большее значение оценки полезности выполнения основной задачи, тогда ИИ может решить активировать аварийный выключатель, чтобы добиться максимальной полезности в минимальное количество времени. Эта суицидальная тенденция известна как "проблема закрытого выхода" [37].

Выводы

Итак, как следует из анализа существующей практики роботизации и развития ИИ, найти баланс между полезностью роботизации и безопасностью общества представляется весьма сложной задачей. В этой связи человечеству предстоит решить весьма трудоемкие научные, правовые и этические проблемы.

В конечном счете, искусственный интеллект – это несомненный прогресс цивилизации, однако имеющиеся неясности, потенциальные опасности, связанные с искусственным интеллектом, сдерживают ученых и исследователей в позитивной оценке будущего развития ИИ. Что нас ждет в будущем – роботы-помощники или заведомо проигранная конкуренция суперинтеллекту? К сожалению, сейчас на этот вопрос однозначно ответить не представляется возможным.

REFERENCES

1. Conn A. On the issue of value, alignment see, for example. How do we align Artificial Intelligence with Human values?" Future of Life Institute, 2017, February 3. Available at: https://futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/?cn-reloaded=1/ (accessed 01.06.2018).

2. The UK Locomotive Act, 1865, p. 3.

3. Walsh T. Android Dreams. London, Hurst & Company, 2017, p. 111.

4. Walsh T. Turing's Red Flag. Communications of the ACM, 2016, July, vol. 59, no. 7, pp. 34–37.

5. Etzioni O. How to Regulate Artificial Intelligence. New York Times, 2017, September 1. Available at: https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulationsrules.html/ (accessed01.06.2018).

6. Gershgorn D. An AI-Powered Design Trick Could Help Prevent Accidents like Uber's Self-Driving Car Crash. Quartz, 2018, March 30. Available at: https://qz.com/1241119/accidents-like-ubers-self-driving-car-crash-could-be-prevented-with-this-ai-powered-designtrick/ (accessed 01.06.2018).

7. The discussion of the AI2 Reasoning Challenge in Will Knight. AI Assistants Say Dumb Things, and We're About to Find Out Why, MIT Technology Review, 2018, March. Available at: https://www.technologyreview.com/s/610521/ai-assistants-dont-have-the-common-senseto-avoid-talking-gibberish/ (accessed 01.06.2018).

8. AI2 Reasoning Challenge Leaderboard, AI2 Website. Available at: http://data.allenai.org/arc/ (accessed 01.06.2018).

9. The proficiency of AI poker players, see Byron Spice, Carnegie Mellon Artificial Intelligence Beats Top Poker Pros, Carnegie Mellon University Website. Available at: https://www.cmu.edu/news/stories/archives/2017/january/AI-beats-pokerpros.html/ (accessed 01.06.2018).

10. Brundage et al. The malicious use of artificial intelligence: forecasting, prevention, and mitigation. 2018, February. Available at: https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads /1c6q2kc4v _50335.pdf/ (accessed 01.06.2018).

11. Driverless vehicles and road safety in New South Wales. Staysafe, Joint Standing Committee on Road Safety, 2016, September 22, p. 2. Available at: https://www.parliament.nsw.gov.au/committees/DBAssets / InquiryReport /ReportAcrobat/6075/Report%20-%20Driverless%20Vehicles%20and%20Road%20Safety%20in%20NSW.pdf/ (accessed 01.06.2018).

12. Selbst A.D., Powles J. Meaningful Information and the Right to Explanation. International Data Privacy Law, vol. 7, no. 4, 2017, November 1, pp. 233-242. DOI: https://doi.org/10.1093/idpl/ipx022/ (accessed 01.06.2018).

13. DARPA Website. Available at: https://www.darpa.mil/ (accessed 01.06.2018).

14. Gunning D. Explainable Artificial Intelligence (XAI). DARPA Website. Available at: https:// www.darpa.mil/program /explainable-artificial-intelligence/ (accessed 01.06.2018).

15. Gunning D. DARPA XAI Presentation. DARPA. Available at: https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf/ (accessed 01.06.2018).

16. Knight W. The Dark Secret at the Heart of AI. MIT Technology Review, 2017, April 11. Available at: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/(accessed01.06.2018).

