Научная статья **https://doi.org/10.21202/2782-2923.2024.1.217-245**

УДК 004.8:340.1:346:347

Л. Парентони[1]

[1] *Федеральный университет Минас-Жерайс (ФУМЖ), г. Белу-Оризонти, Бразилия*

Чего следует ожидать от искусственного интелекта?

Переводчик *Е. Н. Беляева*

Леонардо Парентони, профессор права, Федеральный университет Минас-Жерайс (ФУМЖ)
ORCID ID: https://orcid.org/0000-0002-3593-2831

Аннотация

Цель: изучение несоответствия между ожиданиями от систем искусственного интеллекта (ИИ) и их текущими возможностями. В настоящее время искусственный интеллект является широко распространенной и передовой технологией и используется в различных отраслях, таких как сельское хозяйство, промышленность, торговля, образование, профессиональные услуги, «умные города» и киберзащита. Однако несоответствие между ожиданиями от искусственного интеллекта и его текущими возможностями приводит к двум нежелательным последствиям. Во-первых, от ИИ ожидают результатов, выходящих за рамки его нынешней стадии развития, что приводит к нереалистичным требованиям. Во-вторых, возникает неудовлетворенность существующими возможностями ИИ, хотя во многих контекстах они достаточны.

Методы: для решения проблемы несоответствия в статье используется аналитический подход. Анализируются различные рыночные приложения ИИ, раскрывается их разнообразие. Показано, что ИИ не является однородной, единой концепцией, но включает в себя широкий спектр отраслевых приложений, каждое из которых служит своим целям, обладает присущими ему рисками и соответствует определенным уровням точности.

Результаты: основной вывод статьи заключается в том, что несоответствие между ожиданиями и реальными возможностями ИИ возникает из-за ошибочной предпосылки, что системы ИИ должны всегда достигать точности, значительно превосходящей человеческие стандарты, независимо от контекста. Рассматривая различные рыночные приложения, автор выступает за то, чтобы оценивать потенциал ИИ и приемлемые уровни точности и прозрачности в зависимости от контекста. Результаты показывают, что для каждого приложения ИИ должны быть приняты свои целевые показатели точности и прозрачности, подбираемые в каждом конкретном случае. Следовательно, системы ИИ применимы в различных контекстах, даже если их точность или прозрачность ниже возможностей человека.

Научная новизна: статья опровергает широко распространенное заблуждение о том, что ИИ должен работать со сверхчеловеческой точностью и прозрачностью во всех сценариях. Раскрывая разнообразие приложений ИИ и их задач, автор подчеркивает, что ожидания и оценки должны быть адаптированы к конкретному контексту использования.

Практическая значимость: статья представляет ценность для заинтересованных сторон в области ИИ, включая регулирующие органы, разработчиков и потребителей. Пересмотр ожиданий на основе контекста способствует принятию обоснованных решений и повышению ответственности при разработке и внедрении ИИ. Работа помогает повысить общий уровень использования и принятия технологий ИИ за счет более реалистичного понимания возможностей и ограничений ИИ в различных контекстах. Автор предлагает более широкий взгляд на проблему, призывая к созданию надежной нормативно-правовой базы и ответственному внедрению систем ИИ, что будет способствовать улучшению применения ИИ в различных секторах. Кроме того, более реальные требования позволят обеспечить понимание путей развития и регулирования ИИ.

Впервые статья опубликована на английском языке в журнале *Il Diritto Degli Affari*. По вопросам коммерческого использования обратитесь в редакцию журнала. E-mail: ildirittodegliaffari@gmail.com

Цитирование оригинала статьи на английском: Parentoni, L. (2022). What Should we Reasonably Expect from Artificial Intelligence? *Il Diritto Degli Affari*, 2, 179.

URL публикации: https://www.ildirittodegliaffari.it/articolo/123

Ключевые слова:
точность, искусственный интеллект, цифровые технологии, инновации, право, регулирование, прозрачность

Как цитировать русскоязычную версию статьи: Парентони, Л. (2024). Чего следует ожидать от искусственного интеллекта? *Russian Journal of Economics and Law*, *18*(1), 217–245. https://doi.org/10.21202/2782-2923.2024.1.217-245

Scientific article

L. Parentoni[1]

[1] *Federal University of Minas Gerais – UFMG, Belo Horizonte, Brazil*

What Should we Reasonably Expect from Artificial Intelligence?

Translator *E. N. Belyaeva*

Leonardo Parentoni, Tenured Law Professor, Federal University of Minas Gerais – UFMG
ORCID ID: https://orcid.org/0000-0002-3593-2831

Abstract
Objective: the objective of this article is to address the misalignment between the expectations of Artificial Intelligence (or just AI) systems and what they can currently deliver. Despite being a pervasive and cutting-edge technology present in various sectors, such as agriculture, industry, commerce, education, professional services, smart cities, and cyber defense, there exists a discrepancy between the results some people anticipate from AI and its current capabilities. This misalignment leads to two undesirable outcomes: Firstly, some individuals expect AI to achieve results beyond its current developmental stage, resulting in unrealistic demands. Secondly, there is dissatisfaction with AI's existing capabilities, even though they may be sufficient in many contexts.
Methods: the article employs an analytical approach to tackle the misalignment issue, analyzing various market applications of AI and unveils their diversity, demonstrating that AI is not a homogeneous, singular concept. Instead, it encompasses a wide range of sector-specific applications, each serving distinct purposes, possessing inherent risks, and aiming for specific accuracy levels.
Results: the primary finding presented in this article is that the misalignment between expectations and actual AI capabilities arises from the mistaken premise that AI systems should consistently achieve accuracy rates far surpassing human standards, regardless of the context. By delving into different market applications, the author advocates for evaluating AI's potential and accepted levels of accuracy and transparency in a context-dependent manner. The results highlight that each AI application should have different accuracy and transparency targets, tailored on a case-by-case basis. Consequently, AI systems can still be valuable and welcomed in various contexts, even if they offer accuracy or transparency rates lower or much lower than human standards.
Scientific novelty: the scientific novelty of this article lies in challenging the widely held misconception that AI should always operate with superhuman accuracy and transparency in all scenarios. By unraveling the diversity of AI applications and their purposes, the author introduces a fresh perspective, emphasizing that expectations and evaluations should be contextualized and adapted to the specific use case of AI.

The article was first published in English language by Il Diritto Degli Affari. For more information please contact: ildirittodegliaffari@gmail.com

For original publication: Parentoni, L. (2018). What Should we Reasonably Expect from Artificial Intelligence? Il Diritto Degli Affari, 2, 179.

Publication URL: https://www.ildirittodegliaffari.it/articolo/123

Practical significance: the practical significance of this article lies in providing valuable guidance to stakeholders within the AI field, including regulators, developers, and customers. The article's realignment of expectations based on context fosters informed decision-making and promotes responsible AI development and implementation. It seeks to enhance the overall utilization and acceptance of AI technologies by promoting a realistic understanding of AI's capabilities and limitations in different contexts. By offering more comprehensive guidance, the article aims to support the establishment of robust regulatory frameworks and promote the responsible deployment of AI systems, contributing to the improvement of AI applications in diverse sectors. The author's call for fine-tuned expectations aims to prevent dissatisfaction arising from unrealistic demands and provide solid guidance for AI development and regulation.

Keywords:
accuracy, artificial intelligence, digital technologies, innovation, law, regulation, transparency

For citation of Russian version: Parentoni, L. (2024). What Should we Reasonably Expect from Artificial Intelligence? *Russian Journal of Economics and Law, 18(1)*, 217–245. https://doi.org/10.21202/2782-2923.2024.1.217-245

Введение

Искусственный интеллект (далее – ИИ) – одна из самых широко распространенных и передовых технологий нашего времени. Уже сегодня он представлен в самых разных отраслях, таких как сельское хозяйство, промышленность, торговля, образование, профессиональные услуги, «умные города», киберзащита и т. д.[1] Однако что такое ИИ и чего нам, людям, стоит ожидать от него? Как отмечается в литературе, «этот вопрос легко задать, но на него трудно ответить» (Kaplan, 2016).

Первый шаг к правильному ответу на этот вопрос – признать, что ИИ не является единой концепцией. Продукты и услуги на основе ИИ охватывают широкий спектр отраслевых приложений с различными целями, рисками и уровнем точности. Действительно, универсального определения ИИ не существует[2]. Это одно из самых противоречивых понятий в данной области и его техническое значение в настоящее время обсуждается в рамках законодательного процесса как в ЕС, так и в США. Об этом свидетельствует то, что первая версия текста совместного проекта ЕС и США[3], запущенного в мае 2023 г. и направленного на выработку единых понятий (таксономии) в области ИИ, не содержит определения искусственного интеллекта. Кроме того, определение должно быть достаточно абстрактным и широким, чтобы охватить множество возможных применений этой технологии (включая те, которые еще предстоит разработать), но в то же время достаточно

[1] «Искусственный интеллект (ИИ) – это технология общего назначения, которая способна улучшить благосостояние и благополучие людей, внести вклад в устойчивую глобальную экономическую деятельность, повысить уровень инноваций и производительности, а также помочь в решении ключевых глобальных проблем. Он применяется во многих отраслях – от производства, финансов и транспорта до здравоохранения и безопасности» (OECD. (2019, May). OECD AI Principles overview. https://oecd.ai/en/ai-principles.).

[2] «Используемые [для определения ИИ] методы всегда различны; поиск человекоподобного интеллекта должен быть отчасти эмпирической деятельностью, связанной с психологией, включая наблюдения и гипотезы о поведении и мыслительных процессах человека; с другой стороны, рационалистический подход предполагает сочетание математики и инженерии и связан со статистикой, теорией управления и экономикой» (Russell & Norvig, 2022).

[3] «Выявленная терминология отражает общее для ЕС и США техническое, социотехническое и ценностное понимание систем ИИ и служит основой для будущих определений, а также для будущего трансатлантического сотрудничества в области терминологии и таксономии ИИ. <...> В конечном счете различные терминологии выражают разные “технологические подходы”, тем самым выявляя как через совпадения, так и через расхождения наличие пробелов, расхождений и несоответствий <...>. Кроме того, в представленный ниже список не включены термины, которые в настоящее время обсуждаются и определяются в рамках законодательных процессов в ЕС и/или США, чтобы не мешать этим процессам» (European Union. European Commission. (2023, May 31). EU-U.S. Terminology and Taxonomy for Artificial Intelligence. Brussels. https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence).

узким, чтобы разграничить то, что является ИИ и должно подлежать специальному регулированию. Это не значит, что определений ИИ не существует, но окончательное определение еще не выработано.

Так, по мнению ОЭСР, «система ИИ – это машинная система, способная делать прогнозы, давать рекомендации или принимать решения, влияющие на реальную или виртуальную среду, исходя из заданного набора целей, определенных человеком»[4]. Эксперты-юристы считают, что «лучше всего понимать его как набор методов, направленных на то, чтобы с помощью машин приблизиться к некоему аспекту познания, используемому человеком или животными» (Calo, 2017), или как «машину, способную выполнять задачи, о которых, если бы их выполнял человек, можно было бы сказать, что они требуют наличия разума» (Scherer, 2016). Если говорить о правовом поле, то проект закона ЕС об ИИ[5], рассмотренный ниже в разд. 6, определяет систему ИИ как «программное обеспечение, разработанное с использованием одного или нескольких методов и подходов, перечисленных в Приложении I, и способное при заданном наборе целей, определенных человеком, генерировать такие результаты, как контент, прогнозы, рекомендации или решения, влияющие на среду, с которой они взаимодействуют».

За пределами юридической сферы психологи говорят об интеллекте как о «биопсихологическом потенциале обработки информации, который может быть активирован в культурной среде для решения проблем или создания продуктов, представляющих ценность в данной культуре» (Gardner, 1999). Специалисты в области компьютерных наук, в свою очередь, обычно фокусируются на конкретной сфере применения ИИ[6], такой как экспертные системы, машинное обучение, нейронные сети, робототехника, компьютерное зрение и обработка естественного языка. Наконец, в июне 2022 г. был опубликован *ISO/IEC* 22989 – международный стандарт «концепций и терминологии в сфере искусственного интеллекта»[7].

Независимо от области применения термин «искусственный интеллект» вводит в заблуждение, поскольку напрямую связывает алгоритмические процессы с имитацией человеческого интеллекта. Специалисты по нейросетям категорически отвергают подобные ассоциации[8]. Выражение «искусственный интеллект» не просто терминологически неверно; оно уводит нас от того, что действительно важно, и привносит в дискуссию бессмысленные вопросы. Jerry Kaplan проводит такую аналогию:

> «Чтобы лучше понять, как наши идеи о связи между машиной и человеческим интеллектом запутывают наше понимание этой важной технологии, представьте себе, что могло бы произойти, если бы самолеты с самого начала назывались “искусственными птицами”. Мы отвлекались бы на сравнения между авиацией и птицами и вели философские споры о том, действительно ли самолеты могут “летать”, как птицы, или просто имитировать полет. <...> Если бы эта неуместная формулировка сохранялась, то мы бы начали проводить научные конференции и обсуждать, что произойдет, когда самолеты научатся строить гнезда, разовьют способности проектировать и создавать свое потомство, добывать топливо для кормления потомства и т. д.» (Kaplan, 2016).

Более удачными терминами были бы «аналитические вычисления»[9] или «машинное поведение»[10]. Однако, поскольку термин «искусственный интеллект (ИИ)» закрепился во всем мире, в данном исследовании мы будем использовать именно его. Также следует оговорить, что данная работа не посвящена какому-либо

---

[4] OECD. (2019, May). OECD AI Principles overview. https://oecd.ai/en/ai-principles

[5] European Union. (2021, April 21). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[6] Подробное описание развития ИИ см. (Nilsson, 2010). Список областей использования ИИ см. (Vijipriya et al., 2016).

[7] Универсальным правовым инструментом для регулирования проектирования, разработки и использования систем искусственного интеллекта является CAI, который разрабатывает Совет Европы; в этом документе также будут предложены стандарты в области прав человека и вопросов экологии.

[8] Например, H. L. Dreyfus: «Таким образом, нас вводит в заблуждение утверждение, что если информацию можно обработать, то ее можно обработать цифровым образом» (Dreyfus, 1965). А также (Nicolelis & Cicurel, 2015; Damasio, 2010).

[9] Там же, p. 17.

[10] «…Мы видим примеры поведения машин на трех уровнях исследования: отдельные машины, коллективы машин и группы машин, встроенные в социальную среду с группами людей в гибридных или гетерогенных системах. В качестве отдельной машины изучается сам алгоритм; в случае коллектива машин – взаимодействие между машинами, а в случае гибридного поведения человека и машины – взаимодействие между машинами и людьми» (Rahwan, 2019).

конкретному направлению ИИ в том или ином рыночном секторе. Цель данной статьи – обсудить фундаментальный вопрос, имеющий первостепенное значение для регуляторов, разработчиков и пользователей: чего следует ожидать от ИИ?

Для того чтобы ответить на этот вопрос, в статье представлено следующее. В разд. 1 описываются различные способы предоставления продуктов и услуг на основе ИИ, составляющие контекст применения этой технологии, что важно для анализа конкретных приложений. В разд. 2 показано, что в научной литературе широко обсуждается вопрос уровня точности систем ИИ, подразумевая, что эти системы должны превосходить возможности человека, независимо от контекста. В разд. 3 приводится авторская трехуровневая классификация степени участия ИИ в процессе принятия решений человеком. В разд. 4 обсуждается несоответствие между ожиданиями от ИИ и реальными возможностями этой технологии; предлагаются авторские критерии для определения того, чего следует ожидать от ИИ в каждом контексте, исходя из цели использования этой технологии, уровня вмешательства ИИ в процесс принятия решений человеком, показателей точности, анализа рисков и прозрачности. Наконец, в разд. 5 показано, что этот подход уже используется во многих законодательных инициативах на национальном и международном уровнях.

Результаты исследования

1. Многообразие способов предоставления продуктов и услуг на основе ИИ

Различные типы систем ИИ создают разные возможности и проблемы в области политики.

