МГИМО
УНИВЕРСИТЕТ

НАУЧНЫЙ ЖУРНАЛ

ИССЛЕДОВАНИЯ В ЦИФРОВОЙ ЭКОНОМИКЕ

том 1, выпуск 1

volume 1, number 1

JOURNAL OF DIGITAL ECONOMY RESEARCH

2023

научный журнал

«ИССЛЕДОВАНИЯ В ЦИФРОВОЙ ЭКОНОМИКЕ»

https://jder.mgimo.ru/

Содержание • 1 • 2023

Table of Contents • 1 • 2023

Редакционная статья к первому выпуску журнала Исследования в цифровой экономике

Мы рады представить первый выпуск журнала Исследования в цифровой экономике, рецензируемый междисциплинарный научный журнал.

Стремительное развитие цифровой экономики порождает множество дискуссий. Цель данного журнала стать площадкой для исследователей и практиков, политиков и чиновников в обсуждении стратегий развития и примеров лучших международных практик для обеспечения всестороннего развития общества в условиях цифровой экономики.

Журнал фокусируется на

- результатах исследований в области развития цифровой экономики
- вопросах менеджмента в цифровой трансформации
- анализе результатов и рисков цифровизации
- вызовах международного сотрудничества и
- развитии отдельных отраслей и региональных сегментов в цифровой экономике.

В открытом доступе публикуется четыре выпуска журнала ежегодно. Мы надеемся, что данная стратегия поддержит широкие дискуссии о теоретических и эмпирических аспектах цифровой трансформации среди соответствующих групп участников.

Исследовательские и обзорные статьи, охватывающие примеры из практики, рецензии на книги и конференции, интервью ведущих экспертов будут информировать исследователей и специалистов-практиков по современным проблемам развития.

Благодарность

Мы хотели бы выразить нашу искреннюю признательность членам редакционного совета и авторам первого выпуска журнала за их веру в будущее журнала, ценную поддержку и советы, высококачественный и оригинальный научный вклад.

Мы благодарны руководству МГИМО и национальной программе Приоритет 2030 за поддержку в создании и продвижении журнала Исследования в цифровой экономике.

Главный редактор
Анна Абрамова

Editorial to the first issue of the Journal of Digital economy research

We are pleased to introduce the first issue of the Journal of Digital economy research (JDER), a peer-reviewed interdisciplinary scientific journal.

Digital economy rapid developments raise plenty of discussions. The aim of this journal is to be the forum for researchers and practitioners, politicians, and officials to discuss the development strategies and best international case studies to ensure the comprehensive development of society in the digital economy.

The journal focuses on:

- the research results in the field of the development of the digital economy,
- issues in digital transformation management,
- assessment of results and risks of digitalization,
- the challenges of international cooperation, and
- the development of separate industries and regional segments of the digital economy.

JDER is an open access journal with four issues per year. We do hope that this publishing strategy will support the broader discussions of theoretical and empirical aspects of digital transformation pushing forward cooperation between relevant groups of actors. Research and review articles covering examples from practice and cases, reviews of conferences and books, interviews of leading experts could brief researchers and practitioners with contemporary development issues.

Acknowledgements

We would like to express our sincere appreciation to the editorial board members and the authors of the first issue for their trust in the future of the JDER, precious support and advices, high-quality and original scientific contributions. We thank the anonymous reviewers who have greatly contributed the issue for the comments and valuable time.

We are grateful to MGIMO-University authorities and the national grant programme Proirity 2030 for their support of the creation and promotion of the JDER.

Editor-in-Chief
Anna Abramova

Исследования в цифровой экономике. 2023. 1. P. 6–60
DOI: 10.24833/14511791-2023-1-6-60

Получено: 8 октября 2022 г.,
Пересмотрено 10 ноября,
Принято: 3 декабря 2022 г.

СОЦИАЛЬНО-ЭКОНОМИЧЕСКИЕ АСПЕКТЫ ВНЕДРЕНИЯ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

М.В. Федоров

Аннотация. Данная статья посвящена обзору глобальных эффектов внедрения технологий искусственного интеллекта (ИИ), включая социально-этические аспекты, прямое и непрямое влияние на экономику и нормативно-правовую базу для разработки стратегий устойчивого развития на основе этих технологий. Статья в основном сфокусирована на выявлении глобальных рисков, связанных с интенсивным использованием ИИ. Мы обсудим эти проблемы, рассматривая ИИ как часть общемирового процесса технологического развития, и, следовательно, кратко рассмотрим взаимосвязи между ИИ и соседними областями (вычислительные технологии, методы сбора данных и т.д.). Особое внимание будет уделено вопросам международной стандартизации ИИ и связанных с ним технологий. В разделе, посвященном социальному рейтингу на основе ИИ и больших данных, будут обсуждаться фундаментальные проблемы, присущие таким системам, например предвзятость, непрозрачность и т.д. За этим разделом последует раздел о дипфейках, которые будут детально обсуждаться ввиду их значительного влияния на наши представления о доверии, как на индивидуальном уровне, так и на уровне социума/государства. В статье также будет затронута тема влияния повсеместного внедрения искусственного интеллекта в других областях исследований, такие как химия и молекулярная биология. Мы обсудим пути устойчивого развития «доверенного искусственного интеллекта», которые могут обеспечить желаемый баланс между пользой и рисками применения технологий ИИ в глобальном масштабе. Мы рассмотрим подходы, которые могут привести к разработке стратегических принципов, относящихся к долгосрочным прогнозам эффектов ИИ, за которыми последует внедрение соответствующих регуляторных практик.

Ключевые слова: Искусственный интеллект (ИИ), Машинное обучение, Большие данные, Доверенный ИИ, Дипфейки, Обработка естественного языка, Информационно-коммуникационные технологии (ИКТ), Социальный рейтинг, Глобальные аспекты внедрения технологий ИИ (этические, социальные, экономические и т.д.), ИИ в цифровой фармакологии.

Для цитирования: Федоров М.В. (2023). Социально-экономические аспекты внедрения технологий искусственного интеллекта. – *Исследования в цифровой экономике.* №1, С. 6–60. DOI: 10.24833/14511791-2023-1-66-94

Об авторе:

Федоров Максим Валериевич – д-р хим. наук, канд. физ.-мат. наук, член-корреспондент Российской академии наук, ООО «Синтелли» , Россия.
Тел.: 8 862 241 98 44, E-mail: info@siriusuniversity.ru.
ResearcherID: C-7458-2011

Вступление – социальные и экономические эффекты ИИ

Широкое использование искусственного интеллекта (ИИ) по всей планете в следствие ускоренных темпов цифровизации общества, приводит к «размыванию» национальных границ и формированию некоего общемирового цифрового социума (European Commission, 2021; Федоров, Цветков, 2020а; Федоров, Цветков, 2020б). Широкие технологические возможности систем ИИ были реализованы во многих популярных цифровых сервисах (социальные сети, поисковые системы, агрегаторы новостей, торговые площадки и т.д.), значительно расширяя возможности «традиционных» инструментов информационно-коммуникационных технологий (ИКТ) (Kurzweil, 2006; Юхно, 2022а). Сегодня ИИ является важной частью глобального цифрового пространства (Lindre, 2022; Миронова, 2021; Yukhno, 2022b; Юхно, Умаров, 2022). Однако, по сути, это все еще нерегулируемый (хотя и чрезвычайно эффективный) инструмент, который может обеспечить значительные геополитические и экономические достижения владельцам глобальных цифровых платформ и сервисов и изменить направления современных потребительских, общественно-политических и социокультурных тенденций (Федоров, Цветков, 2020б; Федоров, Цветков, 2021а; Линдре, Капитанов, 2022 год. В этой статье мы рассмотрим потенциальные риски, связанные с ИИ, и обсудим пути устойчивого развития этих технологий, ведущие к “Доверенному ИИ”. Анализ тенденций в патентах и научных публикациях сделан в основном на основе аналитики предоставляемой платформами Statista (www.statista.com) и Dimensions (www.dimensions.ai); также использовались другие источники (будут упомянуты позже).

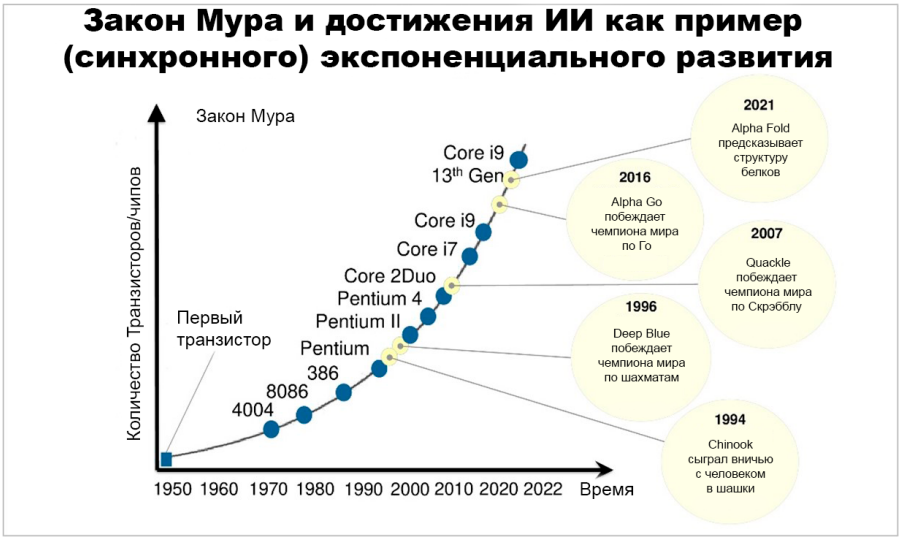

Рис. 1: Сравнение закона Мура и технологий ИИ

Следует отметить, что термин “искусственный интеллект” несколько двусмыслен; в экспертных сообществах существует более 100 определений этого термина, а спектр понимания этого термина общественностью еще шире (Miao Fengchun, et al., 2021). Поэтому, чтобы избежать путаницы, мы будем следовать идеям Курцвейла (Kurzweil, 2006), различая “Слабый ИИ” и “Сильный ИИ”. Под «Слабым ИИ» (также называемым «Узким ИИ») мы подразумеваем технологию ИИ, ограниченную конкретной задачей (например, распознаванием лиц или игрой в шахматы и т.д.). Мы отмечаем, что слабый ИИ может имитировать некоторые конкретные аспекты человеческого поведения и автоматизировать повторяющиеся рутинные задачи. Под «Сильным ИИ» мы будем понимать создание искусственного интеллекта, который превосходит возможности обычного человека в решении ряда различных задач одновременно. Мы отмечаем, что трудно сказать, сможем ли мы создать Сильный ИИ в ближайшем будущем и что будет основой этой технологии, если она будет разработана. Фактически, все технологии ИИ, которые известны в наши дни, можно квалифицировать как слабый/узкий ИИ. Поэтому в статье мы будем использовать термин ИИ, подразумевая, что это слабый ИИ, если не указано иное.

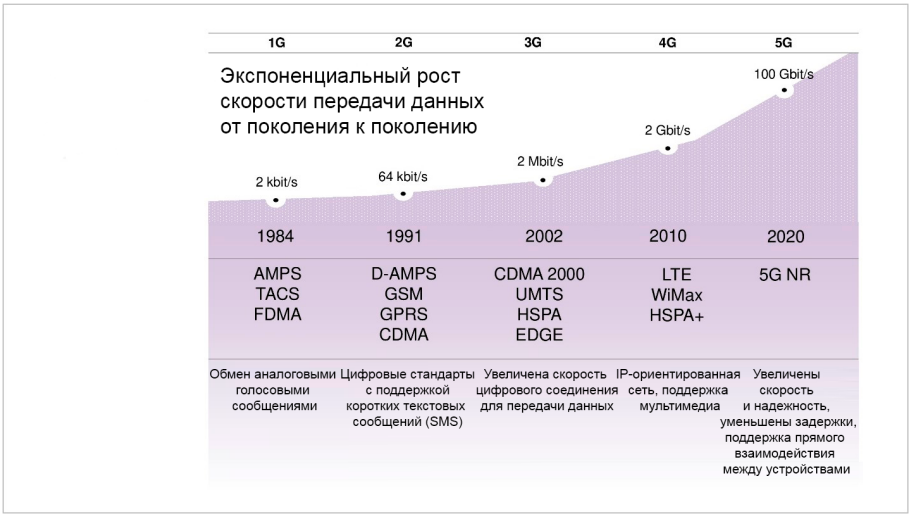

Рис.2 Иллюстрация экспоненциального увеличения скорости передачи данных

С 1950-х годов, когда термин ИИ был введен и стал самостоятельным объектом исследований, по всему миру были запатентованы сотни тысяч изобретений и опубликовано более 1,6 миллиона научных текстов на тему ИИ. В то же время, судя по анализу данных в Statista (www.statista.com) и Dimensions (www.dimensions.ai), более половины всех этих патентов, связанных с ИИ, были опубликованы за последнее десятилетие. Этот экспоненциальный рост приложений искусственного интеллекта сильно коррелирует с экспоненциальным увеличением вычислительной мощности микропроцессоров (так называемый закон

Мура, см. рис. 1) и экспоненциальным увеличением скорости передачи данных (см. рис. 2). Действительно, ИИ тесно взаимосвязан с вычислительными технологиями (Павлов, 2016; Материалы, 2020, 2021;), а также с технологиями обработки и передачи информации (см. рис. 3). Эти три основных стека технологий ИИ действуют вместе и, таким образом, обеспечивают прочную основу для целого ряда приложений (Biamonte, et al., 2017; Sharaev, et al., 2019; Andronov, et al., 2021; Babakov, et al., 2021; Kozlovskii, Popov, 2020; Miao Fengchun, et al., 2021; Morozov, et al., 2021; Khokhlov, et al., 2022), таких как киберспорт, химическая информатика, робототехника, интеллектуальное производство и т.д. (см. рис. 3).