17. Cheng H. et al. Multimedia Event Detection and Recounting. SRI-Sarnoff AURORA at TRECVID, 2014. Available at: http://www-nlpir.nist.gov/projects/tvpubs/tv14. papers/sri_aurora.pdf/ (accessed 01.06.2018).

18. Ehsan U., Harrison B., Chan L., Riedl M. Rationalization: a neural machine translation approach to generating natural language explanations. 1702.07826v2 [cs.AI], 2019, Dec 2. Available at: https://arxiv.org/pdf/1702.07826.pdf/ (accessed 01.06.2018).

19. Winfield A. Artificial intelligence will not turn into a Frankenstein's Monster. The Guardian, 2014, August 10. Available at: https://www.theguardian.com/technology/2014/aug/10/artificial-intelligence-will-not-become-a-frankensteins-monster-ian-winfield/ (accessed 01.06.2018).

20. Bostrom N. Superintelligence. Oxford, Oxford University Press, 2014, pp. 124-125.

21. Musk E. Artificial intelligence is our biggest existential threat. The Guardian, 2014, October 27. Available at: https://www.theguardian.com/ technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat/ (accessed 01.06.2018).

22. Open Letter. Future of Life Institute. Available at: https://futureoflife.org/ai-open-letter/ (accessed 01.06.2018).

23. Hern A. Stephen Hawking: ai will be 'either best or worst thing' for humanity. The Guardian, 2016, October 19. Available at: https://www.theguardian.com/science/2016/oct/19/ Stephen-hawking-ai-best-or-worst-thing-for-humanity-Cambridge/ (accessed 01.06.2018).

24. The Locomotives on Highways Act 1861, the Locomotive Act 1865 and the Highways and Locomotives (Amendment) Act 1878 (all UK legislation).

25. Jones S.E. Against technology: from the luddites to neoluddism. London, Routledge, 2013.

26. Domingos P. The master algorithm: how the quest for the ultimate learning machine will remake our world. New York, Allen Lane, 2015, p. 286.

27. Lewis-Kraus G. The Great A.I. Awakening. The New York Times Magazine, 2016, December 14. Available at: https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html/ (accessed 01.06.2018).

28. Kyoto Protocol to the United Nations Framework Convention on Climate Change, 1997.

29. Paris Climate Agreement, 2016.

30. Dobbs R., Manyika J., Wetzel J. No ordinary disruption: the four global forces breaking all the trends. McKinsey Global Institute, 2015, April. Available at: https://www.mckinsey.com/mgi/no-ordinary-disruption/ (accessed 01.06.2018).

31. Postema G. Coordination and convention at the foundations of law. Journal of Legal Studies, 1982, vol. 165, no. 11, p. 172.

32. Website of the British Medical Association. Available at: https://www.bma.org.uk/advice/career/going-abroad/working-abroad/usa/ (accessed 01.06.2018).

33. Directive 2005/36/EC of the European Parliament and Council of 2005, September 7.

34. The Nazi Doctors and the Nuremberg Code: Human rights in human experimentation.

35. Ryan M. Doctors and the State in the Soviet Union. New York, Palgrave Macmillan, 1990, p. 131.

36. Lewis A. Abroad at Home; A Question of Confidence. New York Times, 1990, September 19. Available at: http://www.nytimes.com/1985/09/19/opinion/abroad-at-home-aquestion-of-confidence.html/ (accessed 01.06.2018).

37. Grace K. The Asilomar Conference: A Case Study in Risk Mitigation. MIRI Research Institute, Technical Report, 2015, no. 9. Berkeley, CA, MIRI, 2015, July 15, p. 15.

YURISPRUDENSIYA

HUQUQIY ILMIY-AMALIY JURNALI

2021/5

ISSN 2181193

9 772181 193800

Таҳририят манзили:
100047. Тошкент шаҳар, Сайилгоҳ кўчаси, 35.
Тел.: (0371) 233-66-36, 233-41-09.
Факс: (0371) 233-37-48.

Веб-сайт: www.tsul.uz
E-mail: lawjournal@tsul.uz
E-mail: tn.tdyu@mail.ru

Журнал 15.12.2021 йилда типографияга топширилди. Қоғоз бичими: А4.
Шартли 23,52 б.т. Адади: 100. Буюртма: № 70.

ТДЮУ типографиясида чоп этилди.