OECD[11]

ИИ охватывает широкий спектр секторов рынка. Кроме того, в каждом из них существует огромное количество способов разработки и внедрения одного и того же приложения, в зависимости от стратегии разработчиков и ретейлеров[12]. Каждый из них преследует свои цели, отличается точностью и рисками. Эти различия необходимо учитывать при ответе на вопрос, чего нам стоит ожидать от ИИ.

Одно из фундаментальных различий касается воплощенного и бестелесного ИИ. Воплощенный ИИ – это приложения, в которых система искусственного интеллекта является неотъемлемой частью материального продукта, например, промышленного оборудования или автономных автомобилей. Для полноценной работы такого рода приложениям необходима заранее определенная физическая структура[13]. Воплощенный ИИ обычно называют роботом[14], однако существуют различные категории этого понятия. Робот, похожий на человека, называется гуманоидом или андроидом, а роботы, предназначенные для специфического социального взаимодействия и вызывающие человеческие эмоции, называются социальными роботами[15]. Большинство из

[11] OECD. (2022, February 22). OECD Framework for the Classification of AI Systems. https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en

[12] «Ценности определяют форму технологий. Ценности, воплощенные в системах и устройствах, не просто являются функцией их объективной формы. Мы также должны изучить сложное взаимодействие между системой или устройством, теми, кто его создал, целями и условиями его использования, а также природным, культурным, социальным и политическим контекстом, в который оно вписано, ведь все эти факторы могут отразиться на ценностях, воплощенных в нем» (Nissenbaum, 2001).

[13] Например, в классических источниках: «Эта книга писалась почти 50 лет, с тех пор как еще дошкольником я понял, что из неодушевленных деталей механического строительного набора получаются одушевленные существа. Этот опыт вылился в статью в 1978 г., книгу в 1988-м и эту работу в 1998-м» (Moravec, 1999).

[14] «Роботы – это физические агенты, которые выполняют задачи, манипулируя физическим миром» (Russell, 2010). «...мы можем определить робота как искусственную систему, которая: (*i*) имеет физическое тело, включающее исполнительные механизмы, датчики и мозг, (*ii*) находится в физической среде и в конечном счете в социальной среде, включающей других роботов и/или людей, и (*iii*) демонстрирует поведение с целью выполнения определенной функции» (Nolfi, 2021).

[15] «Социальный робот – это физически воплощенный автономный агент, который общается и взаимодействует с людьми на социальном уровне. <...> Социальные роботы общаются с помощью социальных сигналов, демонстрируют адаптивное обучающееся поведение и имитируют различные эмоциональные состояния. Наше взаимодействие с ними следует моделям социального поведения и направлено на поощрение эмоциональных отношений. Примерами первых социальных роботов являются интерактивные роботы-игрушки, такие как собака *AIBO* от *Sony* и робот-динозавр *Pleo от Innovo Labs*...» (Darling, 2012).

последних имитируют кукол и домашних животных. Ученые выступают за то, чтобы к социальным роботам относились не как к обычной собственности, а как к «части семьи», как к домашним животным, поскольку они могут создавать глубокие связи с людьми (особенно с детьми), оказывая значительное влияние на их эмоции и социальные взаимодействия. В рамках той же логики считается, что социальные роботы могут стать жертвами жестокого обращения[16].

Напротив, бестелесный ИИ не привязан к какой-либо конкретной физической структуре, по крайней мере, на стороне клиента. Такие приложения могут работать одновременно на многих устройствах почти с одинаковой точностью и охватывать большую аудиторию. Примером могут служить облачные сервисы.

Это различие имеет значение, потому что ущерб от ошибки в воплощенном ИИ будет, скорее всего, локальным, а неисправность в бестелесной системе может вызвать глобальные проблемы, в зависимости от формы управления продуктом или услугой. Кроме того, ошибка в социальном роботе может вызвать длительные психологические проблемы и нарушить социальное взаимодействие, тогда как дефект промышленного оборудования вряд ли нанесет такой же ущерб. Таким образом, *цель* использования каждой системы, *ожидаемая точность* и *прозрачность*, а также допустимые *риски* сильно различаются в зависимости от того, каким образом предоставляется продукт или услуга.

Существует множество других классификаций, но их подробное описание не является целью данной статьи. Достаточно сказать, что способ разработки и использования каждой системы является *одним из основных факторов*, которые необходимо учитывать при оценке перспектив ИИ. Более того, анализ должен проводиться *в каждом конкретном случае* с учетом специфики каждой ситуации.

2. Вопросы точности в работе ИИ

Спросите специалистов, почему вокруг машинного обучения столько шумихи, и ответом будет одно слово – точность.

D. Lehr и P. Ohm (2017)

Как следует из приведенной выше цитаты, вызывает озабоченность уровень точности ИИ, т. е., говоря упрощенно, уровень точности, который может обеспечить система ИИ по сравнению с человеческими стандартами[17]. Чем лучше результаты, предоставляемые системой, тем выше этот показатель. Некоторые исследователи утверждают, что наблюдаемая точность напрямую влияет на доверие людей к системам ИИ, даже если фактическая точность ниже[18]. Другие же указывают на то, что модели машинного обучения могут быть менее надежными, чем кажется на первый взгляд[19].

Таким образом, в научной литературе точность считается ключевым фактором при оценке эффективности системы ИИ. В большинстве работ на эту тему, независимо от направленности или цели статьи, данный вопрос обычно затрагивается если не в качестве основного аргумента, то, по крайней мере, как одна из составля-

[16] «В этом разделе предлагается, чтобы защита социальных роботов от жестокого обращения осуществлялась по аналогии с законами о жестоком обращении с животными. Хотя основания этих законов часто оспариваются и многие из них не совпадают с основаниями для защиты роботов, можно наблюдать как психологические, так и философские параллели» (Там же, p. 16). См. также информацию об основанном в 1999 г. Американском обществе по предотвращению жестокого обращения с роботами (American Society for the Prevention of Cruelty to Robots): http://www.aspcr.com/index.html

[17] См. объяснение понятия точности в классификационных программах: «Точность классификатора – это доля правильных предсказаний на тестовом множестве. <...> Показатель точности дает оценку вероятности правильного предсказания; таким образом, чем выше точность, тем лучше классификатор» (Zaki & Wagner, 2020).

[18] «Мы обнаружили, что на доверие людей к модели влияет как ее заявленная точность, так и наблюдаемая точность и что влияние первой может меняться в зависимости от последней» (Yin et al., 2019).

[19] «Современные самообучающиеся машины успешно решают сложные прикладные задачи, достигая высокой точности и демонстрируя кажущееся разумным поведение. <...> Мы призываем добавить осторожности к продолжающемуся ажиотажу вокруг машинного интеллекта и тщательнее оценить некоторые из этих недавних успехов» (Lapuschkin et al., 2019).

ющих. Это можно наблюдать в научных публикациях из таких областей, как здравоохранение[20], ботаника[21], фондовые рынки[22], военное дело[23], юриспруденция[24] и многие другие.

Точность, безусловно, важна, и опасения по этому поводу часто высказываются в научной литературе. Проблема в том, что чрезмерное внимание к данной характеристике может привести к ложному представлению о том, что ИИ всегда должен превосходить возможности человека, независимо от контекста. Однако это не так.

Действительно, существует множество ситуаций, в которых ИИ не должен превосходить человеческие стандарты или даже приближаться к ним. В этом случае ИИ может сыграть значительную роль, просто заменив человеческий труд, даже ценой существенного снижения точности. Выигрыш в других факторах, таких как предотвращение рисков или прозрачность, может компенсировать потерю точности. Таким образом, ложное представление о том, что ИИ должен превзойти человека, в итоге приуменьшает или вовсе не учитывает важность других значимых факторов. Существуют исследования[25], которые проливают свет на этот вопрос. Некоторые ограничения точности могут быть приемлемыми и в конечном итоге неизбежными. Это не обязательно снижает применимость ИИ.

Таким образом, каждая цель использования системы ИИ (с учетом того, как она была реализована) определяет ожидаемую точность, прозрачность и предотвращение рисков. Именно сочетание этих и других факторов в каждом конкретном случае должно учитываться при оценке пригодности системы[26]. Точность сама по себе является лишь одним из них. Принимая это во внимание, в следующих разделах мы опишем другие технические и юридические факторы, которые также нужно взвесить, чтобы ответить на вопрос, поставленный в заголовке: «Чего нам следует ожидать от ИИ?».

3. Различные уровни вмешательства систем на основе ИИ в процесс принятия решений человеком

Многие исследования показывают, что технологическая эволюция может привести человечество к новой парадигме[27]. В этом контексте системы на основе ИИ будут оказывать все большее влияние на процесс при-

---

[20] «Мы наблюдаем экспоненциальный рост количества проводимых обследований, дальнейшую специализацию областей медицины и повышение точности различных методов визуализации...» (Santos et al., 2019). См. также (Nistal-Nuño, 2021).

[21] «Целью исследования является анализ точности методов машинного обучения (*ML*) в отношении объемной модели и функции конусности для черной акации» (Schikowski et al., 2018).

[22] «В модели представлен и рассмотрен возможный способ прогнозирования движения товаров с высокой точностью» (Patel, A., Patel, D., & Yadav, S. (2022). Prediction of Stock Market Using Artificial Intelligence. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3871022).

[23] «В воздушном бою за пределами визуальной дальности одной из проблем является определение наилучшего времени для запуска ракеты, решение о котором должно приниматься быстро. <...> ИНС была обучена на наборе данных, содержащих 1093 зарегистрированных выстрела на военных учениях, и показала точность 78,0 % согласно кросс-валидации» (Lima et al., 2021).

[24] «В частности, в данной работе показано, что сложные социальные эффекты алгоритмических юридических метрик выходят далеко за рамки опасений по поводу точности, которые до сих пор доминировали в критике таких метрик» (Burk, 2020).

[25] «Машинное обучение сосредоточилось на полезности вероятностных моделей для прогнозирования в социальных системах, но только сейчас приходит к пониманию того, насколько эти модели ошибочны и к каким последствиям приводят их недостатки. В данной статье предпринята попытка дать всеобъемлющий структурированный обзор конкретных концептуальных, процедурных и статистических ограничений моделей машинного обучения в применении к общественной сфере. Разработчики моделей машинного обучения могут использовать описанную иерархию для выявления возможных точек отказа и создания способов их устранения, а пользователям моделей машинного обучения она поможет понять, применять ли машинное обучение, где и как» (Malik, M. M. (2020, February). A Hierarchy of Limitations in Machine Learning. Berkman Klein Center for Internet & Society Research Paper, 1–68. https://arxiv.org/abs/2002.05193).

[26] Вопрос о том, кто может и должен проводить такую оценку, выходит за рамки данного исследования.

[27] Например: «Действительно, мы стремительно переходим от эпохи Интернета к алгоритмическому обществу. Вскоре мы будем вспоминать цифровую эпоху как предшественницу алгоритмического общества. Что я имею в виду под алгоритмическим обществом? Я имею в виду общество, организованное на основе принятия социальных и экономических решений алгоритмами, роботами и агентами искусственного интеллекта, которые не только принимают решения, но и в некоторых случаях исполняют их» (Balkin, 2017). См. также A. C. de M. Cansian: «Если до сих пор машины были вспомогательными элементами для человека, то теперь они перестанут ими быть, они станут главными действующими лицами в различных сценариях человеческих отношений, заменяя собой человека в принятии решений и выполняя сложные задачи, до сих пор немыслимые для компьютера» (Cansian, 2021).

нятия решений человеком. Это, безусловно, несет в себе как новые возможности, так и риски[28]. В данном разделе описана авторская трехуровневая классификация вмешательства ИИ в этот процесс, которую стоит учесть при оценке перспектив искусственного интеллекта.

Существуют различные классификации уровня вмешательства систем на базе ИИ в процесс принятия решений человеком. Например, Национальное управление безопасности движения на трассах США (*NHTSA*) подразделяет беспилотные автомобили на пять уровней, от 0 (отсутствие автоматизации) до 4 (полная автономность)[29]. *NHTSA* использует термин «автоматизация», хотя некоторые авторы утверждают, что термин «автономность»[30] был бы более верным. В этой статье оба выражения используются наравне.

Автор предлагает более простое и интуитивно понятное решение. Кроме того, оно лучше подходит в качестве научного инструмента, чем отраслевые классификации, поскольку применимо к любой ситуации, независимо от типа и цели использования системы. Системы ИИ подразделяются на три категории по уровню вмешательства в процесс принятия решений человеком: 1) вспомогательные системы для автоматизации задач; 2) консультативные системы; 3) полноценное принятие решений (табл.).

Уровни вмешательства систем ИИ в принятие решений человеком
Levels of AI systems interference on human decision-making

Тип системы ИИ / Kind of AI system	Уровень автоматизации / Automation Level	Функционирование / Functioning	Ввод / Input	Вывод / Output	Участие в принятии решений человеком / Interference in human decision-making
Вспомогательная система автоматизации задач / Task-automation auxiliary	Начальный / Initial	Пассивное / Passive	Прошлая информация / Past data	Дает информацию или выполняет указания человека / Provide information or run human commands	Не замещает ни один из этапов принятия решений человеком / Do not replace any stage of human decision
Консультативный / Advisory	Средний / Intermediary	Пассивное / Passive	Прошлая информация / Past data	Рекомендует действие / Recommend an action	Замещает один или более из существенных этапов принятия решений / Replace one or more substantial stages of human decision человеком
Полное принятие решений / Full decision-making	Повышенный / Advanced	Активное / Active	Информация в реальном времени / Real-time data	Принимает и исполняет решения самостоятельно / Makes and executes decisions autonomously	Замещает все или почти все этапы принятия решений человеком / Replaces all or almost all human decision

Поскольку реальность гораздо сложнее теории, конечно, будут пограничные случаи, когда система находится в серой зоне между этими уровнями. Например, таковы системы генеративного ИИ, такие как *ChatGPT* от *Open AI*, *Bard* от *Google* или *Llama* от *Meta*[31]. Но даже в этом случае наша классификация уже послужит

[28] «<...> Мы утверждаем, что люди, взаимодействующие с ИИ, ведут себя как ‘борги’, т. е. существа-киборги с сильными индивидуальными характеристиками, но без человеческой индивидуальности» (Fügener et al., 2021).

[29] Подробный отчет см.: http://www.nhtsa.gov/nhtsa/av/pdf/Federal_Automated_Vehicles_Policy.pdf. Анализ этих шести уровней см. (Lima, 2020).

[30] «Автономность – это способность машины выполнять задачи без участия человека. Она отличается от автоматизации, которая заключается в простом использовании машины для выполнения определенного процесса, в то время как автономность описывает систему, способную работать самостоятельно в течение некоторого периода времени без прямого вмешательства человека. <...> Существует три основных аспекта автономности: тип задачи, которую выполняет машина; отношение человека к машине во время выполнения этой задачи; и сложность принятия решений машиной во время выполнения задачи. Эти параметры независимы, и машина может стать ‘более автономной’, увеличив степень автономности по любому из этих аспектов. В рамках этих задач, или параметров, существуют различные степени автономности, которые определяют отношения между человеком и машиной» (Cherry & Johnson, 2021).

[31] Организация признана экстремистской, ее деятельность запрещена на территории Российской Федерации.

своей цели, проливая свет на большинство ситуаций. Более того, это лишь один из основных факторов, которые необходимо учитывать. Итак, кратко рассмотрим каждый уровень.

Вспомогательные системы представляют собой самый низкий уровень автоматизации. Они работают лишь как инструмент для предоставления информации пользователю (т. е. служат цифровым каталогом) или применяются для выполнения задач на основе команд, заданных пользователем. При этом все решения человек всегда принимает самостоятельно. Кроме того, эти системы не проявляют инициативу и не работают по принципу «от противного», они работают только по требованию пользователя (пассивное функционирование). Эти системы могут работать по решению пользователя, но они не берут на себя ни один из этапов процесса принятия решений. Даже если кажется, что такая система делает что-то самостоятельно, например, бронирует встречу или совершает покупку, это не так, ведь эти действия уже были заданы пользователем. В эту группу входят и голосовые помощники[32], такие как *Siri* и *Alexa*. Например, пользователь может попросить помощника «показать пиццерии рядом с домом». ИИ покажет список результатов, возможно, с оценками других пользователей. Затем пользователю предстоит самостоятельно оценить эту информацию и решить, что делать дальше. Можно отказаться от пиццы, а можно дать помощнику новые команды, например, «заказать столик на двоих в этом ресторане в 7 вечера» и «вызвать такси к дому в 18:30». Именно пользователь выбирает ресторан, время бронирования, количество мест и способ передвижения (на велосипеде, на метро, на своей машине или на такси). Прежде всего, именно пользователь решает, будет ли он ужинать в этот вечер. Вспомогательная система автоматизации задач не скажет: «Съешь салат вместо пиццы, так как ты прибавил в весе». Такие системы работают пассивно, отвечая на вопросы пользователей и выполняя конкретные задачи. Даже если кажется, что они «угадывают», чего хочет пользователь, результат все равно основан на обработке предыдущих параметров и анализе повседневной деятельности.