Рис. 3 Области, связанные с технологиями ИИ

Конечно, сам искусственный интеллект тесно связан с методами анализа и обработки данных, а также со вспомогательными технологиями, такими как высокопроизводительные вычисления, датчики и т.д. Следовательно, современный ИИ можно представить как совокупность трех компонентов: (i) Данные, (ii) Программное обеспечение ИИ (в основном алгоритмы машинного мбучения (в основном алгоритмы машинного обучения и соответствующие библиотеки и фреймворки) и (iii) Поддерживающие (вспомогательные) технологии - вычислительная инфраструктура, инфраструктура сбора и передачи данных и т.д. – см. рис. 4. Можно сказать, что самая "интеллектуальная" часть Слабого ИИ - это фреймворки Машинного Обучения; пример типичного конвейера Машинного Обучения схематично представлен на рис. 5.

Анализ поданных патентов и научных публикаций также демонстрирует устойчивую тенденцию смещения акцента с теоретических исследований на практическую разработку продуктов и услуг искусственного интеллекта в коммерческих целях. Машинное Обучение, глубокое обучение и новые архитектуры нейронных сетей являются наиболее динамично развивающимися областями

технологий искусственного интеллекта в последние годы (Miao Fengchun, et al., 2021). Достижения в этой области приводят к появлению автоматизированных систем, которые могут обучаться и выполнять мыслительные задачи, ранее доступные только людям. В области практического применения ИИ наиболее значительный прогресс демонстрируют компьютерное зрение и обработка естественного языка. Однако, как показано на рис. 3, существует ряд перспективных применений и во многих других областях.

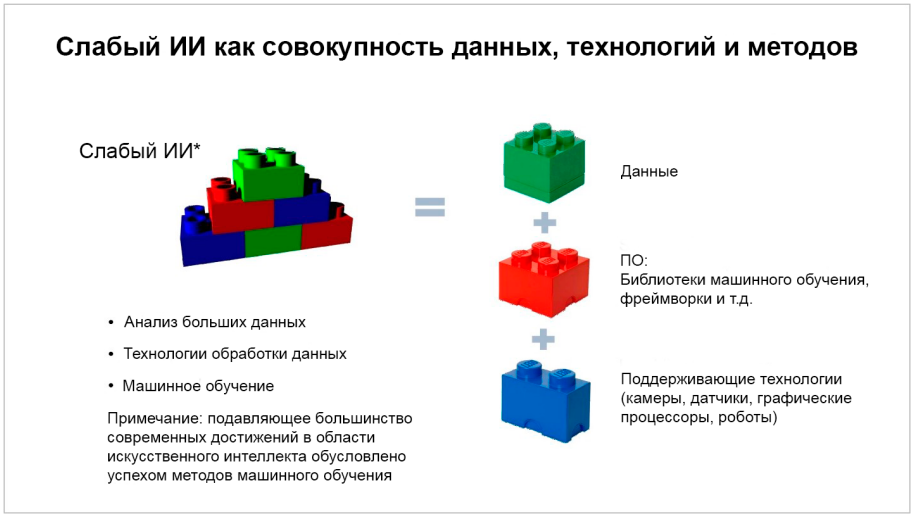

Рис. 4 Основные компоненты Слабого ИИ

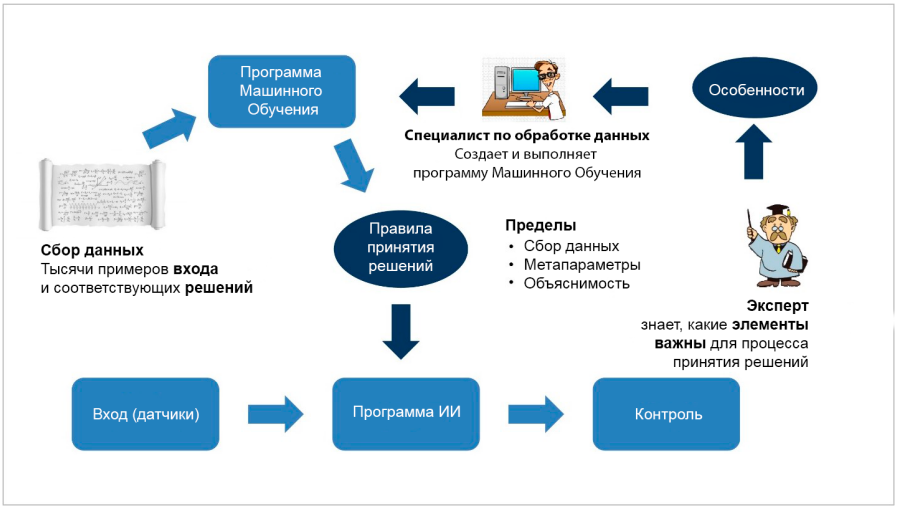

Рис. 5 Схема пути принятия решений в обобщенном алгоритме Машинного Обучения

Очевидно, что такое технологическое развитие будет иметь далеко идущие социальные и культурные последствия (Kurzweil, 2006; UNESCO, 2019; Gurria, 2020; Федоров, Цветков, 2021а; Федоров, Цветков, 2021б; UNESCO, 2021; European Commission, 2021). Необходимо разработать систему критериев и методологическую основу для создания стратегий устойчивого развития для стран и всего человечества в контексте глобального применения и развития технологий искусственного интеллекта (UNESCO, 2019; UNESCO, 2021). Мы отмечаем, что ИИ - это технология общего назначения: он преобразует все секторы экономики и социальную деятельность таким образом, что:

• ИИ повысит производительность в большинстве видов деятельности, где эта технология используется;

• ИИ, равно как и традиционные ИКТ, но в большей степени, влияет на распределение доходов и, следовательно, на социальное равновесие;

• ИИ бросает вызов самоидентификации людей и всего человечества, которая основана на разуме и интеллекте: следовательно, эта технология поднимает сложные социальные и этические проблемы.

Следуя идеям Мироновой (Миронова, 2021) о классификации глобальных рисков, мы можем выделить следующие основные группы глобальных рисков широкого внедрения ИИ.

В группу глобальных *геополитических* рисков входят: кибертерроризм с использованием ИИ, инструменты влияния на внутреннюю и внешнюю политику государств, нарушение цифрового суверенитета государств, постоянно возрастающая угроза использования инструментов ИИ для создания оружия массового уничтожения (в основном химического и биологического) и т.д.

В группу глобальных *экономических* рисков входят: использование инструментов искусственного интеллекта для провоцирования финансового кризиса (посредством манипуляций на фондовой бирже, в криптовалютах и т.д.), высокая ценовая нестабильность на рынке вычислительных компонентов, монополизация рынка вычислительной инфраструктуры, нестабильность на рынке труда и нехватка персонала.

В группу глобальных *технологических* рисков входят: негативные последствия научно-технического прогресса, нарушение работы ведущих информационных систем, растущая экспоненциальная сложность инфраструктуры, накопление ошибок и аварий и т.д.

Группа глобальных *социальных* рисков включает: возможный продовольственный кризис, спровоцированный (неправильно используемым) логистическими цепочками на основе ИИ, возможность «цифровой пандемии», возможные реальные пандемии, спровоцированные новыми инфекционными заболеваниями, «спроектированными" ИИ, повышенная миграционная активность населения из-за удаленной работы, социальное неравенство, порожден-

ное технологиями, огромный разрыв в уровне жизни людей в «цифровых» и «нецифровых» регионах, недоверие к властям, порожденное дипфейками и другими инструментами социальных манипуляций на основе искусственного интеллекта.

Экзистенциальные риски - нарушение баланса между управлением и манипулированием в обществе, возникновение процессов деинтеллектуализации и дезориентации людей в информационном пространстве из-за подмены их интеллектуальных функций искусственным интеллектом, подавления их биологических и психологических потребностей искусственным интеллектом и связанными с ним технологиями (виртуальная реальность, роботы и т.д.).

Рис. 6 Споры по поводу регулирования ИИ

Подводя итог, можно сказать, что одной из главных экзистенциальных угроз является так называемый феномен «опасного знания» (Миронова, 2021). Опасные знания можно определить как информацию, полученную в ходе научных исследований, разработки и анализа больших данных с помощью искусственного интеллекта, негативные последствия которой наше общество не всегда может эффективно контролировать. Это особенно актуально для использования ИИ в таких областях, как генная инженерия, ядерная физика, химические науки и гуманитарные технологии. В конечном счете, это проблема этики науки, ценностей и формирования гуманистических идеалов в области технологий ИИ. Приведенное выше обсуждение приводит нас к выводу, что практически нерегулируемое (пока) развитие искусственного интеллекта, если оно будет продолжаться без надлежащего контроля со стороны регулирующих инструментов, может привести к катастрофическим глобальным последствиям. Однако на многих национальных и международных платформах регулирования ведутся весьма полярные споры по самым основным принципам регулирования, и пока нет четкого

решения, как достичь желаемого баланса между безопасностью и технологической свободой (см. рис. 6). Поэтому в следующем разделе мы обсудим основные международные и национальные мероприятия по разработке нормативно-правовой базы для искусственного интеллекта.

Регулирование технологий ИИ: все еще испытание

Одним из наиболее информативных показателей тенденций в области технологий является содержание соответствующих технологических стандартов и нормативных документов (Materials, 2020, 2021). Анализ документов, созданных для стандартизации и регулирования технологий ИИ и больших данных, позволяет нам увидеть приоритеты разработок в различных странах, а также сделать выводы о деятельности корпораций, разрабатывающих технологии ИИ.

Рис. 7 Основные барьеры между различными заинтересованными сторонами технологий ИИ и способы их устранения с помощью надлежащих механизмов стандартизации

Стремительный рост рынка технологий ИИ потребовал формирование нормативной базы для последующего регулирования использования решений на основе ИИ с основной целью предотвращения вредного воздействия на людей и общество в целом (см. рис. 7). В то же время более или менее осмысленный процесс формирования нормативной базы начался только в 2010-х годах, когда системы ИИ уже были внедрены в ряде областей.

Ключевым направлением в создании интегрированной системы международного регулирования в области искусственного интеллекта и сквозных технологий является стандартизация, результаты которой по-разному воспринимаются обширным экспертным сообществом и различными группами заинтересованных

сторон (разработчики, регулирующие органы, инженеры и пользователи, промышленный и академический секторы, государственные учреждения). Не всегда возможно сохранить баланс интересов, поскольку технологическая повестка дня априори имеет множество явных и скрытых конфликтов интересов, которые возникают при определении приоритетов в большинстве областей стандартизации. Крупные IT-компании являются важными инвесторами в исследования в области стандартизации искусственного интеллекта, включая финансирование пилотных испытаний и проведение кропотливой работы по обобщению мнений международных экспертов. Почти все редакторы и разработчики стандартов искусственного интеллекта и связанных с ними документов связаны с глобальными транснациональными корпорациями, такими как Microsoft и Alphabet.

Документы, требуемые в качестве результата процесса стандартизации, в основном имеют следующий формат: международный стандарт, технический отчет или техническая спецификация. Кроме того, может быть разработан набор документов для формирования серии стандартов, если область стандартизации охватывает аспекты, которые явно разделяются потребителями.

Основные международные инициативы по стандартизации технологий ИИ были реализованы через Подкомитет по искусственному интеллекту (PC 42) и Объединенный технический комитет 1 (OTC 1) «Информационные технологии» Международной организации по стандартизации и Международной электротехнической комиссии (ISO/IEC)[1] созданный в 2017 году. В гораздо меньшей степени они обсуждаются и развиваются в Международном союзе электросвязи (ITU). Экспертную поддержку процессу стандартизации также оказывает Институт инженеров по электротехнике и электронике (IEEE).

В настоящее время основной акцент при разработке стандартов смещен на следующие области применения систем ИИ: Большие данные (БД), интеллектуальные и автономные транспортные средства (IATS). Особое место в стандартизации также занимают вопросы повышения доверия к системам ИИ, управления рисками, этики и проблемы предвзятости искусственного интеллекта.

С 2018 года проводится отдельная работа по терминологической и концептуальной стандартизации систем искусственного интеллекта - основополагающего документа ISO/IEC (ISO/IEC 22989 – AI - Concepts and Terminology) для всех стандартов, связанных с искусственным интеллектом, которые уже приняты или планируются к принятию[2].