Консультационные системы находятся на ступень выше по уровню автоматизации. Они не только помогают автоматизировать задачи, но и напрямую рекомендуют, что делать пользователю, заменяя тем самым часть процесса принятия решений. Хорошим примером может служить программное обеспечение для медицинской диагностики, основанной на визуализации. Оценивая медицинские изображения и сравнивая их с другими базами данных, например, с научными источниками о рекомендуемых методах лечения для каждой стадии заболевания, система не только предоставляет полезную информацию медицинскому персоналу, но и определяет оптимальный ход лечения для каждого пациента. По сравнению со вспомогательными системами автоматизации задач консультационная система идет на шаг дальше, поскольку участвует в процессе принятия решений человеком, автоматизируя значительную его часть. Однако последнее слово все равно остается за человеком. Ведь именно медицинский персонал должен оценить рекомендованное лечение, объяснить его пациенту и выполнить. Или выбрать другой способ лечения, предоставив научное обоснование для отклонения рекомендаций системы. Общим со вспомогательными системами автоматизации задач является тот факт, что рекомендательные системы также работают пассивно (т. е. от медицинской команды и пациента зависит, применять их или нет). И те и другие обычно основываются на данных, полученных ранее, а не поступающих в режиме реального времени.

На высшем уровне автоматизации находятся системы принятия решений. Они полностью заменяют все или почти все этапы процесса принятия решений. Именно система искусственного интеллекта, а не человек принимает решения и выполняет их. Такие системы обычно подключены к актуальным базам данных и активно функционируют в режиме реального времени. Идеальным примером являются самоуправляемые (полностью автономные беспилотные) автомобили. На самом высоком уровне автоматизации *NHTSA* автомобиль принимает все решения в режиме реального времени. Человек – всего лишь зритель, практически не влияющий на решения системы. В некоторых особых ситуациях система предлагает пользователю выбрать

[32] «Автоматизированный голосовой помощник: использует метод обработки естественного языка (*Natural Language Processing*, *NLP*) для преобразования текста или голосового ввода пользователя в исполняемые команды. Многие из этих устройств постоянно обучаются с помощью методов искусственного интеллекта, включая машинное обучение. Некоторые из них, например *Google Assistant* (включая *Google Lens*) и *Samsung Bixby*, также обладают способностью обрабатывать изображения и распознавать объекты на них, чтобы пользователи могли получить более точные результаты» (OECD. (2022, February 22). OECD Framework for the Classification of AI Systems. https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en).

один из вариантов. Например, выбор между более быстрым (с меньшим трафиком) и безопасным (в обход опасных районов), но более длинным путем или если пользователь хочет изменить маршрут в случае непредвиденной аварии, вызвавшей пробку. Отметим, что после того, как пользователь выбрал один из вариантов, все остальное делает самоуправляемый автомобиль. Это существенно отличается от вспомогательных систем автоматизации задач, поскольку они предоставляют информацию на основе того, что пользователь попросил их сделать, в то время как в самоуправляемых автомобилях именно система искусственного интеллекта, а не пользователь определяет, когда, где и какие варианты будут предложены на выбор человеку, основываясь на данных, получаемых в реальном времени. От консультативных систем они отличаются тем, что все или почти все (а не только некоторые) существенные этапы процесса принятия решений человеком осуществляются искусственным интеллектом.

Как уже упоминалось в разд. 2, в настоящее время широко распространено явно выраженное или подразумеваемое представление о том, что системы на основе ИИ должны всегда превосходить возможности человека, независимо от контекста. Другими словами, их точность должна превосходить человеческие стандарты. Это ложное представление не учитывает практические различия в трехуровневой классификации вмешательства ИИ в процесс принятия решений, представленной в данном разделе. Оно также не учитывает тот факт, что каждая цель использования системы ИИ должна предусматривать не только различные показатели точности, но и приемлемые риски и прозрачность в каждом конкретном случае. В следующих разделах этот аспект рассмотрен подробнее.

4. Чего следует ожидать от ИИ?

Все модели ошибочны, но некоторые из них полезны…
G. E. P. Box (1979)

Этой меткой фразой Джордж Эдвард Пелхэм Бокс в 1979 г. подчеркнул, что идеальных систем не существует. Даже самый современный ИИ будет обладать определенным уровнем неточности, неустранимых рисков и непрозрачности. Система, работающая непрерывно, рано или поздно испытает сбой, вызванный внутренними или внешними факторами. Поэтому невозможно гарантировать абсолютную точность работы системы. Признание этого факта позволяет избежать несоответствия между результатами, которые люди ожидают от ИИ, и тем, что эта технология может дать на самом деле. Более того, это показывает, что существует множество контекстов, в которых ИИ не должен превосходить человеческие стандарты. Он может приносить пользу, заменяя человеческий труд, даже при существенном снижении точности, допущении некоторых рисков или уменьшении прозрачности.

Поэтому ключевым моментом для оценки пригодности любой системы ИИ является анализ каждого конкретного случая для предотвращения рисков, определения приемлемой точности и прозрачности для каждой системы с учетом контекста и цели использования этой системы, а также степени участия ИИ в процессе принятия решений человеком, в соответствии с разработанной автором трехуровневой категоризацией, приведенной в разд. 3. В следующих подразделах будут приведены авторские критерии для установления этих параметров, а также ситуации, в которых следует рассмотреть обязательное вмешательство человека («оператор в контуре управления», *human in the loop*, *HITL*). Следует помнить, что универсального решения не существует, а оптимальный результат дает сочетание этих факторов с учетом специфики каждого контекста.

4.1. Ожидаемый уровень точности

Как уже говорилось в разд. 2, ожидание того, что системы ИИ всегда будут достигать высоких показателей точности, является ложным. Точность – это лишь один из факторов, который необходимо учитывать при оценке пригодности системы. Каждый контекст и цель использования ИИ требуют различных показателей точности. Следовательно, ожидать, что любая система на базе ИИ достигнет точности 90 % или выше, независимо от контекста или цели использования системы, нежелательно и не соответствует современному уровню развития технологий. Напротив, анализ должен проводиться индивидуально, чтобы определить, какая точность является приемлемой для конкретного случая. Минимальный уровень в одних ситуациях может быть оптимальным в других.

Например, представим гипотетическую систему распознавания лиц[33], используемую для контроля доступа людей в крупный город (блокировка въезда или выезда). Представьте, что эта система работает с точностью 90 %[34] в г. Сан-Паулу в Бразилии, где проживает около 12 млн человек. Это значит, что она ошибочно заблокирует въезд огромному количеству людей – 1 млн 200 тыс. человек. А если такая система будет работать в Пекине, население которого в два раза больше? В этих примерах даже высокая точность, например, 90 или 95 %, может оказаться недостаточной, поскольку может привести к катастрофическим последствиям. Тот же подход применим и к реальным жизненным ситуациям, например, когда сотрудники правоохранительных органов используют программное обеспечение для распознавания лиц преступников. В этих и многих других контекстах приемлемы только невероятно высокие показатели точности, например, 98 или 99 %.

С другой стороны, есть случаи, когда достаточно точности в 50 % или даже меньше. Например, в деятельности, связанной с высоким риском или вредом для здоровья, замена человеческого труда ИИ оправдана даже за счет существенного снижения точности, поскольку позволяет избежать физического или психологического ущерба. Таким образом, даже менее точный, чем человек, ИИ может быть весьма полезен в некоторых случаях. Более того, неоптимальная система ИИ в сочетании с возможностью вмешательства человека приводит к улучшению результата, причем итоговая точность будет даже выше, чем при использовании одной только системы. Ситуация «человек + машина» [35] – еще один пример ситуации, когда допустимо снижение точности в работе ИИ. Как известно, программа для игры в шахматы обыграла чемпиона мира Каспарова (ПРИЗНАН ИНОАГЕНТОМ В РФ)[36]. Однако менее известно, что лучшие шахматисты – люди, использующие обычный компьютер для анализа ходов, – могут победить ИИ (Brynjolfsson & McAfee, 2016). Таким образом, система «человек + машина» в некоторых ситуациях достигает большего, чем каждый из них смог бы достичь в одиночку[37].

В вышеупомянутых контекстах чрезмерные требования к точности могут привести к таким нежелательным последствиям, как препятствование выходу на рынок новых игроков, если от них требуются показатели точности гораздо выше, чем могут (и должны) обеспечить их продукты или услуги; неоправданное удлинение цикла разработки и тестирования; рост производственных затрат и во многих случаях даже предотвращение внедрения продуктов и услуг, которые могли бы снизить высокий неустранимый риск или вред для здоровья

---

[33] «Технология распознавания лиц позволяет сравнивать цифровые изображения лиц, чтобы определить, являются ли они изображениями одного и того же человека. Сравнение записей, полученных с видеокамер, с изображениями в базах данных называется ‘технологией распознавания лиц в реальном времени’» (O’Flaherty, 2020). См. также (Jain et al., 2011).

[34] «Алгоритмы технологии распознавания лиц дают не окончательный результат, а только вероятность того, что два изображения принадлежат одному и тому же человеку» (Там же, p. 172).

[35] «В существующей литературе основное внимание уделяется описанию типов рабочих мест, которые могут пострадать в результате развития ИИ, а также которые он может создать. Другими словами, смысл существующих исследований в основном сводится к противостоянию человека и машины, соперничеству между ними, а также изучению способов адаптации людей и прогнозам по реорганизации рабочих мест. При этом людям часто приписывается пассивная либо реактивная позиция – они борются с проблемами и ищут новые возможности, открывающиеся в связи с развитием ИИ. Сравнительно мало исследований посвящено тому, как квалифицированные работники могут использовать свой высокий потенциал с помощью технологий ИИ, хотя, как можно предположить, именно это является основной целью при проектировании и разработке ИИ. Цель данного исследования – перейти от соревнования человека и машины к потенциальному балансу ‘человек плюс машина’» (Cao et al., 2021). См. также Yu. N. Harari: «Превращение людей в богов может пойти по одному из трех путей: биологическая инженерия, инженерия киборгов и инженерия неорганических существ» (Harari, 2016).

[36] И многих других профессиональных игроков. Так, совсем недавно ИИ обыграл восьмерых чемпионов мира по игре в бридж: Spinney, L. (2022, Mar. 29). Artificial intelligence beats eight world champions at bridge. The Guardian. London. https://www.theguardian.com/technology/2022/mar/29/artificial-intelligence-beats-eight-world-champions-at-bridge.

[37] «…наше будущее – это слияние с искусственными интеллектуальными машинами! О том, как я пришел к такому выводу, и пойдет речь в этой книге. Я не хочу сказать, что в ближайшие десятилетия мы, люди, будем выглядеть и действовать как роботы на конвейере; скорее, мы будем оснащены таким количеством технологий, включая вычислительные устройства, имплантированные в мозг, что из биологического существа превратимся в технологическое, развивающееся по законам технологии в большей степени, чем по законам биологической эволюции» (Woodrow, 2015). Подробнее об обязательном участии человека (оператор в контуре управления) см. в разд. 4.3. Также нужно учитывать, что система ‘человек + машина’ пока вызывает споры: «Мы анализируем, как рекомендации ИИ влияют на взаимодополняемость между людьми и ИИ, в частности на то, что люди знают, чего не знает ИИ: это уникальные человеческие знания. <…> Результаты моделирования, основанные на наших экспериментальных данных, показывают, что группы людей, взаимодействующие с ИИ, гораздо менее эффективны по сравнению с группами людей без помощи ИИ» (Fügener et al., 2021).

людей. Подводя итог, можно сказать, что преувеличенное внимание к точности может поставить под угрозу инновации, снизить конкурентоспособность и благосостояние, как отмечают в ОЭСР[38].

Далее возникает следующий вопрос: как определить приемлемую точность системы ИИ в каждом конкретном случае? Ответ на него можно получить, взвесив два фактора: 1) уровень вмешательства ИИ в процесс принятия решений (как описано в разделе 3); и 2) присущие автоматизируемой деятельности риски.

С одной стороны, чем выше уровень вмешательства ИИ в процесс принятия решений, тем выше потребность в обеспечении точности системы, превосходящей человеческие стандарты. С другой стороны, чем выше риски, присущие автоматизируемой деятельности, тем ниже допустимая точность системы. Другими словами, между этими факторами существует обратная зависимость. Вмешательство ИИ в процесс принятия решений указывает на необходимость в более высоких показателях точности, в то время как более высокие неустранимые риски могут оправдать более низкую точность. Вот почему так важно взвешивать эти два фактора в каждом конкретном случае.

Что касается первого фактора, то в разд. 3 уже были описаны три различных уровня вмешательства ИИ в процесс принятия решений. В целом более высокие уровни вмешательства ИИ требуют более высокой точности. На уровне 1 (вспомогательные системы для автоматизации задач) более низкие показатели точности приемлемы, поскольку именно человек, а не ИИ, оценивает всю информацию и принимает решение. Система работает только как инструмент для поиска этой информации и в конечном счете автоматизации некоторых задач, основанных на командах человека. На уровне 2 (консультативные системы), по крайней мере, одна существенная часть принятия решений будет полностью заменена результатами работы системы. Поэтому ожидается более высокая точность, чем на уровне 1. Действительно, точность здесь должна быть хотя бы эквивалентна человеческим стандартам, поскольку неправильный вывод может поставить под угрозу последующие шаги. Наконец, на третьем уровне (системы полноценного принятия решений) все или почти все существенные этапы процесса принятия решений заменяются ИИ. В этом контексте логично ожидать, что точность результатов работы системы будет выше, чем у человека. В конце концов, не имеет смысла заменять человеческий труд системой, которая работает хуже, ведь цель автоматизации – повышение эффективности.

Что касается второго фактора, то каждый вид деятельности имеет определенный уровень неустранимых рисков[39], т. е. рисков, естественно связанных с этой деятельностью, которые в конечном итоге могут быть смягчены, но не предотвращены на 100 %[40]. Когда уровень неустранимых рисков высок, имеет смысл принять более низкие (или гораздо более низкие) стандарты точности системы ИИ по сравнению с человеческими. В этом контексте выигрыш в других факторах, таких как предотвращение рисков или прозрачность, может компенсировать потерю точности. В конце концов, сохранение физического и психологического благополучия человека – это цель, которая должна превалировать даже за счет снижения точности. Классический пример – робот, предназначенный для обезвреживания бомб. Даже если его точность ниже человеческих стандартов, все равно оправданно использовать робота вместо эксперта-человека, чтобы снизить риск травм или смерти этого эксперта. В эту линию рассуждений вписывается и ряд других неустранимых рисков или вредных видов деятельности.

Поэтому очень важно взвесить как уровень вмешательства ИИ в процесс принятия решений, так и уровень рисков, присущих той или иной деятельности, чтобы определить, какая точность ИИ является приемлемой в каждом конкретном случае[41]. В некоторых случаях снижение риска может оправдать точность ниже че-

---

[38] «Политики предпочитают использовать подход к регулированию ИИ, основанный на оценке рисков, чтобы сосредоточить возможности для надзора и вмешательства там, где они наиболее необходимы, избегая при этом ненужных препятствий для инноваций» (OECD. (2022, February 22). OECD Framework for the Classification of AI Systems. https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en).

[39] «Риски при использовании любой системы ИИ существенно зависят от сферы применения. Поскольку сложно предугадать и оценить все возможные варианты использования, применяемые системы ИИ следует разделить на несколько групп по уровню риска» (Там же).

[40] «Риск – это возможная опасность, которая более или менее предсказуема, и она отличается от ‘álea’ (непредсказуемого) и от опасности (фактической). Риск абстрактен». (Lopez, 2010). «ИИ может легко стать слоном в посудной лавке, если мы не будем уделять внимание его развитию и применению» (Cath et al., 2017).

[41] Такая интерпретация согласуется с международными инициативами в области регулирования ИИ, такими как предложение ЕС о создании законодательства в области искусственного интеллекта и Бразильская стратегия в области искусственного интеллекта. Обе эти инициативы будут кратко описаны в разд. 5.

ловеческих стандартов; это также может иметь место и на 2-м и 3-м уровнях автоматизации, как в примере с роботом для обезвреживания бомб, принимающим все решения. Если риск высок и его вряд ли можно снизить, представляется лучшим вариантом подвергнуть этому риску робота вместо человека[42], даже за счет существенного снижения точности. В других случаях достаточно, чтобы система обладала точностью, эквивалентной человеческой, и, таким образом, освобождала человека от труда без существенного снижения эффективности. Примером может служить программное обеспечение, используемое в судах для классификации исков в зависимости от обсуждаемого правового вопроса, времени подачи или любого другого параметра. Чтобы приносить пользу, такому программному обеспечению достаточно достичь точности, аналогичной точности сотрудника, которого заменит ИИ, что позволит человеку уделять больше времени другим видам деятельности и, возможно, достичь большей продуктивности. Наконец, если риск, присущий автоматизируемой деятельности, высок и неудача может поставить под угрозу основные права, как, например, при проведении медицинских тестов по визуальным материалам или роботизированной хирургии, максимальная точность становится обязательной. Другими словами, этические и правовые нормы должны запрещать использование систем искусственного интеллекта людьми, если научно доказано, что человек, выполняя ту же задачу, может добиться лучших результатов, не ставя под угрозу ни одну из заинтересованных сторон (что отличает этот пример от робота, обезвреживающего бомбу). Это особенно верно на третьем уровне автоматизации, поскольку именно система, а не пользователь, оценивает данные в реальном времени и принимает решение. Такой подход может предотвратить катастрофические ситуации, как в деле Mracek v. Bryn Mawr Hosp[43].

4.2. Прозрачность/объяснимость[44]

Хотя понятия прозрачности[45] и объяснимости[46] технически различны[47], в данном разделе они рассматриваются вместе, поскольку тесно связаны между собой. Упрощенно говоря, они означают, что человек способен понять, почему система ИИ выдала определенный результат, и объяснить это обычному пользователю

---

[42] С этим может не согласиться Kate Darling и другие энтузиасты создания социальных роботов.

[43] В 2010 г. Роланду К. Мрачеку в больнице Брин-Мор была проведена операция с использованием так называемого хирургического робота да Винчи. Он утверждает, что в начале операции был в сознании и видел, как робот выдает сообщения об ошибках. По его словам, команда врачей пыталась перезагрузить робота, но сообщения об ошибках продолжали появляться на экране. Они также позвонили в службу технической поддержки, и представитель производителя робота приехал в операционную, но не смог решить проблему. В результате сбоя в работе аппарата примерно через 45 минут хирургическая бригада отказалась от попытки роботизированной операции и провела ее вручную. Результат оказался трагическим: Мрачек получил необратимые повреждения и был вынужден жить с постоянной болью. После этого он решил подать иск против больницы, основываясь на положении о строгой ответственности за качество. Он обвинил врачей в халатности, утверждая, что основной причиной стала неисправность робота. В итоге суд удовлетворил ходатайство ответчика, поскольку Мрачек не представил доказательств причинно-следственной связи между неисправностью робота и результатами его операции. (The United States of America. Mracek v. Bryn Mawr Hosp. United States Court of Appeals for the 3rd Circuit. 610 F. Supp. 2d 401, j. 28.01.2010. https://casetext.com/case/mracek-v-bryn-mawr-hosp).

[44] «Одной из проблем, препятствующих установлению общих оснований, является взаимозаменяемое неверное использование в литературе понятий „интерпретируемость“ и „объяснимость“. Между этими понятиями существуют заметные различия. Интерпретируемость – это пассивная характеристика модели, показывающая уровень, на котором данная модель имеет смысл для человеческого наблюдателя. Эту характеристику также можно описать как прозрачность. Напротив, объяснимость рассматривается как активная характеристика модели, обозначающая любое действие или процедуру, выполняемую моделью с целью прояснения или детализации ее внутренних функций» (Arrieta, 2020).

[45] «Модель считается прозрачной, если она понятна сама по себе…» (Там же, p. 83).

[46] «Объяснимость связана с понятием объяснения как интерфейса между человеком и субъектом, принимающим решения, который одновременно точно представляет принимающего решения субъекта и понятен человеку» (Там же). «...объяснимость – [это] способность машинного обучения обосновывать свои оценки» (Lehr & Ohm, 2017).

[47] «Прозрачность означает предоставление человеку возможности понять, как исследуются, проектируются, разрабатываются, внедряются и используются системы ИИ, в соответствии с контекстом использования и чувствительностью системы ИИ. Она также может включать в себя понимание факторов, влияющих на конкретный прогноз или решение, но обычно не включает в себя предоставление конкретного кода или наборов данных. В этом смысле прозрачность – это социотехнический аспект, цель которого – завоевать доверие людей к системам ИИ. Объяснимость – это способность сделать понятной работу систем ИИ и дать представление о ее результатах. Объяснимость моделей ИИ также относится к понятности ввода, вывода данных и действий каждого алгоритмического блока и его вкладу в результат работы модели. Таким образом, объяснимость тесно связана с прозрачностью, поскольку результаты и субпроцессы, приводящие к результатам, должны быть понятны и прослеживаемы, а также соответствовать контексту использования» (UNESCO. (2021). UNESCO Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000380455).

данной системы. Прозрачность, несомненно, является основополагающей ценностью, предусмотренной многочисленными правовыми нормами как для государственного[48], так и для частного[49] секторов по всему миру. Прозрачность имеет первостепенное значение и должна соблюдаться в соответствии с положениями любой правовой системы, поэтому чем она выше, тем лучше[50].

Однако есть случаи, когда система искусственного интеллекта все еще не может обеспечить высокий уровень прозрачности (например, 90 % или выше) в силу текущего уровня развития технологий. Но некоторые системы могут быть весьма полезны (и широко применимы), несмотря на низкий уровень этого параметра. В примере с роботом, обезвреживающим бомбу, ценность предотвращения серьезных травм человека может компенсировать недостаток прозрачности, особенно если точность робота удовлетворительна. Многие другие реальные ситуации с высоким неустранимым риском следуют этой логике.

Таким образом, прозрачность является одним из основных факторов при оценке пригодности ИИ, равно как и точность. Однако ни то ни другое не является самоцелью. Следовательно, вопрос состоит в том, какой уровень прозрачности необходимо требовать от ИИ в каждом конкретном контексте. Чтобы правильно ответить на этот вопрос, необходимо рассмотреть концепцию «черного ящика» и компромисс между точностью и прозрачностью.

В научной литературе довольно активно обсуждается тот факт, что некоторые системы на основе ИИ, в зависимости от того, как они были разработаны, могут стать «черным ящиком»[51]. Это значит, что человеку – даже разработчику системы – будет сложно или даже невозможно понять точную причину, по которой она выдала тот или иной результат[52]. Это может быть крайне вредно для пользователей системы и для общества в целом, например, если система, намеренно или нет, увеличивает уровень дискриминации в обществе[53] или

---

[48] Например, в Бразилии Федеральная конституция 1988 г. и Закон о доступе к информации № 12527/2011 налагают на правительство требование обеспечить прозрачность.

[49] Требование прозрачности также является одним из основных в законодательстве о защите прав потребителей, законах о защите частной жизни/данных, трудовом законодательстве, регулировании искусственного интеллекта и многих других областях в разных странах.

[50] Несомненно, каждая правовая система может установить ограничения прозрачности для защиты других основных ценностей, например, промышленных и коммерческих секретов. Однако существуют и исключения.

[51] Само понятие «черный ящик» является спорным. Существует общее, известное во всем мире определение: «Термин „черный ящик“ – это полезная метафора, учитывая его двойственное значение. Он может означать записывающее устройство, как, например, системы мониторинга данных в самолетах, поездах и автомобилях. Или он может означать систему, функционирование которой скрыто; мы можем наблюдать входные и выходные данные, но не можем сказать, как одно превращается в другое. Мы ежедневно сталкиваемся с этими двумя значениями: за нами все более тщательно следят компании и правительство, но мы не имеем четкого представления о том, как широко может распространяться эта информация, как она используется и каковы последствия этого» (Pasquale, 2015).

Существуют подклассы «черных ящиков»: «Отсутствие прозрачности связано с так называемым черным ящиком (лучше сказать, черными ящиками). Вероятно, можно выделить три различных системы „черных ящиков“: организационную, техническую и юридическую» (La Diega, 2018).

Существует также понятие «черного ящика», связанное с проприетарным контентом, таким как коммерческие тайны: «Моделью „черного ящика“ может быть либо (1) функция, которая слишком сложна для понимания человеком, либо (2) функция, являющаяся собственностью компании» (Rudin, 2019).

[52] «Одним из очевидных системных рисков ИИ и МЛ [машинного обучения] является проблема „черного ящика“. Эта проблема возникает, когда алгоритмическая система принимает решения, которые крайне сложно объяснить обычным людям. По сути, в алгоритмических системах можно наблюдать входные данные (вход) и исходящие данные (выход), но их внутренние операции не очень хорошо понятны» (Black & Murray, 2019).

[53] Некоторые авторы отмечают, что теоретически предотвратить дискриминацию с помощью ИИ проще, чем предотвратить ее в поведении человека, поскольку ИИ можно с самого начала спроектировать так, чтобы его можно было проверить: «Наша главная мысль заключается в том, что, когда задействованы алгоритмы, доказать дискриминацию будет проще – или, по крайней мере, так должно быть и так можно сделать. Закон запрещает дискриминацию в результате действия алгоритма, и этот запрет может быть реализован путем регулирования процесса, в ходе которого разрабатываются алгоритмы. Для этого можно кодифицировать наиболее распространенный подход к созданию алгоритмов машинного обучения и подробные требования к учету. Такой подход обеспечил бы необходимую прозрачность решений и выборов, осуществляемых при создании алгоритмов, а также компромиссов между значимыми параметрами. <...> Создание надлежащей системы регулирования не просто ограничивает возможность дискриминации со стороны алгоритмов; оно способно превратить алгоритмы в мощный противовес человеческой дискриминации и позитивную силу, способствующую социальному благу в самых разных сферах» (Kleinberg et al., 2018).

выдает любые другие незаконные результаты[54]. Поэтому риск превращения ИИ в «черный ящик» вызывает серьезную озабоченность.

Анализируя альтернативные варианты решения этой проблемы, некоторые авторы указывают на неизбежность поиска компромисса между точностью и прозрачностью. Они утверждают, что более высокие показатели точности приводят к появлению «черных ящиков», в то время как полностью объяснимые системы будут менее точными[55]. Другими словами, повышение одного из этих показателей приведет к снижению другого. Другие авторы указывают, что ориентация на прозрачность как основную цель на всех этапах разработки продукта (прозрачность по замыслу)[56] может обеспечить более высокую точность без снижения прозрачности[57]. Исходя из этого, они утверждают, что предлагаемый компромисс – это лишь предлог для оправдания непрозрачных систем. Они добавляют, что большинство систем ИИ, если не все, становятся непрозрачными, потому что разработчики не уделяли должного внимания прозрачности во время цикла разработки. Следовательно, по их мнению, «черные ящики» – это результат некачественной разработки.

Итак, контекстуализировав понятие черного ящика и предполагаемый компромисс между точностью и прозрачностью, рассмотрим, что, по мнению автора, можно ожидать от ИИ в плане прозрачности. И снова ключевым фактором, который следует учитывать в каждом конкретном случае, является уровень вмешательства ИИ в процесс принятия решений. Действительно, ИИ с низким уровнем вмешательства в процесс принятия решений, как правило, допускает меньшую прозрачность без ущерба для пригодности системы. И наоборот, чем больше вмешательство в систему, тем выше ожидаемая прозрачность. Эту мысль мы разовьем далее.

С одной стороны, ожидать высокой прозрачности от систем первого уровня автоматизации[58], таких как голосовые помощники *Siri* и *Alexa*[59], неразумно, по крайней мере, по двум причинам. Во-первых, такие си-

---

[54] «Внутри алгоритмического „черного ящика“ общественные предубеждения становятся невидимыми и неподотчетными. Созданные исключительно для извлечения прибыли алгоритмы неизбежно расходятся с общественными интересами – информационная асимметрия, возможности для давления и внешние эффекты пронизывают эти рынки» (Benkler, 2019).

[55] «Осмысленные объяснения машинного обучения не работают для каждой задачи. <...> задачи, в которых они хорошо работают, часто имеют всего несколько входных переменных, которые комбинируются относительно простыми способами, такими как возрастающие или убывающие отношения. Системы с большим количеством переменных обычно работают лучше, чем более простые системы, поэтому в итоге мы можем получить компромисс между производительностью и объяснимостью. <...> Оптимизация системы объяснения для интерпретируемости человеком неизбежно означает размывание прогностических характеристик, чтобы охватить только основные логические элементы системы» (Edwards & Veale, 2017). «Также ожидается, что точность будет выше, чем результирующая точность с учетом дискриминации, поскольку техника будет снижать уровень дискриминации за счет точности» (Cardoso et al., 2019).

[56] «Сосредоточившись на управлении моделью и не имея возможности рассмотреть этапы машинного обучения на фоне управления данными, ученые вынуждены воспринять слишком узкий взгляд на потенциальный вред и пользу алгоритмов. Большое внимание уделяется неточностям и предвзятостям, и они действительно могут быть частично связаны с низким качеством данных и спецификаций переменных. Но они могут проявляться и на других этапах машинного обучения, а многие вредные факторы возникают почти исключительно на этих этапах. На самом деле некоторые из самых неприятных последствий машинного обучения – его непрозрачность и необъяснимость – возникают при выборе и разработке алгоритмов, а не при сборе данных или определении переменных» (Lehr & Ohm, 2017). «Я обеспокоен тем, что область интерпретируемости/объяснимости/понятности/прозрачности в машинном обучении далеко ушла от реальных потребностей. <...> Недавние работы, посвященные объяснимости „черных ящиков“, а не интерпретируемости моделей, содержат и закрепляют критические заблуждения, которые обычно остаются незамеченными, но которые могут оказывать долговременное негативное влияние на широкое использование моделей машинного обучения в обществе. <...> Неточная (малодостоверная) модель снижает доверие к объяснению и, как следствие, доверие к черному ящику, который она пытается объяснить» (Rudin, 2019).

[57] «Неверно, что между точностью и интерпретируемостью обязательно существует компромисс. Широко распространено мнение, что более сложные модели являются более точными, а это значит, что для высокой эффективности прогнозирования необходим сложный „черный ящик“. Однако зачастую это не так, особенно если данные хорошо структурированы и представлены в виде естественных значимых признаков. При рассмотрении задач с такими данными после предварительной обработки часто нет существенной разницы в производительности между более сложными классификаторами (глубокие нейронные сети, усиленные деревья решений, случайные леса) и гораздо более простыми классификаторами (логистическая регрессия, списки решений)» (Там же).

[58] Как описано в разд. 3.

[59] В научной литературе обычно также приводят пример картографических приложений, таких как *Google Maps*: «Ошибки будут всегда, потому что модели по своей природе являются упрощениями. Ни одна модель не может учесть всю сложность реального мира или нюансы человеческого общения. Какая-то важная информация неизбежно будет упущена. <...> Поэтому, создавая модель, мы выбираем, что достаточно важно включить в нее, упрощая мир до игрушечной версии, которую легко понять и из которой можно сделать вывод о существенных фактах и действиях. Мы ожидаем, что модель будет выполнять только одну

стемы предоставляют меньше технической информации. Именно пользователь оценивает эту информацию и принимает решение. Если вывод системы необоснован, пользователь может сопоставить эту информацию с другими доступными источниками и критически оценить ее, прежде чем принять решение[60]. Во-вторых, предполагается, что эти системы связаны с коммерческими соглашениями, заключаемыми разработчиком. Соответственно, не стоит удивляться, если голосовой помощник *Apple* отдаст предпочтение рекламной информации деловых партнеров компании *Apple* по сравнению с аналогичной информацией ее конкурентов. То же самое касается *Google*, *Amazon* или любого другого крупного игрока на рынке. Пока в процессе конкуренции не нарушен закон[61], такие коммерческие соглашения являются законными.