Формирование ряда международных стандартов в области баз данных идет семимильными шагами. Несмотря на давно назревшую необходимость создания фундаментальной терминологической базы данных, которая заложила бы основу для последующих документов в этой области, других технических областях и

[1] https://www.iso.org/committee/6794475.html

[2] ISO/IEC DIS 22989 Information Technology — Artificial Intelligence — Artificial intelligence concepts and terminology

смежных областях стандартизации, до сих пор не существует всеобъемлющей терминологической и концептуальной основы. Довольно внушительная группа стран (США, Китай, Ирландия, Корея, Япония и т.д.), представленная их научными экспертами и бизнес-сообществом с объективно разными исходными подходами к пониманию базы данных, крайне медленно приходит к каким-то консенсусным решениям. Стандарты появляются по определенным аспектам базы данных, отстающим от времени разработки, например, из документов в области качества систем искусственного интеллекта.

На национальной арене, в контексте больших данных, в России принят аналог базового международного стандарта – предварительный национальный стандарт (ПНСТ) - «Информационные технологии. Большие данные. Типичная архитектура». Этот документ объясняет наиболее важную терминологию и указывает соответствующие области применения.

Технический отчет, который, по сути, представляет собой описание параметров человеческого «доверия» в области применения технологий искусственного интеллекта, был опубликован ISO/IEC в апреле 2020 года. Возникновение человеческого «доверия» к технологиям искусственного интеллекта связано с обеспечением их прозрачности, объяснимости, управляемости и т.д. на трех уровнях: физическом, кибернетическом и социальном. В рамках международной дискуссии по стандартизации предполагается, что «доверие» должно быть обеспечено на всех этапах взаимодействия и использования систем искусственного интеллекта людьми, что приводит к так называемой концепции «Доверенного ИИ» (см. рис. 8).

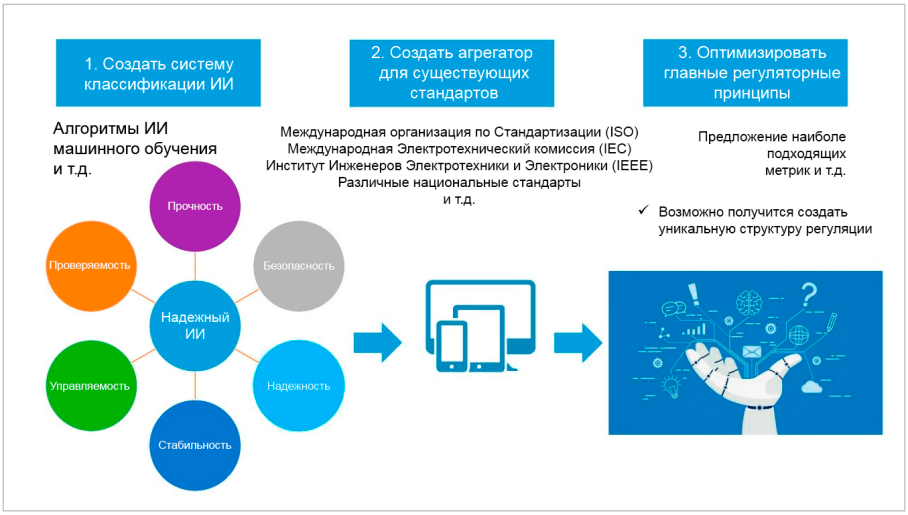

Рис. 8 Компоненты «Доверенного ИИ» и путь создания оптимальной регуляторной структуры

Для сложных систем, основанных на машинном обучении, надежность обычно понимается как абсолютно четкие формализуемые свойства систем, такие как стабильность, надежность (она же устойчивость), гарантированная сходимость, безопасность, устойчивость к атакам – то, что известно как устойчивость, а также набор свойств, таких как конфиденциальность данных, то есть, дифференциальная конфиденциальность (Materials, 2020, 2021).

Отметим, что надежность, также известная как устойчивость, является своего рода метасобственностью, которая описывает сохранение определенного базового свойства, такого как, например, тот же уровень стабильности при наличии некоторых неопределенностей и ошибок в системе. Также этот термин обозначает устойчивость к помехам. Самый банальный пример здесь - когда минимальное искажение визуальной информации приводит к серьезной ошибке классификации: автомобиль может принять пешехода за столб или, скажем, за муху на лобовом стекле. И это не будет классифицировано как некая опасность, которой необходимо избегать только потому, что изображение немного зашумлено (Materials, 2020, 2021).

Дифференциальная конфиденциальность – это свойство, которое заключается в том, что система может обеспечить определенный уровень защиты персональных данных пользователя. Отказоустойчивость означает способность поддерживать нормальное функционирование, в том числе стабильное, выполняющее ограничения и так далее, при наличии неисправности в датчиках или в исполнительных механизмах, то есть в устройствах управления, и так далее (Materials, 2020, 2021).

Устойчивость и дифференциальная конфиденциальность - относительно новые концепции. Это определенные свойства, которые можно рассматривать как подтипы надежности и стабильности, но у них есть некоторая специфика. Возможно, в будущем таких свойств-индикаторов будет еще больше, учитывая специфику искусственного интеллекта (Materials, 2020, 2021).

В ближайшем будущем флагманскими документами в области стандартизации искусственного интеллекта станут стандарт технического регулирования искусственных нейронных сетей и технический отчет об их характеристиках стабильности, принятый в 2021 году. Основной задачей этих документов является регулирование контроля на этапе разработки программного продукта в различных отраслях промышленности и правил сертификации элементов системы для оценки их стабильности, с соответствующим балансом интересов со стороны инженеров и производителей. Например, Американский национальный институт стандартов (ANSI) выступил с предложением заполнить техническую спецификацию ISO/IEC PC 42 в области целей и методов машинного обучения и систем искусственного интеллекта. В нем перечислены методы, которые позволяют повысить объяснимость систем искусственного интеллекта для различных заинтересованных сторон (разработчиков, пользователей, поставщиков, регу-

лирующих органов). Методы различаются для каждой группы заинтересованных сторон. Для разработчиков это означает усиление аспектов безопасности и технической надежности; для пользователей – степень доверия к системе или методу, чтобы субъективность не играла существенной роли; для поставщиков - соблюдение нормативных и технических регламентов, а для регулирующих органов – понимание возможностей и ограничений при создании инновационных фреймворков.

Функциональная безопасность стала актуальной в ходе развития методов машинного обучения и связанных с ними технологий ИИ, которые создали направление интереса в инженерном деле, построении безопасного мира на основе инновационных решений. На 6-м пленарном заседании ISO/IEC SC 42 в апреле 2020 года был предложен проект по функциональной безопасности систем искусственного интеллекта, который после короткого обсуждения было решено реализовать в формате технического отчета ISO/IEC SC 42. Функциональная безопасность играет важную роль во внедрении математических моделей, с помощью которых реализуется применение той или иной технологии искусственного интеллекта. Их свойства отражают соответствие технологии и системы искусственного интеллекта особым правилам безопасности. Если существуют объяснимые и понятные зависимости между ключевыми параметрами, определяющими поведение математической функции, корреляция между наблюдаемыми входными и выходными данными очевидна, то модель будет прозрачной и эффективной. Однако физическое явление, описываемое моделью, может быть очень сложным, иметь очень малый масштаб или не может наблюдаться без влияния экспериментальных данных, которые не могут быть описаны научно обоснованной моделью. Все это вызывает трудности в понимании и верификации модели, вносит двусмысленности и определенную степень неопределенности.

Технический отчет о функциональной безопасности ИИ служит описанием того, как контролировать безопасность, риски безопасности и их связь с системами и технологиями ИИ, а также классифицирует технологии ИИ и уровень их возможного использования. Классы (1-3) отражают возможность большей вариативности при применении в различных областях, от наличия четких рекомендаций по критериям безопасности до невозможности создания требований безопасности вообще. Уровни (A-D) показывают степень влияния компонента системы ИИ на безопасность в целом. Кроме того, в документе использовалась концепция уровня автоматизации и контроля, степени прозрачности и объяснимости решений (от полностью прозрачной системы до «черного ящика»), а также другие аспекты, связанные, в том числе, с физическими компонентами систем.

Таким образом, функциональная безопасность может быть специфичной и зависеть от применения системы искусственного интеллекта. Это открывает возможность реализации его аспектов с использованием различных подходов на том или ином этапе жизненного цикла систем искусственного интеллекта.

Вышеуказанные документы были созданы в ISO / IEC для того, чтобы сформировать обширную и исчерпывающую базу для стандарта или спецификации для тестирования систем искусственного интеллекта.

Искусственный интеллект и социальный рейтинг: основные риски.

Возможно, одной из наиболее спорных областей применения искусственного интеллекта является использование искусственного интеллекта для разработки систем социального рейтинга (Федоров, Цветков, 2020а; Федоров, Цветков, 2021б). Вкратце, систему социального рейтинга можно определить как набор социальных показателей, оцениваемых автоматизированной системой (Бородкин, 2004; Meadows, 1998). На основе выходных данных (рейтингов или индивидуального рейтинга) государственные учреждения и общественные институции формируют модель взаимодействия с конкретным гражданином. Мы отмечаем, что многие такие системы были разработаны в областях, связанных с образованием (Miao Fengchun, et al., 2021). Естественно, что в качестве эффективного инструмента анализа данных ИИ стал популярным инструментом для разработки различных региональных и национальных систем социального рейтинга. Однако недавние исследования (Федоров, Цветков, 2020а; Федоров, Цветков, 2021б; Lindre, 2022), в которых анализируются эффекты систем социального рейтинга, внедренных в разных частях планеты, выявили несколько фундаментальных проблем таких систем, которые мы кратко рассмотрим ниже.

Ниже перечислены основные проблемы концепции социального рейтинга.

1. Непрозрачность правил и алгоритмов, на основе которых система социального рейтинга оценивает поведение человека. Современные системы социального рейтинга непрозрачны, поскольку они, как правило, затрудняют доступ к информации как о процессе разработки правил и критериев оценки действий человека интеллектуальной системой, так и об именах разработчиков. В связи с этим возникает вопрос о степени ответственности разработчиков и владельцев таких систем за принятие решений, которые могут иметь огромное значение для миллионов граждан. В то же время остается неясно, как система социального рейтинга, которая использует ИИ для обработки огромных объемов данных граждан во всей стране или отдельно взятом обществе, на самом деле принимает решения, поскольку большинство алгоритмов ИИ работают по принципу классического «черного ящика». Это означает, что выходные результаты не подразумевают объяснения причин и обстоятельств, по которым система искусственного интеллекта делает выводы и принимает решения. Люди могут только смириться с тем, что «интеллектуальная» машина в режиме реального времени «сортирует» их по группам в зависимости от их поведения и совершенных действий, даже если такие действия и решения не противоречат закону.

2. Никто не застрахован от понижения социального рейтинга. Некоторые лоббисты системы социального рейтинга (наивно) полагают, что следование определенным образцам «хорошего гражданина» является гарантией нахождения человека в высшей привилегированной группе. Однако, как и любая другая сложная технология, социальный рейтинг на основе искусственного интеллекта способен не только отслеживать общий фон вокруг человека на основе определенной системы показателей, но и идти дальше - формировать «мнение» о таком человеке с учетом мультипликативного эффекта восприятия человеческого поведения всем обществом. Например, популярная медийная личность (политик, артист или спортсмен), с одной стороны, имеет больше возможностей поддерживать свой социальный рейтинг на высоком уровне за счет накопительного эффекта своего положительного имиджа. С другой стороны, при спланированной информационной кампании по дискредитации такого человека «интеллектуальная» система социального рейтинга будет учитывать меняющееся в худшую сторону восприятие этого человека в обществе, и такому человеку будет сложнее «скорректировать» свой рейтинг с помощью «хороших поступков» позже. Система дает среднюю оценку человеку с учетом «хорошего» и «плохого» поведения в медиапространстве

3. Система социального рейтинга может иметь большое количество затрудненных результатов. Распределение людей по категориям в зависимости от текущего социального рейтинга может иметь далеко идущие последствия во всех сферах жизни. Например, человека с недостаточно высоким рейтингом вряд ли возьмут на престижную работу, и в конце концов этот человек даже не поймет, почему его заявление о приеме на работу было отклонено. Фундаментальное право каждого человека – право на свободу выбора своей жизненной траектории - будет соразмерно ограничено определенной частью «фазового пространства», занимаемой в социальном рейтинге человека. Неоплаченный штраф за нарушение правил дорожного движения или невыполненная «детская программа» могут самым неожиданным образом проявиться в чьем-то будущем и сыграть решающую роль в судьбе человека при реализации его планов на жизнь.

4. Цифровая дискриминация людей. Принятие системы социального рейтинга на государственном уровне в стране фактически означало бы появление в этой стране «цифрового кодекса законов» и «цифрового процессуального кодекса», которые будут параллельны Конституции страны и другим нормативно-правовым актам (набор социальных показателей и алгоритмов для сбора, анализа и оценки информации о человеке, включенном в систему искусственного интеллекта). Следовательно, вместо того, чтобы предоставлять людям равные права и обязанности, такая система ввела бы полноценный механизм дискриминации и ограничения прав человека, которые могут возникнуть без совершения незаконных действий, которые являются таковыми в соответствии с действующим законодательством

Существует прямая угроза того, что социальный рейтинг превратится в диктатуру определенной системы ценностей, внедренной неизвестными программистами в автоматизированную систему, работающую по каким-то критериям. Это, в частности, является прямым нарушением седьмой статьи Конституции Российской Федерации, которая гарантирует свободное развитие личности.

5. Система социального рейтинга может привести к социальной дестабилизации. Введение даже отдельных элементов социального рейтинга может быть очень криминогенным – потому что такая система автоматически предоставит незаконные способы «подправить» рейтинг за умеренную плату и попасть на более высокий уровень. Кроме того, социальный рейтинг - это явная провокация общественного напряжения. Это, очевидно, вызовет протестные настроения в обществе, что прямо противоречит современному занижению функции демократического государства, где одной из важнейших задач является создание условий для взаимного доверия между ним и обществом. В этом контексте вопрос о «переходе» человека из одной категории социального рейтинга в другую остается открытым. Такие системы создают возможность прямой угрозы юридического «порабощения» человека, т.е. пожизненного пребывания в нижней части рейтинговой пирамиды из-за объективного сужения возможностей последнего улучшить свои показатели. Изначально ориентированная на дискриминацию концепция социального рейтинга будет ущемлять права лиц, которые, хотя и не нарушают закон, не «зарабатывают» достаточно баллов «хорошего гражданина».

Примечательно, что риски систем социального рейтинга были выявлены и осознаны руководящими структурами Европейского Союза на ранней стадии. В апреле 2021 года Европейская Комиссия предложила запретить внедрение технологий искусственного интеллекта (резолюция «О Европейском Подходе к Искусственному Интеллекту»), которые используются для «массовой слежки, применяемой в обобщенной форме ко всем лицам без каких-либо различий». Такие методы наблюдения, как «мониторинг и слежение за отдельными лицами в цифровой или физической среде, а также автоматическое агрегирование и анализ персональных данных из различных источников», станут незаконными[3].

Если эта инициатива будет одобрена Европейским Парламентом и Европейским Советом, Европейский Союз полностью запретит использование систем искусственного интеллекта «высокого риска» и ограничит использование других, если они не соответствуют новым стандартам. Примечательно, что в категорию «высокого риска» входят, среди прочего, технологии искусственного интеллекта в роботизированной хирургии, ИИ в программном обеспечении для найма сотрудников, ИИ для проверки документов и подтверждения доказательств (в судебных разбирательствах), а также ИИ для оценки кредитного рейтинга граждан – первичный элемент всей системы социального рейтинга. Более

[3] https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682

того, в пояснении к документу Европейской Комиссии отмечается, что сама суть социального рейтинга и использование искусственного интеллекта в приложениях, которые манипулируют поведением людей в обход их воли, неприемлемы (European Commission, 2021). Представители бизнеса, которые игнорируют эти новые правила, могут быть оштрафованы на сумму до 20 миллионов евро или 4% от годового оборота.[4]

Мы считаем, что пришло время инициировать полноценное научное исследование феномена социального рейтинга в глобальном масштабе. Полученные результаты должны быть вынесены на широкое общественное обсуждение с участием представителей всех слоев общества (научного и экспертного сообщества, правозащитных организаций, законодателей, деловых кругов, политических партий, гражданского общества и т.д.).

Дипфейки – где правда?

Дипфейки (от англ. *Deepfake* – глубокая подделка) можно определить как синтетические слуховые или визуальные медиа, разработанные с использованием методов глубокого машинного обучения, которые создают настолько реалистичное впечатление, что могут обмануть аудиторию (Hutchinson, 2020; Jaiman, 2020; Линдре, Капитанов, 2022; Prajakta, 2020). Это главная задача авторов дипфейков – создать максимально реалистичную иллюзию. Искусственно созданные медиа широко варьируются по технической сложности и применению, начиная от низкокачественных «дешевых подделок» и заканчивая высококлассными продуктами, которые могут влиять на «восприятие реальности» человеком, а также в определенном смысле изменять процесс принятия решений человеком.

Дипфейки могут быть успешно и с пользой использованы в таких областях, как образование и искусство, но, оказывается, что они скорее нанесут вред отдельным лицам, бизнесу, обществу и демократии в целом, а также ускорят процесс падения общественного доверия к СМИ (Jaiman, 2020; Logacheva, 2021; Babakov, et al., 2021; Dementieva, Panchenko, 2021). Такой подрыв доверия будет способствовать расцвету фактического релятивизма, разрушающего структуру демократии и гражданского общества из-за наплыва ложной информации.

Масштаб использования дипфейков, с которым сталкиваются многие заинтересованные стороны (включая государственные учреждения, компании и частные лица), создает ряд проблем для современного общества. Наиболее беспокоящим является то, насколько дипфейки способствуют развитию эры интернет-дезинформации (Perez, 2019). Действительно, поскольку инструменты для создания искусственной информации для массовой аудитории постоянно совершенствуются на техническом уровне, регулирующие органы по всему миру

[4] https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682

пытаются одновременно решать возникающие социально значимые проблемы, связанные с адекватным реагированием на новые риски и угрозы, создаваемые дипфейками (главным образом, высокий уровень недоверия к публичным СМИ).

Например, существует важный вопрос о защите прав физических лиц в сфере контроля за коммерческим или неправомерным использованием их изображений и других аспектов "цифровой" идентичности (Jaiman, 2020). Методология создания дипфейков уже развилась до такой степени, что многие люди могут полностью воссоздать свой или чей-то другой образ в цифровой форме, чтобы использовать эти «обновленные» изображения в различного рода виртуальных проектах, в которых они не принимают прямого «личного» участия (Davis, 2020). В этом случае какова будет плата за коммерческое использование их изображений и изображения в целом? Кому будут принадлежать права на использование нового изображения (которое может представлять значительную ценность)?

Что еще более важно, в рамках судебного разбирательства дипфейки могут представлять серьезную угрозу для определения подлинности наиболее важных доказательств, поскольку непрерывное развитие технологий значительно усложняет процесс определения реальных цифровых доказательств (изображений, видео, аудио и т.д.) от подделки и, следовательно, ставит под угрозу использование цифровых доказательства (Kietzmann, et al., 2019; Prajakta, 2020). Например, некоторые суды в Соединенных Штатах разрешили фотографировать доказательства без проверки достоверности показаний очевидцев в соответствии с так называемым «правилом молчаливого свидетеля» (использование «замен» при доступе к конфиденциальной информации в Соединенных Штатах в системе открытого суда присяжных). При сохранении текущей динамики распространения дипфейков для юристов может стать обычной практикой заявлять, что цифровые доказательства против их клиента являются подделкой. Это обстоятельство может заставить присяжных усомниться в подлинности реальных доказательств (Prajakta, 2020).

Еще одной проблемной областью использования дипфейков является область кибербезопасности (Gurria, 2020). Действительно, фальсифицированные данные и цифровой след человека создают беспрецедентные риски. Например, данные датчиков могут быть фальсифицированы и затем переданы в системы принятия решений ИИ, что приведет к обману системы искусственного интеллекта с последующим принятием ею неверных решений (Davis, 2020).

Комплексная система защиты общества от вредного использования дипфейков начинается с программного обеспечения, предназначенного для обнаружения таковых или повышения технической способности отличать реальный контент от поддельного. Такие технические решения включают в себя как системы искусственного интеллекта, обученные распознавать аномалии, характерные для дипфейков, так и криптографические методы, которые могут быть интегрированы в оборудование для видео- и аудиозаписи с целью обнаружения

несанкционированного доступа. Однако средства обнаружения дипфейков и противодействия им всегда будут отставать от темпов развития самих технологий подделки. Кроме того, пока не видно решения следующей проблемы: простое знание того, что определенный видеофайл был обработан, не дает информации о том, как идентифицировать создателей и привлечь их к ответственности (Davis, 2020; Jaiman, 2020).

Вторым уровнем противодействия деструктивным дипфейкам является законодательное регулирование. Например, ряд штатов США, например, Вирджиния, приняли местные законы, криминализирующие дипфейки, используемые в порнографии; а штат Техас дополнительно криминализировал дипфейки, созданные для влияния на избирательный процесс. С другой стороны, штаты Массачусетс и Калифорния не смогли одобрить законопроекты, направленные на противодействие дипфейкам из-за опасений по поводу чрезмерного «ужесточения» ИТ-индустрии (Davis, 2020; Jaiman, 2020).

Возможно, наилучшим результатом стала бы разработка технических и юридических инструментов для управления рисками причинения вреда от дипфейков, которые не подавляют инновации и деловую активность, а также не посягают на свободу выражения мнений. Ответственность должна быть возложена на отдельных лиц без введения всеобщих запретов на новые технологии.

Актуальной и практически неразрешимой проблемой в большинстве стран мира сегодня является то, что государство, в первую очередь законодательные органы, не предпринимают достаточных действий для обеспечения подлинности аудиовизуального контента, распространяемого в публичном пространстве. Поощрение цифровых платформ к самоцензуре и самоидентификации деструктивного контента с использованием технологий «Deepdale» имеет очень ограниченный эффект. Например, если происхождение видео невозможно отследить, национальным правительствам рекомендуется создать орган, который может обнаруживать дипфейки с использованием технологии блокчейн. Суть предложения экспертов заключается в том, что цепочки блоков хранят данные в децентрализованной сети, в которой любой желающий может проверить оригинальность информации, сравнив ее с определенным уникальным и неизменяемым ключом. Даже малейшая манипуляция данными приведет к легко обнаруживаемому несоответствию.

Технологический подход к управлению и ограничению вредоносных дипфейков заключается в разработке методов и приемов для проверки подлинности аутентичного контента. Примером может служить технология под названием Amber Authenticate, которая работает на устройствах, воспроизводящих оригинальные, то есть аутентичные фотографии, аудио- и видеоконтент в режиме реального времени по мере записи контента. Программа создает так называемый “слой правды”, который представляет собой исходный контент, криптографически помеченный многочисленными цифровыми отпечатками пальцев, а затем заархивированный в общедоступной цепочке блоков. Эта дактилоскопия циф-

рового контента используется для отслеживания его происхождения по мере распространения в цифровом пространстве, а также для выявления попыток манипулирования исходным контентом и реагирования на них.

Администрация киберпространства Китая разработала правила (вступили в силу 1 января 2020 года), запрещающие любую публикацию в цифровом информационном пространстве с измененным контентом, созданным с использованием технологий искусственного интеллекта, без соответствующего предупреждения. Эти меры были приняты по соображениям национальной безопасности, чтобы предотвратить тиражирование фальшивых новостей и распространение политической дезинформации на ранней стадии (Jaiman, 2020).

Новые правила позволяют правоохранительным органам преследовать в судебном порядке лиц, которые создают дипфейки и видеоплатформы, на которых они размещены. В документе конкретно не упоминается термин deepfake — вместо этого запрещается публиковать в целом вводящий в заблуждение контент, который может нанести ущерб репутации, например, политического деятеля или обмануть избирателей (Babakov, et al., 2021). В то же время правила регулирования дипфейков носят длительный характер, за нарушение которых формально не предусмотрено никаких конкретных административных или уголовных мер (Jaiman, 2020).

В конце раздела мы хотели бы привести предложения из исследования Университета Индианы «Deepfakes: Trick or treat?» о том, как регулировать дипфейки и бороться с вредоносным контентом на основе этой технологии (Kietzmann et al., 2019).

1. Исправление исходного контента, гарантирующего обнаружение недостоверности дипфейка. Эффективная борьба с дипфейками, созданными с целью обманного или манипулятивного воздействия на человека, предполагает систему раннего обнаружения и блокировки такого контента. В связи с этим становятся востребованными технологии, которые отслеживают и фиксируют жизнь человека, включая его геолокацию, взаимодействие с другими людьми и учреждениями, а также информацию о любой другой «внешней» деятельности. Несмотря на прямую угрозу неприкосновенности частной жизни и конфиденциальности действий человека, с технической точки зрения сбор такой информации с современных мобильных устройств кажется довольно простым и удобным. Однако такие личные и конфиденциальные данные должны быть зашифрованы, сохранены и использованы для сравнительного анализа только в случае необходимости для “разоблачения” глубоких подделок.

2. Разоблачение вредных дипфейков. Наряду с развитием технологий искусственного интеллекта, которые упрощают процесс создания и улучшения конечного результата аудиовизуальных подделок, также разрабатываются технологические инновации для обнаружения и классификации дипфейков. Лидерами в этой области являются Соединенные Штаты. Например, Агентство перспективных исследовательских проектов Министерства обороны США (DARPA)

располагает инновационными средствами обнаружения глубоких подделок, в частности, программой судебной экспертизы цифрового медиаконтента, а также сервисами обнаружения дипфейков, такими как Truepic. Глобальные корпорации и цифровые платформы также вкладывают серьезные ресурсы в идентификацию и обнаружение дипфейков.

3. Защита прав в рамках закона. Жертвы дипфейков должны иметь предусмотренную законом систему защиты от материального или морального вреда, причиненного им в случае клеветы, злого умысла, нарушения неприкосновенности частной жизни или морального ущерба, вызванного дипфейками, а также в случаях нарушения авторских прав, имитации и мошенничества, связанных с дипфейками.