С другой стороны, в случае ИИ, который играет значительную роль в принятии решений, например, на 2-м и 3-м уровнях автоматизации, отсутствие прозрачности может стать проблемой даже при высокой точности системы. Непрозрачность сама по себе может поставить под угрозу защищаемые законом ценности, поскольку пострадавшие имеют право знать[62], почему система приняла то или иное решение, особенно если они подозревают, что был использован какой-то незаконный параметр, независимо от того, было ли решение точным или нет. Рассмотрим, например, программное обеспечение для визуализации медицинских данных, которое дает индивидуальные рекомендации по лечению пациента. Даже если медики согласны с рекомендациями программы, они все равно должны быть в состоянии объяснить пациенту, почему система сделала именно такой выбор. Более того, они должны уметь выявлять несоответствия, такие как предвзятость или необоснованные результаты, и предлагать меры для их исправления.

Подводя итог, можно сказать, что более низкая степень участия в процессе принятия решений, как правило, допускает меньшую прозрачность, поскольку пользователям легче предотвратить вред, изучив другие источники информации, прежде чем принять решение, и ожидается, что такая система законно участвует в коммерческих соглашениях разработчика. Иначе обстоит дело с системами, которые играют значительную роль в принятии решений человеком, например, на 2-м и 3-м уровнях автоматизации, поскольку недостатки прозрачности могут быть незаконными сами по себе, независимо от точности системы. Более того, пользователи должны быть полностью информированы о коммерческих соглашениях, связанных с этими системами, и в этих соглашениях должны соблюдаться строгие этические и правовые границы. Поэтому в этом втором сценарии максимальный уровень прозрачности обязателен.

Завершая этот раздел, стоит отметить, что регулирование должно быть направлено на соблюдение минимальных стандартов прозрачности на протяжении всего цикла разработки любой системы ИИ, чтобы облегчить оценку рисков и способствовать прозрачности по замыслу. С этим согласны большинство специалистов

задачу, и допускаем, что иногда она будет вести себя как невежественная машина с огромными слепыми зонами. <...> Иногда эти слепые зоны не имеют значения. Когда мы просим *Google Maps* подсказать дорогу, он моделирует мир как ряд дорог, туннелей и мостов. Он игнорирует здания, потому что они не имеют отношения к задаче» (O´Neil, 2016).

[60] Известно, что реальный мир настолько сложен, что бывают случаи, когда неверная подсказка голосового помощника может оказаться губительной. Однако такие случаи являются исключениями.

[61] Например, в 2022 г. в совместном отчете Университета Вашингтона, Калифорнийского университета в Дэвисе, Калифорнийского университета в Ирвине и Северо-восточного университета указывалось, что данные от голосового помощника *Alexa* незаконно передаются коммерческим партнерам *Amazon*, что противоречит американским законам о конфиденциальности. См. Tuohy, J. P. (2022, Apr. 28). Researchers find Amazon uses Alexa voice data to target you with ads. The Verge. https://www.theverge.com/2022/4/28/23047026/amazon-alexa-voice-data-targeted-ads-research-report.

[62] В нормативных актах о защите персональных данных по всему миру это положение обычно называется «правом на пересмотр автоматизированного принятия решений» или «правом на разъяснение». Например, ст. 22 Европейского общего регламента по защите данных от 2016 г. (GDPR – Regulation 2016/679); ст. 20 Закона Бразилии о защите данных от 2018 г. (LGPD – Federal Law № 13,709/2018); ст. 16 Закона Уругвая № 18,331 от 2008 г. и многих других. Упомянутое положение вызывает много споров. С одной стороны, специалисты отмечают, что оно не дает заинтересованному лицу права знать, как именно работает алгоритм, поскольку это противоречило бы правам интеллектуальной собственности владельца алгоритма: «Описание объяснений в статье 71 не содержит требования открыть „черный ящик“. Понимание внутренней логики алгоритмической системы принятия решений не требуется в явном виде» (Wachter et al., 2018). С другой стороны, некоторые авторы выступают за более широкое право на объяснение, особенно когда речь идет о государственных органах: «...важно, чтобы каждый из вовлеченных субъектов, как государственные органы, так и частные лица, участвующие в процедуре, могли знать механизм работы алгоритма, хотя и по разным причинам: для государственных органов необходимо понять, позволяет ли этот алгоритм законно и справедливо достичь целей, которые должны быть достигнуты с помощью административной процедуры; для гражданина полезно знать, как был сделан административный выбор, чтобы исключить возможность стать жертвой несправедливости с ущербом для основных прав» (Ferrari, 2020).

в юридической сфере. Автор считает, что эти стандарты не обязательно должны быть одинаковыми для всех видов систем. Напротив, для каждой системы требуется свой уровень прозрачности в каждом конкретном случае. Действительно, такие факторы, как различные способы предоставления продуктов и услуг на основе ИИ, цель использования каждой системы, различные уровни вмешательства ИИ в процесс принятия решений, уровень точности, неустранимые риски автоматизируемой деятельности и прозрачность, служат для оценки пригодности ИИ, позволяя сбалансировать интересы и ожидания. Среди этих параметров прозрачность является, несомненно, ключевым, но не единственным фактором.

4.3. Запрет на автономность ИИ или обязательность участия человека («оператор в контуре управления»)

Даже если система на основе ИИ достигает высокой точности и прозрачности, как в примерах, приведенных в разд. 4.1 и 4.2, существуют условия, в которых из-за юридических или социальных причин может быть наложен (посредством регулирования[63]) запрет на использование этой системы или как минимум введено требование, чтобы она обеспечивала возможность значимого вмешательства человека при необходимости отмены решения системы[64]. Полный запрет на использование ИИ в том или ином контексте называется запретом на автономность ИИ. Альтернативный вариант, когда использование системы разрешено при условии возможности значимого вмешательства человека, получил название «система с оператором в контуре управления» (*human in the loop*, *HITL*).

Запрет на автономность ИИ – радикальная мера, допустимая только в тех случаях, когда использование системы на базе ИИ по своей сути несовместимо с фундаментальными ценностями, т. е. система по своей природе противоречит правам человека, независимо от цели ее использования. Классический пример – смертоносное автономное оружие (*Lethal Autonomous Weapon*, *LAW*[65]), система третьего уровня для военных целей[66], например, беспилотник, ракета, транспортное средство или другой вид воплощенного ИИ[67], оснащенный оружием, который после активации сам начинает поиск цели и атакует, в итоге приводя к серьезным ранениям или смерти[68]. Именно такой ИИ больше всего напоминает «машины зла», изображенные в фильмах, подобных «Терминатору».

Хотя в научной литературе[69] нет единого мнения по этому вопросу, Министерство обороны США считает основным признаком автономного оружия способность после активации выбирать, отслеживать и поражать

---

[63] «…мы можем выделить два типа определений регулирования, которые пересекаются в разных дисциплинах, – основанное на сущности и основанное на модели. <…> Сущностное определение направлено на то, чтобы охватить минимальную сущность понятия. Это классическое определение в том смысле, что оно включает в себя только те элементы, без которых регулирование теряет свою идентичность <…>. Соответственно, регулирование можно определить как намеренное вмешательство в деятельность целевой группы населения. Вмешательство, о котором идет речь в этом определении, может быть прямым и/или косвенным, деятельность может быть экономической и/или неэкономической, регулятор может быть государственным или частным субъектом, а регулируемый субъект также может быть государственным или частным. <…> Наше определение, основанное на модели, не менее всеобъемлющее, чем определение, основанное на сущности, но оно дает представление о тех проявлениях, которые в основном интересуют исследователей регулирования и которые мы считаем более важными для концепции в целом. Мы придаем большее значение различиям и определяем регулирование как намеренное вмешательство в деятельность целевой группы населения, где вмешательство, как правило, является прямым (т. е. включает обязательное установление стандартов, мониторинг и санкции) и осуществляется субъектами государственного сектора в отношении экономической деятельности субъектов частного сектора» (Kopp & Lodge, 2015). См. также (Parentoni, 2020).

[64] Приблизительный перевод с португальского: «…даже если данная автономная система достигнет приемлемого уровня количества попаданий и промахов, будет ли этически законным делегировать определенные типы решений полностью автоматизированным системам, без соответствующего вмешательства человека?» (Wimmer & Doneda, 2021).

[65] Также известное как автономная система вооружений.

[66] В соответствии с классификацией, представленной в разд. 3.

[67] Согласно определению воплощенного ИИ, приведенному в разд. 1.

[68] «После первоначального запуска или активации человеком-оператором сама система вооружения с помощью своих сенсоров, программного обеспечения и оружия берет на себя функцию поиска цели, которая в противном случае контролировалась бы человеком. Это рабочее определение относится к любой системе вооружения, которая может самостоятельно выбирать и атаковать цели, включая некоторые существующие виды оружия и потенциальные будущие системы» (Davison, 2018).

[69] «Само понятие автономных систем вооружений не имеет четкого международного определения…» (Reeves et al., 2021). «В международном правовом и политическом сообществе существует множество определений автономных систем вооружений, однако государства не могут договориться об общем определении» (Cherry & Johnson, 2021).

цели без дальнейшего вмешательства человека[70]. В Китае был предложен более подробный список из пяти основных признаков:

«1) смертоносность, что означает достаточную мощность заряда и наличие средств для причинения смерти;

2) автономность, что означает отсутствие вмешательства и контроля со стороны человека в течение всего процесса выполнения задачи;

3) невозможность прекращения действия, означающая, что после запуска нет возможности остановить работу устройства;

4) неизбирательное действие, означающее, что устройство будет выполнять задачу по убийству и нанесению увечий независимо от условий, сценариев и целей;

5) эволюция, означающая, что благодаря взаимодействию с окружающей средой устройство может автономно обучаться, расширять свои функции и возможности таким образом, чтобы превзойти ожидания человека»[71].

Независимо от принятого определения, решающим здесь является тот факт, что решение о лишении жизни, даже во время войны, должно приниматься людьми, и только людьми. Этому есть как этические, так и юридические причины[72]. Поэтому смертоносное автономное оружие представляет пример системы, на которую должен распространяться запрет на автономность ИИ[73]. Специалисты в области гуманитарного права выступают за заключение международного договора о внесении поправок в Конвенцию ООН об обычных вооружениях, чтобы решить этот вопрос на глобальном уровне[74]. Общественные организации, такие как *Human Rights Watch*, также поддерживают это предложение[75].

Другие примеры автономности ИИ еще более спорны. Один из них – решение суда, подразумевающее, что судья-человек может быть полностью заменен системой ИИ, которая будет вести процесс и принимать все решения[76]. Lawrence Solum идет еще дальше, предлагая обсудить гипотетическое «право искусственного интеллекта»[77], т. е. правовую систему, управляемую ИИ. Учитывая, что в судебном процессе фигурируют субъективные по своей сути суждения и что прецедентное право должно создаваться людьми (а не машинами), мы считаем, что в этой области запрет на автономность ИИ оправдан.

Менее спорный пример относится к сфере здравоохранения – это автономные роботизированные операции, т. е. операции, полностью проводимые роботами, без значимого вмешательства человека. В этом контексте запрет на автономную работу ИИ имеет смысл, особенно когда показатели точности низки или когда нет

---

[70] The United States of America. Department of Defense – DoD Directive 3000.09. https://www.hsdl.org/?abstract&did=726163

[71] China. CCW/GGE.1/2018/WP.7. https://undocs.org/Home/Mobile?FinalSymbol=CCW%2FGGE.1%2F2018%2FWP.7&Language=E&DeviceType=Desktop&LangRequested=False

[72] «Защитники прав человека и специалисты в области компьютерных наук утверждают, что системы „оружия с оператором в контуре управления“ нарушают международное право, поскольку системы ИИ не могут адекватно учитывать нормы различения („которые требуют, чтобы вооруженные силы проводили различие между комбатантами и некомбатантами“) и пропорциональности» (Citron & Pasquale, 2014).

[73] При всем уважении автор выражает свое несогласие с лицами, которые поддерживают использование этих систем, например: «Я хочу определенно заявить <...>, что существует множество условий, при которых использование автономной системы вооружений уместно не только с рациональной и стратегической, но и с моральной точки зрения» (Macintosh, 2021).

[74] The United Nations. Background on LAWS in the CCW. https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/

[75] «Большинство государств – участников Конвенции о негуманном оружии и движения „Остановить роботов-убийц“, глобальной коалиции гражданского общества, координируемой организацией *Human Rights Watch*, призывают к переговорам о юридически обязывающем документе, запрещающем или ограничивающем смертоносные автономные системы вооружений. Движение выступает за заключение соглашения, обеспечивающего значимый человеческий контроль над применением силы и запрещающего системы оружия, которые действуют без такого контроля» (They published a manifesto called “Stop Killer Robots”: Human Rights Watch. New Weapons, Proven Precedent – Elements of and Models for a Treaty on Killer Robots. https://www.hrw.org/report/2020/10/20/new-weapons-proven-precedent/elements-and-models-treaty-killer-robots).

[76] «В этом смысле ИИ будет обучен выполнению конкретной задачи по обеспечению принятия судебных решений» (Fortes et al., 2022).

[77] «В данной статье рассматривается серия мысленных экспериментов, которые постулируют существование „права на основе искусственного интеллекта“. Такая правовая система определяется как система, обладающая тремя функциональными возможностями: 1. Способность генерировать правовые нормы. 2. Способность применять генерируемые правовые нормы. 3. Способность использовать глубокое обучение для модификации генерируемых правовых норм. <...> Делегирование функции законотворчества искусственному интеллекту качественно отличается от любого известного мне в настоящее время использования искусственного интеллекта» (Solum, 2019).

достаточной прозрачности в отношении того, как работает система[78]. В данном исследовании не рассматриваются эти и другие возможные случаи запрета на автономность ИИ, поскольку они потребовали бы более глубокого анализа, выходящего за рамки нашего исследования.

Кратко описав контекст запрета на автономность ИИ, рассмотрим его главную альтернативу, известную как «система с оператором в контуре управления» (*human in the loop*, *HITL*).

Сегодня люди принимают непосредственное участие в создании самого алгоритма, т. е. в первоначальном программировании[79]. В этом смысле в контуре управления всегда будет присутствовать человек, поскольку, независимо от результата, полученного ИИ, ему будет предшествовать определенная деятельность человека. В данном исследовании *HITL* понимается иначе. По сути, *HITL* – это возможность человека наблюдать за системой ИИ, в любой момент приостановить работу системы или отменить ее решения. В этом и состоит обеспечение значимого вмешательства человека в работу системы во время или после ее функционирования[80]. Различают также аспект мониторинга – *human on the loop* – и аспект вмешательства – *human in the loop*[81]. Однако наиболее употребительное выражение, охватывающее оба значения, – «оператор в контуре управления» (*human in the loop*, *HITL*). В конечном счете важно то, что окончательное решение, когда это необходимо, остается за человеком.

Альтернативой запрета системы является ее использование при условии наличия *HITL*. Например, при принятии необратимых или труднообратимых решений, таких как прием на работу или увольнение сотрудника, поскольку здесь требуется сбалансировать факторы как объективные (экономическое положение компании, прогулы работника, производительность труда), так и субъективные (социальное окружение, семейное положение и состояние здоровья сотрудника и т. д.). Такой баланс требует «человеческого подхода» в принятии решения. Поэтому *HITL* является обязательным как по этическим, так и по юридическим причинам. Принятый в 1997 г. Международной организацией труда Кодекс практических мер по защите персональных данных работников предписывает, что прием или увольнение работников должны осуществляться только при условии значимого участия человека[82]. Недавно Комитет Совета Европы призвал страны-члены запретить использование систем распознавания лиц без должного участия человека[83].