4. Принятие мер по укреплению доверия. Глобальные компании, которые утверждают, что придерживаются универсальных моральных принципов и прав человека, в контексте дипфейков должны удвоить свои усилия по укреплению доверия со своими клиентами и установлению прочных эмоциональных связей с ними. В том случае, если компании или их бренды оказываются вовлеченными в скандалы на основе дипломатических угроз, рядовые клиенты и партнеры, скорее всего, критически воспримут информационный шум и сохранят свою лояльность.

ИИ в Цифровой Фармакологии и Молекулярной Биологии

Фармацевтические науки и молекулярная биология в настоящее время претерпевают “фазовый переход” из-за широкого использования методов машинного обучения на основе больших данных в этих областях (Sosnin, et al., 2018b; Kozlovskii, Popov, 2020; Sosnina, et al., 2020; Zaretckii, et al., 2022). Издалека это может показаться странным, потому что эти области традиционно ассоциировались с интенсивным использованием экспериментальных методов и клинических испытаний. Однако современные исследовательские программы в области биомолекулярных наук имеют в своей основе экспериментальные данные (например данные клинических испытаний), а также сгенерированные данные с помощью вычислений (например, данные по молекулярному моделированию); таким образом, анализ данных и исследования в области машинного обучения на основе сгенерированных данных могут в значительной степени способствовать прогрессу в этих областях. Цифровая фармацевтика является одним из замечательных примеров, когда высокопроизводительные платформы для скрининга in vitro или in silico используются для получения прогностических моделей, которые могут быть применены для решения фундаментальных и практических проблем в области открытия современных лекарств и молекулярного дизайна (Sosnin, et al., 2018b; Sosnina, et al., 2020; Andronov, et al., 2021; Khokhlov, et al., 2022).

Рис. 9 Проблемы цифровой фармакологии

Каждый год фармакологические компании по всему миру тратят сотни миллиардов долларов США на разработку и дизайн лекарств. Чтобы справиться с кризисом, связанным с экспоненциальным ростом затрат на исследования и разработки в этих областях, необходимо разработать подходы, и для этого большие перспективы имеют молекулярные технологии с интенсивным использованием данных. Действительно, недавно в престижных журналах был опубликован ряд работ по применению искусственного интеллекта в цифровой фармакологии и молекулярной биологии, которые являются отличными примерами пересечения современных вычислительных методов с интенсивным использованием данных и «влажной» биологии (Young, et al, 2018; Popov, et al., 2019c; Popov, et al., 2019b; Popov, et al., 2019c; Karlov, et al., 2020; Kozlovskii, Popov, 2020; Kozlovskii, Popov, 2021a; Kozlovskii, Popov, 2021b; Morozov, et al., 2021; Zaretckii, et al., 2022).

Прогресс в экспериментальных методах уже привел к накоплению большого объема соответствующих данных, и скорость генерации новых данных все еще растет (Sosnin, et al., 2018b). Эффективная интеграция крупномасштабного анализа данных с «влажными» биомолекулярными науками требует разработки новых подходов и новых инструментов, а также квалифицированных исследователей и инженеров как с IT, так и с химическим / биологическим образованием для интеграции инструментов на основе искусственного интеллекта с экспериментальными исследовательскими программами. Более того, химическое пространство обладает сложной структурой и огромной размерностью; таким образом, простые алгоритмы машинного обучения обычно имеют очень ограниченную область применимости, что делает их бесполезными на практике. Использование современных методов статистического анализа с интенсивным использованием данных наряду с подходами молекулярного моделирования и машинного обучения позволяет расширить область применимости «чистых» методов машин-

ного обучения и разработать более надежные решения (Соснин и др., 2018а). Эта стратегическая область направлена на разработку комплексных технологий, использующих основанные на знаниях, последовательности, графах, структуре и многомерных представлениях интенсивных молекулярных данных (Andronov, et al., 2021)..

Рис. 10 Задачи для ИИ в цифровой фармакологии

Рис. 11 Проблемы структурной биологии

Два ключевых направления развития в этой области:

- разрабатывать и применять инструменты на основе искусственного интеллекта для исследования целевых областей в химическом пространстве и создания новых химических веществ с желаемыми свойствами
- разрабатывать и применять инструменты на основе искусственного интеллекта для исследования целевых областей в биологическом пространстве с акцентом на молекулярную биологию и молекулярные механизмы взаимодействия лекарства с макромолекулой и макромолекулы с макромолекулой.

Рис. 12. Задачи ИИ в структурной биологии

Рисунки 9-10 иллюстрируют первый подход. Рисунки 11-12 иллюстрируют второй подход.

Вероятно, можно оценить, что применение ИИ в бимолекулярных науках может (i) снизить затраты на НИОКР для разработки и проектирования на 30-50 процентов; (ii) сократить время разработки новых лекарств и вакцин на 20 (пессимистично) - 70 (оптимистично) процентов; (iii) значительно сократить потенциальные риски новых лекарств (побочные эффекты, проблемы со стабильностью и т.д.) из-за тщательной проверки препаратов-кандидатов in silico. Следовательно, при широком внедрении эти инструменты потенциально могут сэкономить эквиваленты сотен миллиардов долларов США в год и, что более важно, спасти миллионы человеческих жизней и внести значительный вклад в улучшение здоровья людей на всей планете. Однако, несмотря на положительные эффекты уже многих примененных в цифровых молекулярных науках методов, основанных на искусственном интеллекте, возникают опасения по поводу потенциальных рисков, связанных с внедрением этих методов. Например, недавнее исследование показало, что методы, основанные на искусственном интеллекте, могут быть использованы для быстрого получения высокотоксичных соедине-

ний (Urbina, et al., 2022). Таким образом, двойственная природа ИИ проявляется и в этих областях, что поднимает вопросы о надлежащем регулировании применения методов ИИ в междисциплинарных областях.

Обсуждение

Из-за многочисленных социальных и экономических последствий этих технологий ряд тем, связанных с искусственным интеллектом, был включен в политическую и гуманитарную повестку дня ООН в последние годы (UNESCO; UNESCO, 2021). Значительную роль в этом сыграли Цели Устойчивого Развития (SDGs), принятые на саммите Всемирной организации в 2015 году, которые представляют собой список из 17 глобальных целей и 169 задач (United Nations, 2015). Успешная реализация большинства из них связана с необходимостью консолидации усилий человечества для достижения нового уровня развития общества. Предполагается, что переход от третьей к четвертой промышленной революции (характеризующейся развитием и массовым внедрением киберфизических систем в глобальное производство) позволит аккумулировать и использовать дополнительные ресурсы с целью стимулирования роста мировой экономики, повсеместной ликвидации бедности, снижения негативного воздействия человека на окружающую среду и т.д. Предпосылками такой трансформации является внедрение принципиально новой технологической базы во все ключевые сферы жизни общества, способной ускорить эволюцию основных социально-экономических процессов, а также создать новые и повысить эффективность существующих производственных цепочек. Ведущая роль в этом отводится «умным» высокотехнологичным системам, которые используют в своей работе ИИ или его отдельные компоненты. Однако, как указано во введении, массовое внедрение искусственного интеллекта создает ряд экзистенциальных рисков, которые могут повлиять на судьбу всей цивилизации. В то же время, в области оценки рисков, связанных с ИИ, нам необходимо перейти от дискуссий и постфактумных наблюдений к фундаментальным исследованиям, направленным на разработку набора стратегических принципов для упреждающей оценки рисков и моделирования/оценки соответствующих мер по снижению этих рисков. Задачи действительно сложные из-за экспоненциальной скорости развития искусственного интеллекта и технологий больших данных (см. рис. 1 и рис. 2 для иллюстрации).

Тем не менее, можно извлечь уроки из истории фармацевтической науки, которая в течение 20-21 веков эволюционировала от нерегулируемого дикого поля до стадии, когда все этапы процесса разработки лекарств строго регулируются как на основе науки и техники, так и этики. Настало время вспомнить старейший принцип биомедицинской этики, primum non nocere («не навреди»), который может быть применен в качестве основного подхода к регулированию развития технологий искусственного интеллекта. В целом, несмотря на множество отличительных особенностей искусственного интеллекта, общая проблема регу-

лирования потенциально опасных технологий (например, ядерных технологий, клонирования, разработки биологически активных соединений и т.д.) не является чем-то уникальным для человечества. Можно перенять множество полезных практик и методов из других областей, которые могут оптимизировать ресурсы, затрачиваемые на разработку нормативно-правовой базы для искусственного интеллекта. Такой подход может привести к разработке адаптивной нормативно-правовой базы, которая будет соответствовать принципам прозрачности, сотрудничества, подотчетности и этики, чтобы способствовать безопасной и устойчивой разработке инновационных продуктов на основе искусственного интеллекта.

Благодарность:

Автор выражает признательность Ксении Поляковой (Университет Сириус) за большую помощь в подготовке статьи. Мы высоко ценим многочисленные дискуссии с Юрием Линдре (Сколтех) и Анной Ждановой (Сколтех) по различным аспектам глобальных эффектов ИИ; многие идеи, представленные в данной работе, были вдохновлены этими дискуссиями. Большое спасибо Наталье Струшкевич (Сколтех), Петру Попову (Сколтех) и Сергею Соснину (Синтелли) за совместные работы и дискуссии по ИИ в биомолекулярных науках и цифровой фармакологии. Автор благодарен многим своим коллегам из Сколтеха и членам Комитета по этике искусственного интеллекта Комиссии Российской Федерации по делам ЮНЕСКО за обсуждения различных тем, связанных с искусственным интеллектом. Особая благодарность Александру Кулешову, Ректору Сколтеха (председатель Комитета). Автор очень ценит дискуссии с Игорем Ашмановым (Kribrum), Натальей Касперской (Infowatch) и Станиславом Ашмановым (Nanosemantica) о дипфейках и рисках, связанных с цифровыми технологиями.

Ссылки

1. (Andronov, et al., 2021) Andronov M., et al., «Exploring Chemical Reaction Space with Reaction Difference Fingerprints and Parametric t-SNE», ACS Omega, 03.11.2021 (https://pubs.acs.org/doi/10.1021/acsomega.1c04778)

2. (Babakov, et al., 2021) Babakov N., et al., «Detecting Inappropriate Messages on Sensitive Topics that Could Harm a Company's Reputation», Trustworthy AI Conference speech, 2021 (https://www.aclweb.org/anthology/2021.bsnlp-1.4/)

3. (Biamonte, et al., 2017) Biamonte J., et al., «Quantum machine learning», Nature, 14.09.2017 (https://www.nature.com/articles/nature23474?error=cookies_not_supported&code=8a773c58-d06c-4938-a3a4-ea5ef568463a)

4. (Davis, 2020) Davis R., «Technology Factsheet: Deepfakes», Policy Brief, spring 2020 (https://www.belfercenter.org/publication/technology-factsheet-deepfakes)

5. (Dementieva, Panchenko, 2021) Dementieva D., Panchenko A., «Cross-lingual Evidence Improves Monolingual Fake News Detection», Association for Computational Linguistics, 2021 (https://aclanthology.org/2021.acl-srw.32/)

6. (European Commission, 2021) European Commission, «Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in AI», 21.04.2021 (https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682)

7. (Gurria, 2020) Gurria A. speech «Munich Cyber Security Conference (MCSC) Fail Safe-Act Brave: Building a Secure and Resilient Digital Society», OECD Secretary-General, 13.02.2020 (https://www.oecd.org/about/secretary-general/fail-safe-act-brave-building-a-secure-and-resilient-digital-society-february-2020.htm)

8. (Hutchinson, 2020) Hutchinson A., «Snapchat and TikTok are Both Reportedly Working on New 'Deepfake' Type Features», Content and Social Media Manager, 04.01.2020 (https://www.socialmediatoday.com/news/snapchat-and-tiktok-are-both-reportedly-working-on-new-deepfake-type-feat/569792/)

9. (Jaiman, 2020) Jaiman A., «Debating the ethics of deepfakes», Digital Frontiers, 27.08.2020 (https://www.orfonline.org/expert-speak/debating-the-ethics-of-deepfakes/)

10. (Jumper, et al., 2021) Jumper J., et al., «Highly accurate protein structure prediction with AlphaFold», Nature, 15.07.2021 (https://www.nature.com/articles/s41586-021-03819-2)

11. (Karlov, et al., 2019) Karlov D.S., et al., «Chemical space exploration guided by deep neural networks», RSC Advances, 11.02.2019 (https://pubs.rsc.org/en/content/articlelanding/2019/RA/C8RA10182E)

12. (Karlov, et al., 2020) Karlov D.S., et al., «GraphDelta: MPNN Scoring Function for the Affinity Prediction of Protein–Ligand Complexes», ACS Omega, 09.05.2020 (https://pubs.acs.org/doi/10.1021/acsomega.9b04162)

13. (Khokhlov, et al., 2022) Khokhlov I., et al., «Image2SMILES: Transformer-Based Molecular Optical Recognition Engine», Chemistry Europe, 11.01.2022 (https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cmtd.202100069)

14. (Kietzmann, et al., 2019) Kietzmann J., et al., «Deepfakes: Trick or treat?», Business Horizons, 2019 (https://www.researchgate.net/publication/338144721_Deepfakes_Trick_or_treat)