Выше уже упоминалась роботизированная хирургия как возможный пример замены запрета автономности ИИ на *HITL*. Учитывая, что сбои в работе ИИ здесь чреваты трагическим исходом, как, например, в деле *Mracek vs Bryn Mawr Hospital*[84], осмысленное вмешательство человека может решить вопрос жизни и смерти. Поэтому регулирование должно предусматривать *HITL*, а не запрещать использование автономных хирургических роботов третьего уровня.

Третий пример *HITL* – область административных решений, т. е. решений, принимаемых органами государственного управления при осуществлении ими своих институциональных полномочий. Как известно, конституция многих стран требует, чтобы такого рода решения были направлены на удовлетворение обще-

---

[78] Как показано в разд. 4.1 и 4.2.

[79] По крайней мере, на некоторое время, пока эволюционные алгоритмы не будут способны развиваться самостоятельно.

[80] Несмотря на то, что для правильной работы HITL должна быть реализована согласно проекту.

[81] «При контролируемом автономном управлении, или “системе с оператором в контуре управления”, машина способна принимать решения и действовать самостоятельно, но пользователь может наблюдать за ее действиями и при необходимости останавливать их. Примером такой системы является контролируемая автономная роботизированная хирургия. При полной автономии система может принимать решения и действовать без вмешательства человека. Человек не находится в контуре управления, поскольку машина работает без обратной связи с пользователем. Примером полностью автономной системы является пылесос Roomba» (Cherry & Johnson, 2021).

[82] «5.5. Автоматизированные процессы не освобождают работодателя от необходимости ознакомиться со всеми данными, необходимыми для правильной оценки результатов. Таким образом, кодекс не поддерживает чисто механический процесс принятия решений, но делает выбор в пользу четко индивидуализированной оценки работников» (The International Labour Organization. Code of Practice on the Protection of Workers Personal Data. https://www.ilo.org/global/topics/safety-and-health-at-work/normative-instruments/code-of-practice/WCMS_107797/lang--en/index.htm).

[83] «Уровень вмешательства систем распознавания лиц и связанного с ним нарушения прав на неприкосновенность частной жизни и защиту данных будет варьироваться в зависимости от конкретной ситуации их использования, и будут случаи, когда национальное законодательство будет строго ограничивать или даже полностью запрещать его, если демократический процесс приведет к такому решению» (European Union. (2021, Jan. 28). Council of Europe. Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. Brussels. http://rm.coe.int/0900001680a134f3).

[84] Как описано в разд. 1.

ственных интересов и их принятие было обосновано государственным органом. Поэтому, даже если и возможно использовать искусственный интеллект для принятия решений, по-прежнему необходимо обеспечить участие человека, чтобы устранять сбои в системе, способные нанести ущерб гражданам или государственному управлению как таковому[85].

Четвертый, крайне спорный, пример – это модерация контента в социальных сетях, таких как *Facebook*, *Instagram* или *Twitter*[86]. Из-за большого объема данных эта модерация уже сейчас осуществляется на основе алгоритмов, в соответствии с условиями использования каждой платформы. Однако как должно происходить вмешательство человека в случае сбоя – например, когда блокируется легитимный контент? Эта тема обсуждается во всем мире и имеет явную политическую составляющую[87].

Еще один сбой произошел с транспортным приложением *Uber* в 2021 г. Чтобы проверить, является ли человек, пытающийся войти в систему, зарегистрированным водителем, компания использовала систему распознавания лиц, которая требовала отправить свое фото в режиме реального времени. Однако система не смогла распознать фотографии цветных людей. Только после вмешательства сотрудника *Uber* пострадавшие водители смогли войти в систему[88].

Наконец, существует область, где *HITL* уже активно используется, – это защита персональных данных[89]. Так, в соответствующем европейском регламенте (*GDPR*) этот вопрос рассматривается следующим образом:

> «Статья 22. Автоматизированное принятие индивидуальных решений, включая профилирование.
> 1. Субъект данных имеет право не подчиняться решению, основанному исключительно на автоматизированной обработке данных, включая профилирование, которое влечет за собой юридические последствия в отношении его или аналогичным образом существенно влияет на него.
> 2. Пункт 1 не применяется, если указанное решение:
> (a) необходимо для заключения или исполнения договора между субъектом данных и контролером данных;
> (b) санкционировано законодательством Европейского союза или государства-члена, которому подчиняется контролер и которое также устанавливает соответствующие меры по защите прав, свобод и законных интересов субъекта данных; или
> (c) основывается на явном согласии субъекта данных.
> 3. В случаях, указанных в пунктах (a) и (c) параграфа 2, контролер данных должен принять соответствующие меры для защиты прав, свобод и законных интересов субъекта данных, по меньшей мере, права на вмешательство человека-контролера, на выражение своей точки зрения и на оспаривание решения»[90].

Аналогичное положение содержится и в Директиве № 2016/680, касающейся «защиты физических лиц в отношении (...) предотвращения, расследования, обнаружения или преследования уголовных преступлений или исполнения уголовных наказаний»[91], а также в других нормативных актах.

---

[85] Более глубокое обсуждение ИИ в государственном секторе см. (Ferrari, 2020).

[86] *Facebook*, *Instagram* – соцсети, принадлежащие *Meta* – организации, признанной экстремистской, деятельность которой запрещена на территории Российской Федерации. *Twitter* – социальная сеть, заблокированная на территории Российской Федерации за распространение незаконной информации.

[87] Хорошим примером служит Наблюдательный совет *Facebook* (соцсеть принадлежит *Meta* – организации, признанной экстремистской, деятельность которой запрещена на территории Российской Федерации.) Созданный в 2020 г., он объединяет известных международных экспертов с разным опытом работы и разными культурами. Его задача состоит в том, чтобы: «[использовать] свое независимое суждение для поддержки права людей на свободу выражения мнений и обеспечения надлежащего соблюдения этих прав. Решения совета о поддержке или отмене решений *Facebook* по контенту имеют обязательную силу, т. е. *Facebook* обязан их выполнять, если это не противоречит закону». См. подробнее: https://oversightboard.com/

[88] Некоторые профсоюзы Великобритании подали в суд на *Uber* за дискриминацию с помощью алгоритма. См.: https://www.business-humanrights.org/en/latest-news/uk-drivers-couriers-sue-uber-over-allegedly-racist-facial-recognition-checks/

[89] Приблизительный перевод с португальского: «...именно в области защиты персональных данных можно более четко проследить попытки внедрить элементы участия человека в автоматическое принятие решений» (Wimmer & Doneda, 2021).

[90] European Union. (2016, Apr. 27). European Parliament. Regulation nº 2016/679. Brussels. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1528874672298&uri=CELEX%3A32016R0679 . В Бразилии ст. 20 Закона о защите данных (Lei 13,709/2018, LGPD) содержит аналогичное положение. Однако в ходе законодательного процесса Конгресс убрал положение о «пересмотре физическим лицом», открыв возможность для обсуждения пересмотра автоматизированного решения программным обеспечением, без участия оператора в контуре управления.

[91] European Union. (2016, Apr. 27). European Parliament. Directive nº 2016/680. Brussels. https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A32016L0680

Рассмотрев контекст понятия «система с оператором в контуре управления», прокомментируем различные точки зрения на него. С одной стороны, некоторые авторы указывают на то, что взаимодействие человека и машины имеет первостепенное значение для достижения максимального эффекта от ИИ[92], как, например, в приведенных ранее примерах. Во всех этих случаях обеспечение значимого вмешательства человека в работу системы имеет решающее значение для нахождения баланса между использованием ИИ и соблюдением базовых ценностей, охраняемых законом. С другой стороны, ученые предупреждают, что большой объем данных и скорость их обработки превосходят возможности человеческого мозга, что делает вмешательство человека медленным, неэффективным или даже невозможным[93]. По их мнению, *HITL* противоречит основной цели автоматизации с помощью ИИ – достижению более быстрых и качественных результатов, поскольку участие человека замедляет работу системы и снижает ее эффективность[94]. На самом деле обе линии рассуждений верны. За использование *HITL*, несомненно, приходится чем-то платить. Этот метод позволяет избежать запрета на автономность ИИ в определенных ситуациях. Поэтому он не должен быть обязательным для любой системы ИИ, независимо от контекста и цели ее использования. Как будет показано ниже, большинство систем должны быть полностью автоматизированы без *HITL*.

Действительно, и запрет на автономность ИИ, и «система с оператором в контуре управления» должны рассматриваться как исключения и быть ограничены системами, принимающими решения с высокими ставками или высокими рисками. Хотя люди иногда предпочитают, чтобы значимые для них решения принимал человек[95], стандартное использование ИИ наиболее эффективно при полной автоматизации[96]. Ralf Poscher так резюмирует это положение:

> «...Перенос внимания с абстрактной опасности на конкретное, материальное выражение основных прав человека позволяет обсудить пороговые значения. В аналоговом мире закон также не реагирует на каждый риск, возможный в современном обществе. Не каждый возможный риск превышает порог нарушения основных прав. В жизни существуют риски, которые юридически не являются значимыми. <...>
>
> Порог для рисков повседневной жизни существует в аналоговом мире и должен существовать и в цифровом мире. В нашем цифровом обществе мы должны осознать, что существует – и, вероятно, постоянно изменяется – порог повседневных цифровых рисков, которые не являются нарушением основных прав, даже при хранении или обработке персональных данных. Для технологий искусственного интеллекта это означает, что их разработка и внедрение могут оставаться ниже порога риска для повседневной цифровой жизни» (Poscher, 2021).

Исходя из вышесказанного, зададим следующий вопрос: каковы пороговые значения для того, чтобы система ИИ подчинялась оператору в контуре управления? В научной литературе нет единого мнения по этому

---

[92] «…создание совместных коллективов из людей и машин дает необычайные преимущества, примером чему могут служить победители открытых шахматных турниров. В целях оптимизации человеко-машинные системы следует воспринимать как социотехнические системы, которые используют вклад человека и социума в автоматизацию и наоборот. <...> Подобно тому, как люди работают с другими людьми (при поддержке автоматизированных и неавтоматизированных инструментов) для выполнения многочисленных задач и достижения различных целей в своей повседневной жизни, люди будут работать в обязательном содружестве с роботизированными и интеллектуальными системами. Участники таких циклов будут тесно взаимодействовать между собой» (Ambrose, 2014).

[93] «Многие считают, что лучший способ обеспечить честность или справедливость – это внедрить человека в процесс принятия решений, возможно, с правом вето, чтобы отменить решение неодушевленного сотрудника. Мы опасаемся, что если мы просто поставим человека на выходе работающей модели, он мало что сможет сделать для ликвидации предвзятости. Человек становится образцом для машины, и его участие может лишь замаскировать проблему. Возможно, есть более эффективные и продуктивные способы обеспечить человеческий надзор на других стадиях процесса» (Lehr & Ohm, 2017).

[94] «Устранение недостатков надлежащей правовой процедуры вряд ли будет заключаться в дополнительном человеческом участии, а скорее в улучшении алгоритмического дизайна» (Huq, 2020).

[95] См., например, D. E. Bambauer и M. Risch: «В исследованиях, включающих репрезентативные сценарии с различными ставками, выясняли, предпочитают ли люди, чтобы результат, влияющий на их благосостояние, определялся алгоритмом или человеком» (Bambauer & Risch, 2021).

[96] Эту точку зрения разделяют другие авторы: «Авторы предлагают рассматривать риск как качество, разграничивающее классы ИИ. <...> Нельзя требовать одинаковых мер регулирования для каждой категории, а некоторые приложения могут вообще не требовать мер регулирования на данном этапе. Принятие регулирующих мер может потребоваться только при повышении уровня риска приложения ИИ» (Guihot et al., 2017).

вопросу[97]. Мы считаем целесообразным решать этот вопрос отдельно в каждом конкретном случае, принимая во внимание следующее: 1) каков средний уровень риска системы ИИ, определенный с помощью надлежащей оценки риска (низкие или даже средние уровни должны быть исключены из *HITL*, если иное не рекомендовано при оценке риска); 2) поставит ли сбой в ИИ под угрозу фундаментальные ценности; 3) является ли решение системы обратимым или необратимым (или, по крайней мере, труднообратимым); 4) может ли определение «верного» или «неверного» решения быть закодировано математически или оно по своей сути субъективно (как, например, при моральных суждениях); 5) каковы общественные и экономические последствия ошибки.

Учитывая эти факторы в каждом конкретном случае, можно достичь баланса между технологическим развитием и инновациями, с одной стороны, и защитой прав человека и предотвращением системных рисков – с другой. Независимо от того, какие критерии используются для оценки необходимости запрета на автономность ИИ или использования *HITL*, крайне важно, чтобы регулирующие органы четко донесли принятые критерии до ключевых заинтересованных сторон, включая разработчиков, пользователей и исследователей.

5. Международная дискуссия по теме исследования

Глобальный ландшафт регулирования ИИ все еще представляет собой лоскутное одеяло инициатив, исходящих из огромного количества источников, таких как передовые страны, компании и международные организации[98]. В этом разделе мы рассмотрим ряд важных источников и покажем, что идеи, развиваемые в нашем исследовании, интересуют не одного лишь автора, а широко обсуждаются во всем мире. В рамках данной работы мы не предполагаем углубленно изучать какой-либо из этих источников, поскольку каждый из них потребовал бы отдельной статьи для полного анализа. Кроме того, существует множество других стран, заслуживающих изучения, таких как Китай, Великобритания, Австралия и Уругвай. Остановимся лишь на некоторых выбранных автором пунктах.

ОЭСР. Организация экономического сотрудничества и развития (далее – ОЭСР) опубликовала два важнейших документа: Принципы ИИ (май 2019 г.) и Рамочную программу классификации систем ИИ (2022 г.). Документ «Принципы ИИ»[99] считается основополагающим. Это первый межгосударственный нормативный документ в этой сфере, цель которого – установление «практичных и достаточно гибких стандартов для ИИ, чтобы выдержать испытание временем». Документ включает в себя пять принципов разработки и использования систем ИИ, основанных на ценностях и применяемых как в государственном, так и в частном секторе, независимо от вида ИИ, а также пять рекомендаций для выработки глобальной политики в данной области. Эти принципы таковы: «1) инклюзивный рост, устойчивое развитие и благосостояние; 2) человекоориентированные ценности и справедливость; 3) прозрачность и объяснимость; 4) надежность, безопасность и защищенность; 5) подотчетность». Данное исследование соответствует всем принципам ОЭСР, и обсуждаемые вопросы напрямую связаны с ними, особенно с принципами 3, 4 и 5.

Еще один источник, имеющий огромное значение, – это Рамочная программа классификации систем искусственного интеллекта, принятая в феврале 2022 г.[100] ОЭСР описывает ее как «ориентированный на пользователей основной документ для политиков, регуляторов, законодателей и других лиц для определения характеристик систем ИИ в конкретных проектах и контекстах». Рамочная программа увязывает характеристики систем ИИ с Принципами ИИ ОЭСР; указанные принципы являются первым набором стандартов ИИ, на который правительства обязуются опираться в процессе разработки политических мер, чтобы содействовать

[97] Например, ОЭСР рекомендовала следующие критерии: «Независимо от количества уровней риска и от того, какая организация их предлагает, типичными критериями для определения уровня риска приложения или системы ИИ являются следующие: масштаб, т. е. серьезность негативных последствий (и их вероятность); сфера, т. е. широта применения, например, количество людей, которые подвергаются или будут подвергаться его воздействию; факультативность, т. е. возможность выбора, подвергаться ли воздействию системы ИИ» (OECD. (2022, February 22). OECD Framework for the Classification of AI Systems. https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en).

[98] Например, ОЭСР ведет репозиторий, содержащий «более 700 политических инициатив в области ИИ из 60 стран» (OECD. (2021). National AI policies & strategies. https://oecd.ai/en/dashboards).

[99] Подробное описание каждого принципа см.: OECD. (2019, May). OECD AI Principles overview. https://oecd.ai/en/ai-principles

[100] OECD. (2022, February 22). OECD Framework for the Classification of AI Systems. https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en

инновационному и надежному использованию ИИ»[101]. Между Принципами 2019 г. и Рамочной программой 2022 г. заметна разница в масштабах. В то время как принципы остаются намеренно гибкими, чтобы охватить все виды ИИ (как и ожидалось от основополагающего документа), программа сосредоточивается на конкретных проектах и контекстах[102]. Эти контексты (обозначенные в документе как «направления») следующие: 1) люди и планета Земля; 2) экономический контекст; 3) исходные данные и информация; 4) модель ИИ; 5) задачи и результат. Кроме того, Рамочная программа может применяться как для разработок, так и на практике[103]. В ближайшем будущем ОЭСР планирует дополнить ее за счет анализа большого количества реальных систем и разработать систему показателей для оценки влияния этих документов на реализацию прав человека и благосостояние. Следующим шагом могло бы стать создание системы оценки рисков в соответствии с идеями, развитыми в разд. 5 данной статьи.

ЮНЕСКО. В ноябре 2021 г. Генеральная конференция ЮНЕСКО утвердила Рекомендацию по этике искусственного интеллекта[104], предложив ряд принципов, многие из которых совпадают с принципами ОЭСР, например, справедливость, прозрачность, объяснимость, безопасность и подотчетность. Кроме того, в документе рекомендуется проводить постоянную оценку рисков ИИ. Две части этой рекомендации особенно созвучны идеям, развиваемым в нашем исследовании. Во-первых, речь идет о балансе прозрачности и объяснимости, о поиске практических алгоритмов[105]. Во-вторых, несмотря на отсутствие термина «оператор в контуре управления», подобный механизм, по-видимому, рекомендован для внедрения[106].

Европейский союз. ЕС – один из самых плодотворных источников регулирования в сфере ИИ, поэтому здесь было бы уместно упомянуть многие из их инициатив, однако для краткости мы рассмотрим только две. В феврале 2020 г. был опубликован документ «Искусственный интеллект: Европейский подход к совершенству и доверию»[107]. В нем отмечается, что «нынешний и будущий устойчивый экономический рост Европы и благосостояние общества все больше опираются на ценности, создаваемые за счет информации»[108], а «ИИ является одним из наиболее важных приложений экономики данных»[109]. Цель документа – способствовать развитию и этическому использовании ИИ в рамках Евросоюза на общих основаниях, избежать фрагментации между государствами-членами и укрепить ЕС в качестве глобального лидера в экономике данных. Рекомендуется использовать существующую промышленную инфраструктуру и человеческие ресурсы, а также укреплять сотрудничество между государствами-членами для создания общей нормативной базы в области ИИ. В документе много точек соприкосновения с данным исследованием; упомянем три из них. Во-первых, в нем перечислены некоторые критерии оценки рисков ИИ, такие как ценности, риск использования систем,

---

[101] Там же, p. 6.

[102] Там же, p. 16. «Основная цель – охарактеризовать применение системы ИИ, развернутой в конкретном проекте и контексте, хотя некоторые аспекты относятся и к общим системам ИИ».

[103] Там же, p. 7. «Лабораторное использование ИИ относится к созданию и разработке системы ИИ до ее развертывания. Сюда включаются такие аспекты, как “Входные данные” (например, отбор данных), “Модель ИИ” (например, обучение исходной модели) и “Задачи и результаты” (например, задача персонализации). Это особенно актуально для подходов и требований к управлению рисками на начальном этапе. Полевая работа ИИ относится к использованию и развитию системы после развертывания и включает все аспекты. Сюда входят подходы и требования к управлению рисками по факту».

[104] «В Рекомендациях рассматриваются этические вопросы, связанные с ИИ. Этика ИИ рассматривается как целостная система взаимозависимых ценностей, принципов и действий, которыми могут руководствоваться общества в жизненном цикле систем ИИ, учитывая человеческое достоинство и благополучие в качестве ориентиров для ответственного подхода к известным и неизвестным воздействиям систем ИИ в их взаимодействии с людьми и окружающей средой» (UNESCO. (2021). UNESCO Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000380455).

[105] Там же, p. 17. «Осуществимость: многие алгоритмы ИИ все еще не поддаются объяснению; для других объяснимость создает значительные дополнительные расходы при реализации. Пока полная объяснимость не станет технически возможной с минимальным влиянием на функциональность, будет существовать компромисс между точностью/качеством системы и уровнем ее объяснимости».

[106] Там же, p. 9. «Возможно, иногда людям придется делить контроль с системами ИИ по соображениям эффективности, но решение об уступке контроля в ограниченных контекстах остается за людьми, поскольку системы ИИ должны исследоваться, проектироваться, разрабатываться, внедряться и использоваться для помощи людям в принятии решений и действиях, но никогда не заменять собой конечную ответственность человека».

[107] European Union. (2020, February 19). White Paper on Artificial Intelligence: A European approach to excellence and trust. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

[108] Там же, p. 1.

[109] Там же.

защита потребителей, персональных данных и других основных прав[110]. Во-вторых, в нем подчеркивается, что конкретная цель использования каждой системы имеет решающее значение для определения ожидаемого показателя точности[111]. Подчеркивается, что в некоторых случаях, например в ситуациях с высоким риском или способных нанести вред, приемлемы низкие показатели точности. В-третьих, рекомендуется обязательное участие человека в контуре управления для ИИ с высоким уровнем риска, тогда как приложения со средним и низким уровнями риска остаются вне сферы действия этой нормы[112]. Эти три аспекта полностью совпадают с предложениями нашего исследования.

Основываясь на идеях указанного документа и после широких консультаций с государствами-членами, гражданским обществом, компаниями, работающими в сфере ИИ, учеными, практиками и другими ключевыми заинтересованными сторонами[113], в апреле 2021 г. ЕС опубликовал предложение гармонизированных правил в области искусственного интеллекта под названием «Акт об искусственном интеллекте» («Акт об ИИ») [114], в который в июне 2023 г. были внесены поправки Европейского парламента. В настоящее время это основная инициатива ЕС в области регулирования ИИ, хотя до вступления в силу нового закона еще необходимо предпринять ряд шагов.

Некоторые положения данного закона вызывают немало споров. Например, не определены ситуации с генеративными системами ИИ, такими как *ChatGPT* от *Open AI*, *Bard* от *Google* или *Llama*[115] от *Meta*[116], с запретом на использование систем распознавания лиц в общественных местах[117] или предполагаемыми негативными последствиями для инновационной деятельности и конкуренции в ЕС[118]. Существуют и другие проблемы, однако их углубленный разбор выходит за рамки данного исследования.

Достаточно отметить, что в законе используется подход, основанный на оценке рисков, и рассмотрены многие аспекты, содержащиеся в нашем исследовании. Во-первых, в законе сделана попытка найти баланс между рыночными инновациями и защитой основных прав[119]. Во-вторых, предусмотрены обязательные требования и предварительные меры, ориентированные на системы с высоким уровнем риска, но исключающие системы со средним и низким уровнями риска[120]. В-третьих, в разделе «Запрещенные практики искусственного интеллекта» вводится запрет на автономность ИИ для некоторых систем, таких как кредитный рейтинг в государственном секторе или правоохранительная деятельность на основе дистанционной биометрии в ре-

---

[110] Там же, p. 17

[111] Там же, p. 20. «Обеспечение предоставления четкой информации о возможностях и ограничениях системы ИИ, в частности, о цели, для которой предназначены системы, условиях, при которых они будут функционировать так, как задумано, и ожидаемом уровне точности в достижении указанной цели».

[112] Там же, p. 21. «Человеческий надзор помогает гарантировать, что система ИИ не подорвет автономность человека и не вызовет других негативных последствий. Цель создания надежного, этичного и ориентированного на человека ИИ может быть достигнута только путем обеспечения надлежащего участия людей в работе с высокорискованными приложениями ИИ».

[113]«С 19 февраля по 14 июня 2020 г. проходили онлайн-консультации по Белой книге по искусственному интеллекту, в ходе которых приняли участие 1215 представителей самых разных заинтересованных сторон (граждане – 33 %, бизнес и промышленность – 29 %, гражданское общество – 13 %, научные круги – 13 %, органы государственной власти – 6 %; 84 % материалов поступило из государств-членов, остальные – из-за пределов ЕС)» (Hubert, D. (2021). Initial Appraisal of a European Commission Impact Assessment. https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2021)694212).

[114] European Union. (2021, April 21). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[115] Volpicelli, G. (2023, Mar. 3). ChatGPT broke the EU plan to regulate AI. Politico. https://www.politico.eu/article/eu-plan-regulate-chatgpt-openai-artificial-intelligence-act/

[116] Организация признана экстремистской, ее деятельность запрещена на территории Российской Федерации.

[117] Volpicelli, G. (2023, Jun. 14). Forget ChatGPT: Facial recognition emerges as AI rulebook's make-or-break issue. Politico. https://www.politico.eu/article/facial-recognition-artificial-intelligence-act-ai-issue-european-parliament/?mc_cid=0debf02283&mc_eid=8af43a1913

[118] Weatherbed, J. (2023, Jun. 30). European companies claim the EU's AI Act could 'jeopardise technological sovereignty'. The Verge. https://www.theverge.com/2023/6/30/23779611/eu-ai-act-open-letter-artificial-intelligence-regulation-renault-siemens

[119] Там же, p. 3. «…данное предложение основано на сбалансированном и пропорциональном горизонтальном подходе к регулированию ИИ, который ограничивается минимально необходимыми требованиями для решения рисков и проблем, связанных с ИИ, без неоправданного сдерживания или препятствования технологическому развитию или иного непропорционального увеличения стоимости вывода решений в области ИИ на рынок».

[120] Там же.

альном времени в общественных местах, в связи с «неприемлемым уровнем риска»[121]. В-четвертых, в законе подчеркивается, что цели использования каждой системы и способ ее развертывания являются ключевыми факторами для оценки риска[122], как это подробно описано в разд. 2 и 5 нашего исследования. Наконец, в законе указано, что могут возникать различные проблемы в зависимости от степени участия ИИ в процессе принятия решений человеком, как мы описываем в разд. 4.

Соединенные Штаты Америки. Являясь одной из ведущих стран в области исследований и разработок ИИ, США представляют собой обширный источник нормативных инициатив. В данной статье мы рассмотрим законопроект Сената под названием «Закон о подотчетности алгоритмов» (*Algorithmic Accountability Act*)[123], предложенный в апреле 2019 г. и обновленный в феврале 2022 г. В целом он «требует от компаний оценивать воздействие автоматизированных систем, которые они используют и продают, создает условия прозрачности в отношении того, когда и как используются автоматизированные системы, и дает потребителям возможность делать осознанный выбор в отношении автоматизации принятия важных решений»[124]. Как и другие акты, упомянутые выше, Закон о подотчетности алгоритмов также фокусируется на системах высокого риска и недвусмысленно исключает малые и средние компании, поскольку он распространяется на компании со среднегодовой валовой выручкой более 50 млн долларов США, размером акционерного капитала более 250 млн долларов США или работающие с идентифицирующей информацией о более чем миллионе человек или устройств[125]. Одним из наиболее спорных аспектов законопроекта является его возможное влияние на модерацию контента в социальных сетях, о чем кратко говорится в разд. 5.3 нашего исследования.

Бразилия. Хотя Бразилия не является ведущим игроком в области ИИ, эта тема в настоящее время актуальна в этой стране, и некоторые местные источники заслуживают упоминания. Первые серьезные усилия можно проследить на примере бразильской Стратегии цифровой трансформации (*E-Digital*)[126] от марта 2018 г., которая была направлена на гармонизацию и координацию правительственных проектов по цифровым вопросам в целом. Хотя в *E-Digital* не упоминается ИИ, в нем заложена основа для будущих инициатив.

После выхода *E-Digital* в Бразилии мало что происходило с политикой в области ИИ, пока в сентябре и октябре 2019 г. в Сенат не были представлены два законопроекта. Они соответствовали международным стандартам, таким как Принципы ОЭСР, и касались всех видов ИИ, независимо от сектора экономики, а также от того, используется ли система государственными или частными структурами. По совпадению, оба документа состояли всего из семи статей, т. е. были гораздо короче, чем обычные бразильские законодательные акты. Предполагалось, что они должны дополнять друг друга. Законопроект 5051/2019[127] определял принципы разработки и использования ИИ, а в законопроекте 5691/2019[128] предлагалась национальная стратегия в этой области. Из-за недостаточной технической и политической поддержки[129] они были быстро заменены другим законодательным предложением, о котором речь пойдет ниже.

В Палату депутатов 4 февраля 2020 г. был представлен Законопроект 21/2020[130], более подробный и технически более обоснованный, чем законопроекты Сената, и направленный на их замену. Он в целом соответствует мировым стандартам, предусматривая, что использование ИИ должно основываться на уважении прав человека и демократических ценностей, равенстве, недискриминации, плюрализме, прозрачности, автономии и конфиденциальности данных. Законопроект 21/2020 также вводит обязательную оценку влияния

---

[121] Там же, p. 12.

[122] Там же, p. 13. «…отнесение к категории высокого риска зависит не только от функции, выполняемой системой искусственного интеллекта, но и от конкретной цели и условий, в которых эта система используется».

[123] The United States of America. (2022). Senate – The Algorithmic Accountability Act. https://www.congress.gov/bill/117th-congress/house-bill/6580/text?r=2&s=1

[124] Из резюме законопроекта Сената.

[125] Там же, p. 3.

[126] Presidency of the Republic. (2018, March). Decree № 9319/2018. http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/decreto/D9319.htm

[127] Federal Senate. (2019). Senate Bill № 5051/2019. https://www25.senado.leg.br/web/atividade/materias/-/materia/138790

[128] Federal Senate. (2019). Senate Bill № 5691/2019. https://www25.senado.leg.br/web/atividade/materias/-/materia/139586

[129] Для детального анализа Закона 5051/2019 см. (Parentoni et al., 2020).

[130] Federal Senate. (2020). Chamber of Deputies Bill PL 21/2020. https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2236340

ИИ. Однако он имеет три серьезных недостатка. Во-первых, в нем не используется подход, основанный на оценке рисков, как в упомянутых выше международных источниках. Во-вторых, законопроект не исключает ответственности малых и средних предприятий, что могло бы стимулировать инновации. В-третьих, в нем недостаточно разграничены два типа агентов ИИ – «разработчики» и «операторы». К первым относятся организации, задействованные в планировании, проектировании, сборе и обработке данных и создании модели ИИ, а также в ее проверке и валидации, а ко вторым – организации, участвующие в мониторинге и эксплуатации систем ИИ. Поскольку законопроект устанавливает различные правила ответственности для каждого типа агентов, их точное определение крайне важно.

При этом исполнительная власть попыталась вернуть себе ведущую роль в управлении ИИ, воспользовавшись задержкой в законодательном процессе и разработав собственную стратегию, которая была опубликована в апреле 2021 г. Бразильская стратегия в области ИИ[131] включает в себя девять основных направлений, которые сгруппированы по трем горизонтальным и шести вертикальным осям. К трем горизонтальным (или тематическим) осям относятся: (*i*) законодательство, регулирование и этическое использование; (*ii*) управление ИИ и (*iii*) международные аспекты. Шесть вертикальных (или прикладных) осей следующие: (*i*) образование; (*ii*) рабочая сила и обучение; (*iii*) НИОКР и предпринимательство; (*iv*) применение в производственных секторах; (*v*) применение в правительстве и (*vi*) общественная безопасность. Основная критика заключается в том, что стратегия слишком абстрактна, вплоть до невозможности ее применения для выработки практических мер.

На основе этих инициатив в мае 2023 г. был представлен законопроект Сената 2338/2023[132], который в настоящее время является основным предложением Бразилии в этой области. Этот законопроект в значительной степени основан на Законе ЕС об искусственном интеллекте[133], в котором применяется подход оценки рисков (ст. 13, 18 и 20). Он содержит некоторые моменты, рассмотренные в данном исследовании, например, обязательное участие человека в принятии решений ИИ, которые могут привести к «значимым юридическим последствиям» или «существенно затронуть интересы человека», если вмешательство человека не является «невозможным» (ст. 10). В нем также предусматриваются максимальные показатели точности и прозрачности, доступные для медицинских приложений ИИ, учитывая связанные с ними риски и возможность ущемления основных прав в случае сбоя в работе (ст. 17, IX). Кроме того, сенатский законопроект вводит обязательную оценку воздействия ИИ, что соответствует нормам США и ЕС.