15. (Kozlovskii, Popov, 2020) Kozlovskii I., Popov P., «Spatiotemporal identification of druggable binding sites using deep learning», Nature, 27.10.2020 (https://www.nature.com/articles/s42003-020-01350-0)

16. (Kozlovskii, Popov, 2021a) Kozlovskii I., Popov P., «Protein–Peptide Binding Site Detection Using 3D Convolutional Neural Networks», J. Chem. Inf. Model., 22.06.2021 (https://pubs.acs.org/doi/full/10.1021/acs.jcim.1c00475)

17. (Kozlovskii, Popov, 2021b) Kozlovskii I., Popov P, «Structure-based deep learning for binding site detection in nucleic acid macromolecules», NAR Genomics and Bioinformatics, 26.11.2021 (https://academic.oup.com/nargab/article/3/4/lqab111/6441762?login=false)

18. (Krasnov, et al., 2021) Krasnov L., et al., «Transformer-based artificial neural networks for the conversion between chemical notations», Scientific Reports, 20.07.2021 (https://www.nature.com/articles/s41598-021-94082-y)

19. (Kurzweil, 2006) Kurzweil R., «The Singularity Is Near: When Humans Transcend Biology», Penguin Books, 26.09.2006 (https://www.amazon.com/Singularity-Near-Humans-Transcend-Biology/dp/0143037889)

20. (Lindre, 2022) Lindre Yu., Collection of articles for the 12th Russian Internet Governance Forum RIGF 2022: «Fragmentation of the Internet space and control of information flows - necessary conditions for effective manipulation of public consciousness in the interests of global IT giants», 29.09.2022

21. (Logacheva, 2021) Logacheva V., «Detoxification: NLP task showcase», Trustworthy AI Conference speech, 25.05.202. https://events.skoltech.ru/ai-trustworthy

22. (Materials, 2020, 2021) Materials of the «Trustworthy AI» conference, 2020, 2021 (https://events.skoltech.ru/ai-trustworthy)

23. (Meadows, 1998) Meadows D., «Indicators and Information Systems for Sustainable Development», The Sustainability Institute, September 1998 (https://edisciplinas.usp.br/pluginfile.php/106023/mod_resource/content/2/texto_6.pdf)

24. (Miao Fengchun, et al., 2021) Miao Fengchun, et al., «AI and education: guidance for policy-makers», UNESCO, 2021 (https://unesdoc.unesco.org/ark:/48223/pf0000376709)

25. (Morozov, et al., 2021) Morozov A., et al., «Equidistant and Uniform Data Augmentation for 3D Objects», IEEE Access, 23.12.2021 (https://ieeexplore.ieee.org/document/9662385)

26. (Perez, 2019) Perez S., «Twitter drafts a deepfake policy that would label and warn, but not always remove, manipulated media», TechCrunch, 11.11.2019 (https://techcrunch.com/2019/11/11/twitter-drafts-a-deepfake-policy-that-would-label-and-warn-but-not-remove-manipulated-media/)

27. (Popov, et al., 2019a) Popov P., et al., «Computational design for thermostabilization of GPCRs», Current opinion in structural biology, 23.03.2019 (https://www.sciencedirect.com/science/article/pii/S0959440X18301374)

28. (Popov, et al., 2019b) Popov P., et al., «Controlled-advancement rigid-body optimization of nanosystems», Journal of computational chemistry, 29.06.2019 (https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.26016)

29. (Popov, et al., 2019c) Popov P., et al., «Prediction of disease-associated mutations in the transmembrane regions of proteins with known 3D structure», PloS one, 10.07.2019 (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0219452)

30. (Prajakta, 2020) Prajakta P., «AI Deepfakes. The Goose Is Cooked?», University of Illinois Law Review, 04.10.2020 (https://illinoislawreview.org/blog/ai-deepfakes/)

31. (Sharaev, et al., 2019) Sharaev M., et al., «Functional Brain Areas Mapping in Patients with Glioma Based on Resting-State fMRI Data Decomposition», IEEE, 2019 (https://ieeexplore.ieee.org/abstract/document/8637450)

32. (Sosnin, et al., 2018a) Sosnin S., et al., «3D matters! 3D-RISM and 3D convolutional neural network for accurate bioaccumulation prediction», Journal of Physics: Condensed Matter, 19.07.2018 (https://iopscience.iop.org/article/10.1088/1361-648X/aad076)

33. (Sosnin, et al., 2018b) Sosnin S., et al., «A Survey of Multi-task Learning Methods in Chemoinformatics», Molecular Informatics, 28.11.2018 (https://onlinelibrary.wiley.com/doi/10.1002/minf.201800108)

34. (Sosnin, et al., 2018c) Sosnin S., et al., «Comparative Study of Multitask Toxicity Modeling on a Broad Chemical Space», J. Chem. Inf. Model., 27.12.2018 (https://pubs.acs.org/doi/10.1021/acs.jcim.8b00685)

35. (Sosnina, et al., 2020) Sosnina E.A., et al., «Recommender Systems in Antiviral Drug Discovery», ACS Omega, 21.06.2020 (https://pubs.acs.org/doi/10.1021/acsomega.0c00857)

36. (Tejal, et al., 2016) Tejal A. Patel MD, et al., «Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods», ACS Journals, 29.08.2016 (https://acsjournals.onlinelibrary.wiley.com/doi/full/10.1002/cncr.30245)

37. (UNESCO, 2019) UNESCO, «Beijing Consensus on Artificial Intelligence and Education», 2019 (https://unesdoc.unesco.org/ark:/48223/pf0000368303)

38. (UNESCO, 2021) UNESCO, «Recommendation on the Ethics of Artificial Intelligence», 2021 (https://unesdoc.unesco.org/ark:/48223/pf0000380455)

39. (United Nations, 2015) United Nations, «Transforming Our World: The 2030 Agenda for Sustainable Development», 2015(https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf)

40. (Urbina, et al., 2022) Urbina Fabio, et al., «Dual use of artificial-intelligence-powered drug discovery», Nature Machine Intelligence, 07.03.2022 (https://www.nature.com/articles/s42256-022-00465-9?fbclid=IwAR11_V1cd9SUxEvUfwrWMA7TUcroyYIY1nBDUL3KaS-8B4rG5MIqZCmjm0M&utm_medium=affiliate&utm_source=commission_junction&utm_campaign=CONR_PF018_ECOM_GL_PHSS_ALWYS_PRODUCT&utm_content=productdatafeed&utm_term=PID100032693&CJEVENT=25bc2722a9b311ec814303660a18050f#Fig1)

41. (Yang, 2018) Yang Sh., «Crystal structure of the Frizzled 4 receptor in a ligand-free state», Nature, 22.08.2018 (https://www.nature.com/articles/s41586-018-0447-x)

42. (Yukhno, 2022b) Yukhno A., «Digital Transformation: Exploring big data Governance in Public Administration», Public Organization Review, 13.12.2022 (https://link.springer.com/article/10.1007/s11115-022-00694-x)

43. (Zaretckii, et al., 2022) Zaretckii M., et al., «3D chemical structures allow robust deep learning models for retention time prediction», Digital Discovery, 30.08.2022 (https://pubs.rsc.org/en/content/articlelanding/2022/dd/d2dd00021k)

44. (Бородкин, 2004) Бородкин Ф.М., «Социальные индикаторы – что это такое?», Мир России, 2004 (https://cyberleninka.ru/article/n/sotsialnye-indikatory-chto-eto-takoe/viewer)

45. (Линдре, Капитанов, 2022) Линдре Ю.А., Капитанов А., «Дипфейк: невинная технология для развлечения или угроза современному обществу?», РСМД, 23.09.2022 (https://russiancouncil.ru/analytics-and-comments/analytics/dipfeyk-nevinnaya-tekhnologiya-dlya-razvlecheniya-ili-ugroza-sovremennomu-obshchestvu/)

46. (Миронова, 2021) Миронова Н.В., «Специфика цивилизационных рисков и механизмы их преодоления», Манускрипт, 2021 (https://cyberleninka.ru/article/n/spetsifika-tsivilizatsionnyh-riskov-i-mehanizmy-ih-preodoleniya/viewer)

47. (Павлов, 2016) Павлов А.В., «Архитектура вычислительных систем», Университет ИТМО, 2016 (https://books.ifmo.ru/file/pdf/2074.pdf)

48. (Федоров, Цветков, 2020а) Федоров М.В., Цветков Ю.А., «Этические вопросы технологий Искусственного Интеллекта – как избежать судьбы Вавилонской башни», CDO2DAY; Digital Russia, 11.11.2020 (https://cdo2day.ru/vid-sverhu/jeticheskie-voprosy-tehnologij-iskusstvennogo-intellekta-kak-izbezhat-sudby-vavilonskoj-bashni/), (https://d-russia.ru/jeticheskie-voprosy-tehnologij-iskusstvennogo-intellekta-kak-izbezhat-sudby-vavilonskoj-bashni.html)

49. (Федоров, Цветков, 2020б)) Федоров М.В., Цветков Ю.А., «Этика искусственного интеллекта в деятельности ЮНЕСКО», РСМД, 18.11.2020 (https://russiancouncil.ru/analytics-and-comments/analytics/etika-iskusstvennogo-intellekta-v-deyatelnosti-yunesko-voprosy-politiki-prava-i-perspektivy-ravnopra/)

50. (Федоров, Цветков, 2020а) Федоров М.В., Цветков Ю.А., «Искусственный интеллект как инструмент борьбы за сознание людей», РСМД, 21.05.2021 (https://russiancouncil.ru/analytics-and-comments/columns/cybercolumn/iskusstvennyy-intellekt-kak-instrument-borby-za-soznanie-lyudey/?sphrase_id=94902947)

51. (Федоров, Цветков, 2020б)) Федоров М.В., Цветков Ю.А., «Искусственный интеллект и социальный рейтинг: начало эпохи цифрового концентрационного лагеря «в интересах человечества»?», РСМД, 25.06.2021 (https://russiancouncil.ru/analytics-and-comments/analytics/iskusstvennyy-intellekt-i-sotsialnyy-reyting-nachalo-epokhi-tsifrovogo-kontsentratsionnogo-lagerya-v/?sphrase_id=94902947)

52. (Юхно, 2022а) Юхно А.С., «Обзор тенденций развития рынка метавселенной», Вестник Института Экономики Российской Академии Наук No. 6/2022 (https://vestnik-ieran.ru/index.php/mh-currentissue/22-stati/ekonomika-i-upravlenie/151-yukhno-a-s-obzor-tendentsij-razvitiya-rynka-metavselennoj)

53. (Юхно, Умаров, 2022) Юхно А.С., Умаров Х.С., « Перспективы развития метавселенной: эмпирические наблюдения», Управленческое консультирование, 2022 (https://www.acjournal.ru/jour/article/view/2097)

Journal of Digital Economy Research. 2023. 1. P. 6–60
DOI: 10.24833/14511791-2023-1-6-60

Received: 1 October 2022
Revised: 7 November 2022
Accepted: 15 December 2022

SOCIO-ECONOMIC ASPECTS OF THE INTRODUCTION OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES

by Maxim V. Fedorov

Abstract. The main objective of the paper is to give an overview of global effects of AI technologies, including socio-ethical principles, direct and non-direct economic impact and regulatory frameworks for developing strategies of sustainable development based on AI technologies. We will discuss these problems considering AI as a part of the global process of technological development, and, therefore, will briefly overview relationships between AI and other close fields (computational technologies, data acquisition techniques etc). A particular focus will be on global risks associated with the intensive use of AI technologies. Special attention will be given to the issues of international standardization of AI and related technologies. A section on AI-based social ranking will discuss fundamental problems inherent for such systems (biases, non-transparency etc). That section will be followed by a section on deepfakes which will be discussed in view of their dramatic effect on the conception of trust, both on individual and population/state levels. The paper will also discuss effects of widespread introduction of AI on other fields of research, such as chemical sciences and molecular biology. We will discuss pathways for sustainable development of "Trustworthy AI" which may achieve the desired balance between the benefits and risks of using these technologies and a global scale. We will discuss approaches that may lead to development of strategic principles for accessing long term effects of AI followed by relevant regulatory approaches.

Keywords: Artificial Intelligence (AI), Machine Learning, Big Data, Trustworthy AI, Deepfakes, Natural Language Processing, Information&Communication Technology (ICT), Social Ranking, Global Aspects of implementation of AI technologies (ethical, social, economical etc.); AI-powered Digital Pharma

For citation: Fedorov Maxim V. (2023) Socio-economic aspects of the introduction of artificial intelligence technologies. *Journal of Digital Economy Research,* vol. 1, no 1, pp. 6–60. (in English).
DOI: 10.24833/14511791-2023-1-6-60

Information about the author:

Maxim V. Fedorov – DSc in Chemistry, PhD in Physics and Mathematics, Corresponding member of the Russian Academy of Sciences, LLC Sintelly, Russia.
Ph.: 8 862 241 98 44, E-mail: info@siriusuniversity.ru
ResearcherID: C-7458-2011
ORCID: 0000-0003-3901-3565

Introduction – economic and social impact of AI

The widespread use of artificial intelligence (AI) over the whole planet as a consequence of the accelerated rate of digitalization of society, is leading to the 'smearing' of national borders (European Commission, 2021; Fedorov, Tsvetkov, 2020a; Fedorov, Tsvetkov, 2020b). The huge technological capabilities of AI systems have been implemented in many digital services (social networks, search engines, news aggregators, marketplaces, etc.), significantly expanding capabilities of 'traditional' information&communication technology (ICT) tools (Kurzweil, 2006; Yukhno, 2022a). Today, AI is an important part of the global digital space (Lindre, 2022; Mironova, 2021; Yukhno, 2022b; Yukhno, Umarov, 2022). However, in fact, it is still an unregulated (albeit extremely effective) tool which can provide significant geopolitical and economic advances to owners of global digital platforms and services and change directions of modern consumer, socio-political and socio-cultural trends (Fedorov, Tsvetkov, 2020b; Fedorov, Tsvetkov, 2021a; Lindre, Kapitanov, 2022. In this paper, we will overview potential risks associated with AI and discuss pathways for sustainable development of these technologies, leading to "Trustworthy AI". Analysis of trends in patents and scientific publications will be provided mostly based on analytics by Statista (www.statista.com) and Dimensions (www.dimensions.ai) as well as other sources (will be referred later).

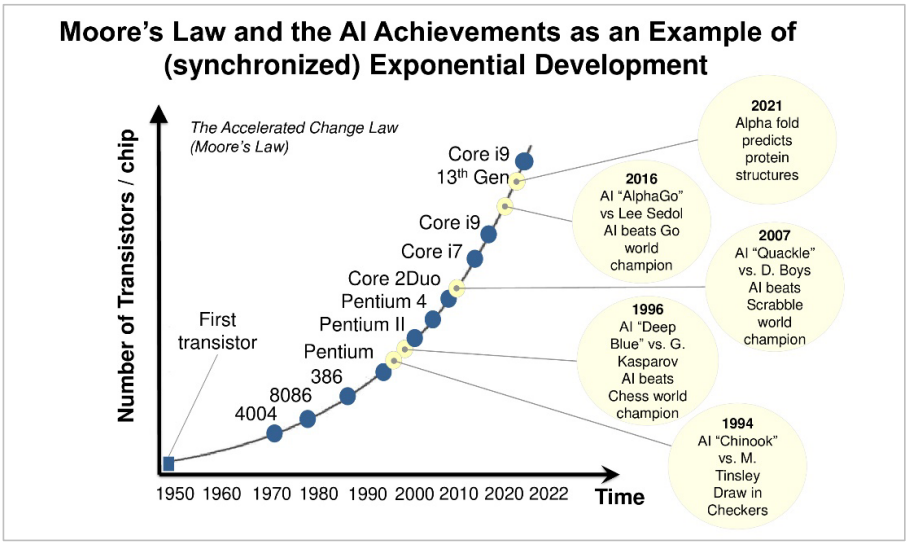

Fig. 1 Moor's law vs AI challenges

One has to note that the term "Artificial Intelligence" is somewhat ambiguous; more than 100 definitions of this term exist in expert communities, and the spectrum of understanding of this term by public is even wider (Miao Fengchun, et al., 2021). Therefore, to avoid confusion, we will follow ideas presented by Kurzweil (Kurzweil,

2006) and make a distinction between "Weak AI" and "Strong AI". By Weak AI (also called Narrow AI) we mean an AI technology limited to a specific task (e.g. face recognition or playing chess etc). We note that weak AI can mimic some particular aspects of human behavior and automate repetitive routine tasks. By Strong AI we will mean the creation of an artificial intelligence that outperforms the capacity that of a regular human being in a number of different tasks simultaneously. We note that it is hard to say whether we are going to achieve Strong AI in the near future and what would be the base of this technology, if developed. In fact, all AI technologies that are known these days can be qualified as Weak/Narrow AI. Therefore, through the manuscript, we will use the term AI meaning that it is the Weak AI if not mentioned otherwise.

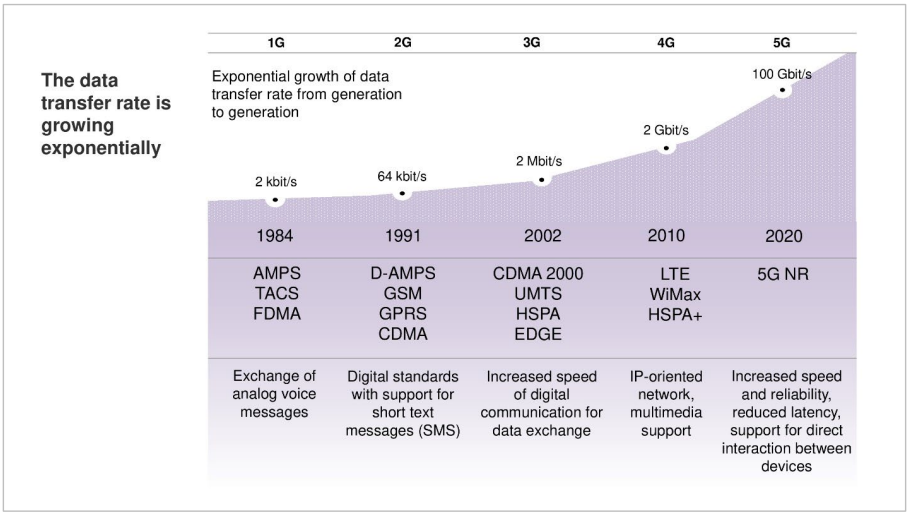

Fig. 2 Illustration of exponential increase of data transfer rate

Since the 1950s, when the term AI for the first time was introduced and became an independent object for research, hundreds thousand inventions have been patented over the world and more than 1.6 million scientific materials on the subject of AI have been published. At the same time, judging by the analytics provided by Statista (www.statista.com) and Dimensions (www.dimensions.ai) more than a half of all these AI-related patents have been published over the past decade. This exponential growth of AI applications strongly correlates with exponential increase of computational power of microprocessors (so-called Moor' law, see Fig. 1) and exponential increase of the rate of data transfer (see Fig. 2). Indeed, AI is strongly interconnected with computational technologies (Pavlov, 2016; Materials, 2020, 2021;) as well as with technologies for processing and transmission of information (see Fig. 3). These three pillars make a synergy together and thus, provide a solid background for a number of applications (Tejal, et al., 2016; Pavlov, 2016;Biamonte, et al., 2017; Sharaev, et al., 2019; Andronov, et al., 2021;Babakov, et al., 2021; Kozlovskii, Popov, 2020; Miao Fengchun, et al., 2021; Morozov, et al., 2021; Khokhlov, et al., 2022) such as E-sport, Chemical Informatics, Robotics, Smart Manufacturing etc (see Fig. 3).

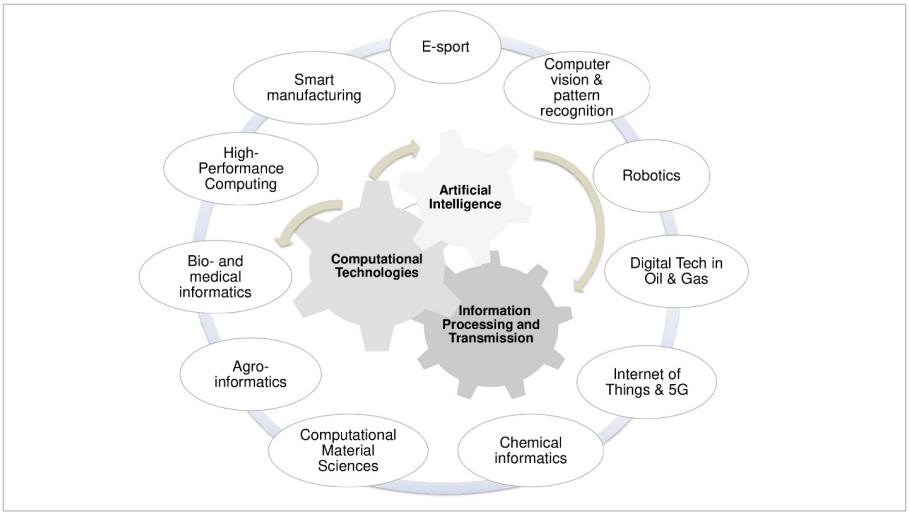

Fig.3 Areas related to AI technology

Indeed (Weak) AI itself strongly relates with data analysis and data processing techniques as well as with supporting technologies such as high-performance computing, sensors etc. Therefore, one may present the modern AI as an assembly of three components: (i) Data, (ii) AI Software (mostly Machine Learning Algorithms and corresponding libraries and frameworks) and (iii) Supporting Technologies (computational and data acquisition and transfer infrastructure etc) – see Fig. 4. The most "intelligent" part of this (arguably) is the Machine Learning frameworks; an example of a typical Machine Learning pipeline is schematically presented on Fig. 5.

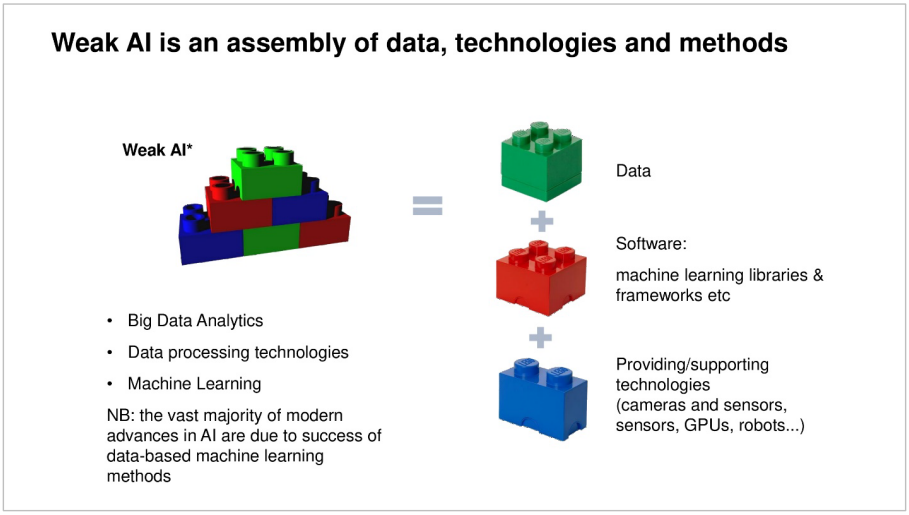

Fig.4 Main components of weak AI

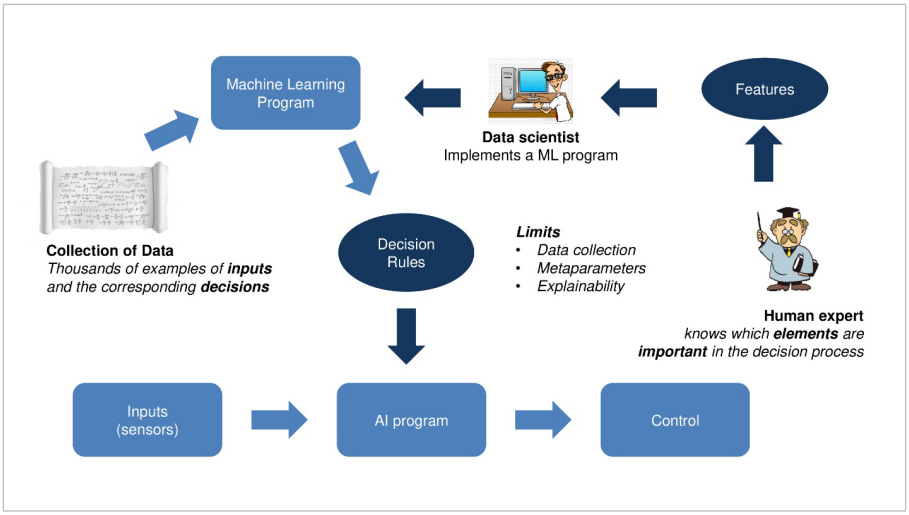

Fig.5 Overview of a Machine Learning pipeline for decision making

Analysis of patents and scientific publications in the filed also demonstrate a steady trend of shifting emphasis from theoretical research to the practical development of AI products and services for commercial purposes. Machine learning, deep learning and new neural network architectures are the most dynamically developing areas of AI technologies in recent years (Miao Fengchun, et al., 2021). Achievements in this area lead to the emergence of automated systems that can learn and perform cognitive tasks previously accessible only to people. In the field of practical applications of AI, computer vision and natural language processing demonstrate most significant progress. However, as presented by Fig. 3, there is a number of promising applications in many other areas as well.

Obviously, such technological development will have far-reaching social and cultural implications (Kurzweil, 2006; UNESCO, 2019; Gurria, 2020; Fedorov, Tsvetkov, 2021a; Fedorov, Tsvetkov, 2021b; UNESCO, 2021;European Commission, 2021). It is necessary to elaborate a system of criteria and a methodological basis for creating strategies of sustainable development for countries and all of humanity in the context of global application and development of AI technologies (UNESCO, 2019; UNESCO, 2021). We note that AI is a General Purpose technology: it transforms all economic sectors and social activities in such a way that:

- AI will boost productivity in most activities it touches;
- AI, like traditional ICT but to a larger extent, will affect income distribution, hence social equilibria;
- AI challenges the very identity of humanity, which is based on intelligence: hence it raises complex societal and ethical issues.