Однако в сенатском законопроекте заложен ошибочный подход о том, что системы ИИ всегда должны обеспечивать высокую точность и прозрачность, во много раз превышающую человеческие стандарты, независимо от контекста. Он просто не признает множество ситуаций, описанных в данном исследовании, в которых более низкие показатели не только оправданы, но и необходимы.

Итак, обзор некоторых важных глобальных инициатив по регулированию ИИ показывает, что положения, разработанные в рамках данного исследования, не являются лишь частным мнением автора, но находят подтверждение в многочисленных международных источниках.

Заключение

Как мы видим, существует несоответствие между тем, что ИИ в настоящее время может дать человечеству, и тем, что некоторые ожидают от него. В связи с этим возникает главный вопрос данного исследования: чего мы должны ожидать от ИИ? Автор попытался дать научный и обоснованный ответ на этот вопрос, который должен способствовать пересмотру ожиданий в отношении систем ИИ как в науке, так и в практике, учитывая их текущее развитие.

---

[131] Administrative Rule 4,617/2021. (2021). PORTARIA GM Nº 4.617, DE 6 DE ABRIL DE 2021 (*) – DOU – Imprensa Nacional. Portal da Imprensa Nacional do Brasil. Diário Oficial da União. https://www.in.gov.br/en/web/dou/-/portaria-gm-n-4.617-de-6-de-abril-de-2021-*-313212172

[132] Federal Senate. (2023). Senate Bill № 2338/2023. https://www25.senado.leg.br/web/atividade/materias/-/materia/157233

[133] Настолько, что определение системы искусственного интеллекта в законопроекте частично копирует первую версию Закона ЕС об искусственном интеллекте.

Чтобы ответить на этот вопрос, в данном исследовании был проведен систематический анализ ряда ключевых факторов. Во-первых, это то, что ИИ не представляет собой единую концепцию. Напротив, он охватывает широкий спектр приложений в различных секторах рынка и основан на огромном количестве методов и моделей, используемых для самых разных целей. Поэтому оценка должна производиться в каждом конкретном случае с учетом конкретной системы и стратегии разработчиков и ретейлеров по ее внедрению на рынок. Ведь разные цели и стратегии могут быть связаны с разными видами рисков. Во-вторых, при оценке необходимо определить степень участия ИИ в процессе принятия решений человеком, поскольку каждый уровень создает различные проблемы и риски. Для этого автор предлагает трехуровневую категоризацию в качестве инструмента для оценки большинства ситуаций.

В исследовании также рассматриваются три основных критерия оценки системы ИИ: 1) уровень точности; 2) уровень прозрачности/объяснимости и 3) особые ситуации вмешательства регулятора, чтобы запретить использование некоторых систем из-за неприемлемых рисков (запрет на автономность ИИ) или, по крайней мере, предусмотреть участие человека, способного при необходимости отменить решение системы (оператор в контуре управления). При этом очевидно, что эти критерии должны оцениваться совместно, поскольку они противоречат друг другу. Мы также подчеркиваем, что чрезмерный акцент только на одном из них (обычно на точности или прозрачности) может не только поставить под угрозу инновации, но и снизить конкурентоспособность и благосостояние, о чем свидетельствуют упомянутые в тексте примеры, признанные также ОЭСР и другими международными источниками.

Наконец, понимая, что ИИ постоянно меняется и развивается, в данном исследовании мы попытались представить более абстрактные и проверенные временем критерии, позволяющие определить, чего следует ожидать от ИИ в каждом контексте. Безусловно, в этой области есть место для дальнейших разработок, и наше исследование вносит свой вклад в обсуждение этой актуальной темы.

Список литературы / References

Ambrose, M. L. (2014). Regulating the Loop: Ironies of Automation Law. In *WeRobot 2014*. Miami: University of Miami. https://perma.cc/E9KL-CSNK

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Balkin, J. M. (2017). The Three Laws of Robotics in the Age of Big Data. *Yale Law School Research Paper*, *592*, 01–28.

Bambauer, D. E., & Risch, M. (2021). Worse Than Human? *Arizona State Law Journal*, *53*(4), 1091–1151. https://ssrn.com/abstract=3897126

Benkler, Y. (2019, May). Don't let industry write the rules for AI. *Nature*, 569, 01. https://doi.org/10.1038/d41586-019-01413-1

Black, Ju., & Murray, A. (2019). Regulating AI and Machine Learning: Setting the Regulatory Agenda. *European Journal of Law and Technology – EJLT*, *10*(3), 1–21. https://www.ejlt.org/index.php/ejlt/article/view/722

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In *Robustness in Scientific Model Building* (pp. 201–236). Academic Press, Inc. https://gwern.net/doc/statistics/decision/1979-box.pdf

Brynjolfsson, E., & McAfee, A. (2016). *The Second Machine Age*: *Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: Norton & Company, 2016. https://edisciplinas.usp.br/pluginfile.php/4312922/mod_resource/content/2/Erik%20-%20The%20Second%20Machine%20Age.pdf

Burk, Dan L. (2020, February 13). Algorithmic Legal Metrics. *Notre Dame Law Review*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3537337

Calo, R. (2017). Artificial Intelligence Policy: A Primer and Roadmap. *University of Washington Research Paper*, 01–28. http://dx.doi.org/10.2139/ssrn.3015350

Cansian, A. C. de M. (2022). *Aspectos Jurídicos Relevantes da Internet das Coisas (IoT): Segurança e Proteção de Dados*. Tese (Doutorado). Universidade de São Paulo, São Paulo. (In Portuguese).

Cao, Sean S., Jiang, Wei, Wang, Junbo L., & Yang, Baozhong. (2021, May). From Man vs. Machine to Man Machine: The Art and AI of Stock Analyses. *Columbia Business School Research Paper*. http://dx.doi.org/10.2139/ssrn.3840538

Cardoso, R. L., Meira Jr, W., Almeida, V., & Zaki, M. J. (2019). A Framework for Benchmarking Discrimination-Aware Models in Machine Learning. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA* (pp. 437–444). ACM, New York, NY, USA. https://doi.org/10.1145/3306618.3314262

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and Engineering Ethics*, *24*(2), 505–528. https://doi.org/10.1007/s11948-017-9901-7

Cherry, J., & Johnson, D. (2021). Maintaining command and control (C2) of lethal autonomous weapon Systems: Legal and policy considerations. *Southwestern Journal of International Law*, *27*(1), 1–27. https://www.swlaw.edu/sites/default/files/2021-03/1.%20Cherry%20%5Bp.1-27%5D.pdf

Citron, D. K., & Pasquale, F. A. (2014). The Scored Society: Due Process for Automated Predictions. *University of Maryland School of Law Research Paper*, 2014-8, 01–34. https://ssrn.com/abstract=2376209

Damasio, A. (2010). *Self Comes to Mind*: *Constructing the Conscious Brain*. New York: Pantheon.

Darling, K. (2012). Extending Legal Protection to Social Robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In *WeRobot 2012*. Miami: University of Miami. http://dx.doi.org/10.2139/ssrn.2044797

Davison, N. (2018). A legal perspective: Autonomous weapon systems under international humanitarian law. *UNODA Occasional Papers*, 30, 01–14. https://doi.org/10.18356/29a571ba-en

Dreyfus, H. L. (1965). Alchemy and Artificial Intelligence. *Rand Corporation Report Papers*, 01-90.

Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a 'Right to an Explanation' is probably not the remedy you are looking for. *Duke Law & Technology Review*, *16*(01), 18–84. https://doi.org/10.31228/osf.io/97upg

Ferrari, M. (2020). L'uso degli algoritmi nella attività amministrativa discrezionale. *Il Diritto Degli Affari*, 1, 58–82. (In Portuguese). https://hdl.handle.net/10281/272405

Fortes, P. R. B., Baquero, P. M., & Amariles, D. R. (2022). Artificial Intelligence Risks and Algorithmic Regulation. *European Journal of Risk Regulation*, *13*(3), 357–372. https://doi.org/10.1017/err.2022.14

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Information Systems Quarterly – MISQ*, *45*(3), 1527–1556. https://doi.org/10.25300/misq/2021/16553

Gardner, H. (1999). *Intelligence Reframed: Multiple Intelligences for the 21st Century*. New York: Basic Books.

Guihot, M., Matthew, A. F., & Suzor, N. P. (2017). Nudging robots: Innovative solutions to regulate artificial intelligence. *Vanderbilt Journal of Entertainment and Technology Law*, *20*(2), 385–456. https://doi.org/10.31228/osf.io/5at2f

Harari, Yu. N. (2016). *Homo Deus: A Brief History of Tomorrow*. New York: Harper Collins.

Huq, A. Z. (2020). Constitutional Rights in the Machine-Learning State. *Cornell Law Review*, *105*(7), 1875–1954. https://doi.org/10.2139/ssrn.3613282

Jain, A. K., Ross, A. A., & Nandakumar, K. (2011). *Introduction to Biometrics*. New York: Springer. https://doi.org/10.1007/978-0-387-77326-1

Kaplan, J. (2016). *Artificial Intelligence: What everyone needs to know*. Oxford: Oxford University Press.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, *10*(1), 113–174. https://doi.org/10.1093/jla/laz001

Kopp, Ch., & Lodge, M. (2015). What is regulation? An interdisciplinary concept analysis. *Regulation & Governance*, *11*(1), 1–43. https://doi.org/10.1111/rego.12094

La Diega Guido Noto. (2018). Against the Dehumanisation of Decision-Making: Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law – JIPITEC*, *9*(1), 3–34. https://doi.org/10.31228/osf.io/s2jnk

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*, 1–9. https://doi.org/10.1038/s41467-019-08987-4

Lehr, D., & Ohm, P. (2017). Playing with the Data: What Legal Scholars Should Learn About Machine Learning. *University of California Davis Law Review*, *51*(2), 653–717. https://lawreview.law.ucdavis.edu/sites/g/files/dgvnsk15026/files/media/documents/51-2_Lehr_Ohm.pdf

Lima, C. R. P. de. (2020). *Sistemas de Responsabilidade Civil para Carros Autônomos*. 2020. 422 f. Tese – Faculdade de Direito. Universidade de São Paulo, Ribeirão Preto.

Lima, F. G. M. De, Medeiros, F. L. L., & Passaro, A. (2021). Decision Support System for Unmanned Combat Air Vehicle in Beyond Visual Range Air Combat Based on Artificial Neural Networks. *Journal of Aerospace Technology and Management*, *13*, 1–18. https://doi.org/10.1590/jatm.v13.1228

Lopez, T. A. (2010). *Princípio da Precaução e Evolução da Responsabilidade Civil*. São Paulo: Quartier Latin.

Macintosh, D. (2021). Fire and Forget: A Moral Defense of the Use of Autonomous Weapons in War and Peace. In J. Galliott, D. Macintosh, J. D. Ohlin (Eds.), *Lethal Autonomous Weapons: Re-Examining the Law and Ethics of Robotic Warfare* (pp. 9–23). Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780197546048.003.0002

Moravec, H. (1999). *Robot: Mere Machine to Transcendent Mind*. Oxford: Oxford University Press.

Nicolelis, M. A. L., & Cicurel, R. (2015). *The Relativistic Brain: How it Works and why it cannot be simulated by a Turing Machine*. Montreux: Kios Press.

Nilsson, N. J. (2010). *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge: Cambridge University Press. https://ai.stanford.edu/~nilsson/QAI/qai.pdf

Nissenbaum, H. (2001). How Computer Systems Embody Values. *Computer*, *34*(3), 118–120. https://doi.org/10.1109/2.910905

Nistal-Nuño, B. (2021). Artificial intelligence forecasting mortality at an intensive care unit and comparison to a logistic regression system. *Einstein*, *19*, 1–8. https://doi.org/10.31744/einstein_journal/2021ao6283

Nolfi, St. (2021). *Behavioral and Cognitive Robotics: An adaptive perspective*. https://www.researchgate.net/publication/351093674_Behavioral_and_Cognitive_Robotics_An_Adaptive_Perspective

O´Neil, C. (2016). *Weapons of Math Destruction*. New York: Crown.

O'Flaherty, M. (2020). Facial Recognition Technology and Fundamental Rights. *European Data Protection Law Review*, *6*(2), 170–173. https://doi.org/10.21552/edpl/2020/2/4

Parentoni, L. (2022). What Should we Reasonably Expect from Artificial Intelligence? *Il Diritto Degli Affari*, 2, 179. https://www.ildirittodegliaffari.it/articolo/123

Parentoni, L. (2020). Artificial Intelligence. In M. Sellers, S. Kirste et al. (Eds.). *Encyclopedia of the Philosophy of Law and Social Philosophy*. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-6730-0_745-1

Parentoni, L., Valentini, R. S., & Alves, T. C. O. (2020). Panorama da Regulação da Inteligência Artificial no Brasil: com ênfase no PLS n. 5.051/2019. *Revista Eletrônica do Curso de Direito UFSM. Santa Maria: Universidade Federal de Santa Maria – UFSM*, *15*(2), 1–29, ago. 2020. https://doi.org/10.5902/1981369443730

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Poscher, R. (2022). Artificial Intelligence and the Right to Data Protection. In S. Voeneky, Ph. Kellmeyer, O. Mueller, & W. Burgard (Eds.), *The Cambridge Handbook of Responsible Artificial Intelligence. Interdisciplinary Perspectives* (pp. 281–289). Cambridge University Press. https://doi.org/10.1017/9781009207898.022

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'Sandy' … Wellman, M. (2019). Machine behaviour. *Nature*, *568*, 477–486. https://doi.org/10.1038/s41586-019-1138-y

Reeves, Sh. R., Alcala, R. T. P., & McCarthy, Amy. (2021). Challenges in regulating lethal autonomous weapons under international law. *Southwestern Journal of International Law*, *27*(1), 101–118.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence*: *A Modern Approach* (3d ed.). New Jersey: Prentice-Hall. https://people.engr.tamu.edu/guni/csce421/files/AI_Russell_Norvig.pdf

Russell, S. J., & Norvig, P. (2022). *Artificial Intelligence: A Modern Approach*. (4th ed.). London: Pearson.

Santos, M. K., Júnior J. R. F., Wada, D. T., Tenório, A. P. M., Nogueira-Barbosa, M. H., & Marques, P. M. de Azevedo. (2019). Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine. *Radiologia Brasileira*, *52*(6), 387–396. https://doi.org/10.1590/0100-3984.2019.0049

Scherer, M. U. (2016). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies and Strategies. *Harvard Journal of Law & Technology*, *29*(2), 353–400. Spring. http://dx.doi.org/10.2139/ssrn.2609777

Schikowski, A. B., Corte, A. P. D., Ruza, M. S., Sanquetta, C. R., & Montaño, R. A. N. R. (2018). Modeling of stem form and volume through machine learning. *Anais da Academia Brasileira de Ciências*, *90*(4), 3389–3401. (In Portuguese). https://doi.org/10.1590/0001-3765201820170569

Solum, L. B. (2019). Artificially Intelligent Law. *BioLaw Journal*, 1. https://doi.org/10.15168/2284-4503-351

Vijipriya, J., Ashok, J., & Suppiah, S. (2016). A Review on Significance of Sub Fields in Artificial Intelligence. *International Journal of Latest Trends in Engineering and Technology – IJLTET*, *6*(3), 542–548. https://www.ijltet.org/journal/86.pdf

Wachter, S., Mittelstadt, B., & Russell, Ch. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841–887. https://doi.org/10.2139/ssrn.3063289

Wagner, Meira J., & Zaki, M. J. (2020). *Data Mining and Machine Learning*: *Fundamental Concepts and Algorithms*. (2d ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108564175

Wimmer, M., & Doneda, D. (2021). "Falhas de IA" e a Intervenção Humana em Decisões Automatizadas: Parâmetros para a Legitimação pela Humanização. *Revista Direito Público*, *18*(100), 374–406. (In Portuguese). https://doi.org/10.11117/rdp.v18i100.6119

Woodrow, B. (2015). *Cyber-Humans*: *Our Future with Machines*. New York: Springer.

Yin, Ming, Vaughan, J. W., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland.*

История статьи / Article history

Дата поступления / Received 01.12.2023
Дата одобрения после рецензирования / Date of approval after reviewing 14.01.2024
Дата принятия в печать / Accepted 15.01.2024