Following ideas of Mironova (Mironova, 2021) on classification of global risks, we may select these main groups of global risks of widespread implementation of AI.

The group of global *geopolitical* risks includes: cyberterrorism using AI, instruments of influence on the domestic and foreign policy of states, violation of the digital sovereignty of states, ever increasing threat of the use of AI tools for the creation of weapons of mass destruction (mainly chemical and biological), etc.

The group of global *economical* risks includes: the use of AI tools to provoke a financial crisis (through manipulation on the stock exchange, cryptocurrencies, etc.), high price instability in the market of computing components, monopolization of the market of computing infrastructure, instability in the labor market and shortage of personnel.

The group of global *technological* risks includes: negative consequences of scientific and technological progress, disruption of leading information systems, exponential complexity of infrastructure, accumulation of errors and accidents, etc.

The group of global *social* risks includes: potential food crisis provoked by (misused) AI-based logistics tools, the possibility of a "digital pandemic", potential real pandemics provoked by new infectious diseases "engineered" by AI, increased migration activity of the population due to remote work, social inequality generated by technology, a huge gap in the standards of living of people in "digital" and "non-digital" regions, distrust of power generated by deepfakes and other AI-based tools for social manipulations.

Existential risks – violation of the balance between management and manipulation in society, the emergence of processes of de-intellectualization and disorientation of people in the information space due to substitution of their cognitive functions by AI, suppression of their biological and psychological needs by AI and related technologies (virtual reality, robots etc).

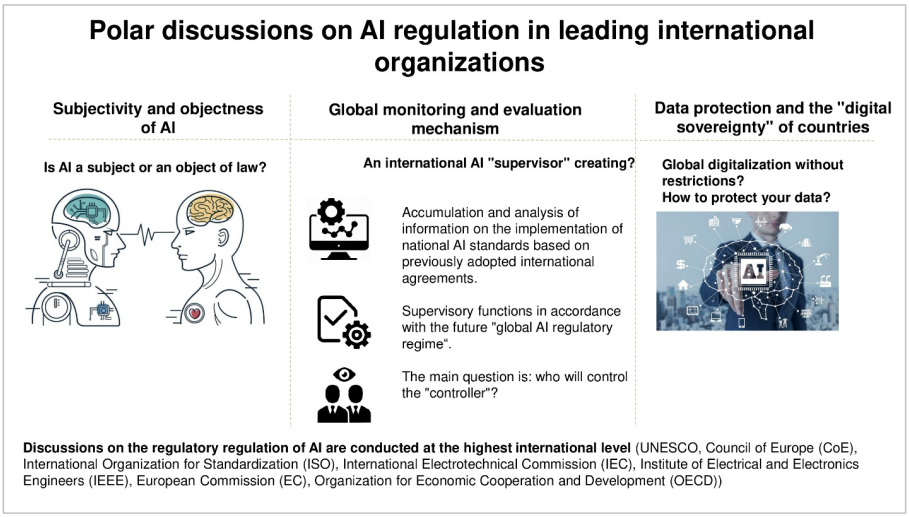

Fig.6 Discussions on the regulation of AI

To summarize, one of the main existential threats is the phenomenon of "dangerous knowledge" (Mironova, 2021). Dangerous knowledge can be defined as information obtained in the course of scientific research, development and analysis of big data by AI, which negative consequences our society cannot always effectively control. This is especially true for the use of AI in such fields as genetic engineering, nuclear physics, chemical sciences and humanitarian technologies. Ultimately, this is a problem of the ethics of science, values and the formation of humanistic ideals in the field of AI technologies. The discussion above leads us to a conclusion, that the almost unregulated development of AI, if it continues without being overseen by a proper set of regulatory instruments, may lead to catastrophic global consequences. However there are very polar debates on many national and international regulatory platforms on very basic principles of regulations and there is no clear solution how to reach the desired balance between safety and technological freedom yet (see Fig. 6). Therefore, in the next section we will discuss main international and national activities on development of regulatory frameworks for AI.

Regulation of AI technologies: still a challenge.

One of the most informative indicators of trends in technology is the content of relevant technological standards and regulatory documents (Materials, 2020, 2021). The analysis of documents created for the standardization and regulation of AI and big data technologies allows us to see the priorities of developments in various countries, as well as draw conclusions about the activities of corporations developing AI technologies.

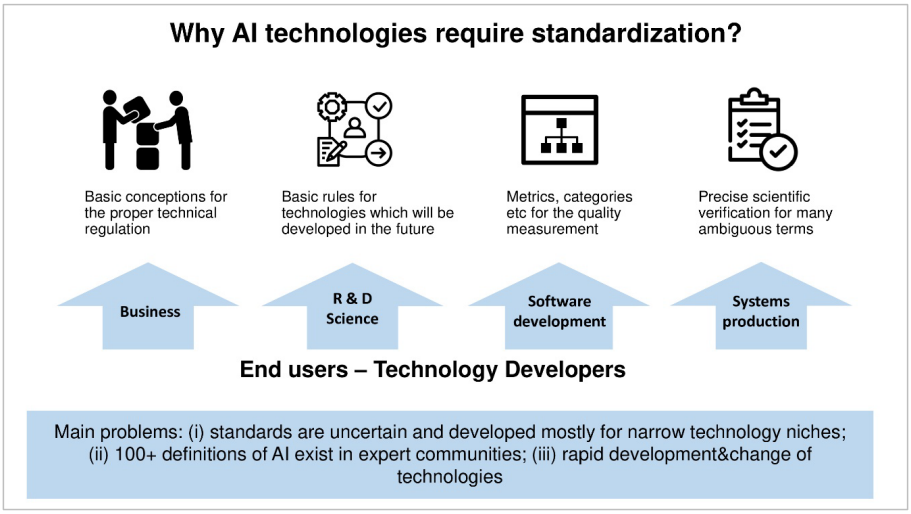

Fig.7 Main barriers between different stakeholders of AI technologies and ways of their settlement through proper standardization mechanisms

The rapid growth of the AI technology market has requested the formation of a regulatory framework for the subsequent regulation of the use of AI-based solutions with the main goal of preventing harmful effects on humans and society as a whole (see Fig. 7). At the same time, a more or less meaningful process of forming the regulatory framework started only in the 2010s, when AI systems had been already introduced in a number of fields.

The key direction in creating an integrated system of international regulation in the field of AI and end-to-end technologies is standardization, the results of which are perceived differently among the vast expert community and various groups of stakeholders (developers, regulators, engineers and users, industrial and academic sectors, government agencies). It is not always possible to maintain a balance of interests, since the technological agenda a priori has many obvious and hidden conflicts of interest that arise when prioritizing most areas of standardization. Large IT companies are essential investors in research in the field of AI standardization, including financing pilot tests and conducting painstaking work on generalizing international expert thought. Almost all editors and developers of AI standards and related documents have affiliations with global multinational corporations such as Microsoft and Alphabet.

The documents required as the output of standardization process mostly have the following format: an international standard, a technical report or a technical specification. In addition, a set of documents may be developed to form a series of standards, if the standardization area covers aspects that are clearly shared by consumers.

The main international initiatives on standardization of AI technologies have been carried out through the Artificial Intelligence Subcommittee (PC 42) and the Joint Technical Committee 1 (OTC 1) "Information Technologies" of the International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC)[1] established in 2017. To a much lesser extent, they are discussed and developed in the International Telecommunication Union (ITU). Expert support for the standardization process is also provided by the Institute of Electrical and Electronics Engineers (IEEE).

Currently, the main focus of standards development is shifted to the following areas of application of AI systems: Big Data (DB), Intelligent and Autonomous Vehicles (IATS). A special place in standardization is also occupied by issues of increasing confidence in AI systems, risk management, ethics and problems of AI bias.

Since 2018, separate work has been carried out on the terminological and conceptual standardization of AI systems – the fundamental ISO/IEC document (ISO/IEC 22989 – AI – Concepts and Terminology) for all AI-related standards that have already been adopted or are planned to be adopted.[2]

[1] https://www.iso.org/committee/6794475.html

[2] ISO/IEC DIS 22989 Information Technology — Artificial Intelligence — Artificial intelligence concepts and terminology

The formation of a series of international standards in the field of *databases* is proceeding in leaps and bounds. Despite the long overdue need to create a fundamental terminological database framework that would establish the basis for subsequent documents in this area, other technical areas and related areas of standardization, there is still no comprehensive terminological and conceptual framework. A fairly impressive group of countries (USA, China, Ireland, Korea, Japan, etc.) represented by their scientific expert and business community with objectively different starting approaches to understanding the database are extremely slow to reach some consensus decisions. Standards appear on certain aspects of the database, lagging behind the development time, for example, from documents in the field of quality of AI systems.

On the national arena, in the context of big data, Russia has adopted an analogue of the basic international standard – the preliminary national standard (PNST) – "Information technology. Big data. Typical architecture". This document focuses on the description of the most important terminology and indicates the relevant areas of application.

The technical report, which is essentially an overview of the *parameters of human "trust"* in the field of AI technology applications, was published by ISO/IEC in April 2020. The emergence of human "trust" in AI technologies is linked to ensuring their transparency, explainability, manageability, etc. at three levels: physical, cybernetic and social. As part of the international discussion on standardization, it is assumed that "trust" should be ensured at all stages of interaction and use of AI systems by humans leading to the so-called conception of "Trustworthy AI" (see Fig. 8).

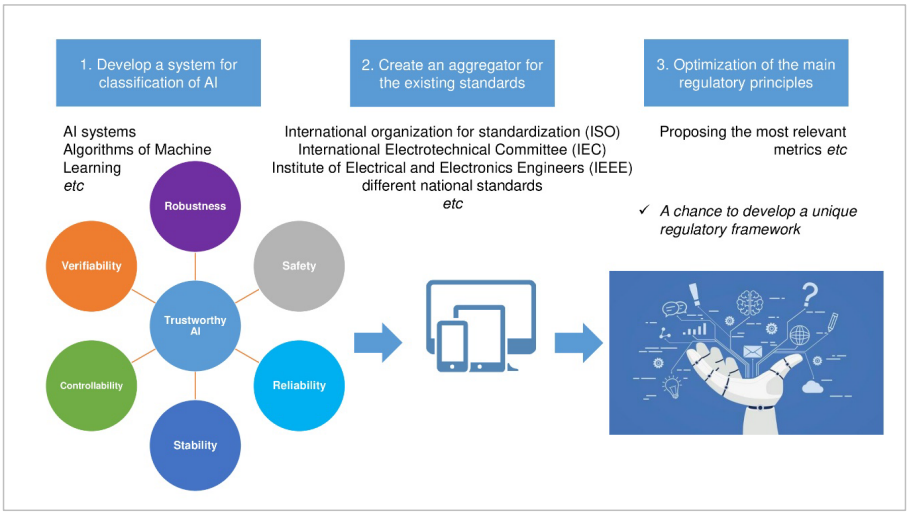

Fig.8 Components of "Trustworthy AI" and a pathway to an optimal regulatory framework
(UNESCO, 2019; UNESCO, 2021; Fedorov, Tsvetkov, 2020)

For complex systems based on machine learning, trustworthiness is generally understood to mean absolutely clear formalizable properties of the systems, such as stability, reliability (aka robustness), guaranteed convergence, security, adversarial robustness – what is known as adversarial robustness, as well as a set of properties like data privacy, that is, differential privacy (Materials, 2020, 2021).

We note that reliability, also known as robustness, is a kind of meta–property that describes the preservation of a certain basic property, such as, for example, the same level of stability in the presence of some uncertainties and disturbances in the system. Another term which is widely used to describe trustworthiness is the adversarial robustness which is resistance to adversarial attacks. The most banal example here is when some minimal distortion of visual information leads to severe misclassification: a car can mistake a pedestrian for a pole or, say, for a fly on the windshield. And it will not be classified as a kind of danger that needs to be avoided, just because the image is a little noisy (Materials, 2020, 2021).

Differential privacy, aka differential privacy, is a property that consists in the fact that the system can provide a certain level of protection of the user's personal data. Fault tolerance means the ability to maintain normal functioning, including stable, performing restrictions, and so on, in the presence of a malfunction in sensors, a malfunction in actuators, that is, control devices, and so on (Materials, 2020, 2021).

Adversarial robustness and differential privacy are relatively new concepts. These are certain properties that can be considered as subtypes of reliability and stability, but they have some specifics. Maybe there will be even more of these indicating properties in the future, taking into account the specifics of AI (Materials, 2020, 2021).

In the near future, flagship documents in the field of AI standardization will be the standard for technical regulation of artificial neural networks and the technical report on their stability characteristics adopted in 2021. The main task of these documents is to regulate the control at the stage of software product development in various industries and the rules for certification of system elements for assessing their stability, with appropriate balance of interests on the part of engineers and manufacturers. For example, the American National Standards Institute (ANSI) made a proposal, to fill in the technical specification of ISO/IEC PC 42 in the field of machine learning goals and methods and AI systems. It lists methods that make it possible to increase the explainability of AI systems for various stakeholders (developers, users, providers, regulators), and the methods differ for each group of stakeholders. For developers, it means strengthening the aspects of security and technical reliability; for users – the degree of trust in the system or method, so that bias does not play a significant role; for providers – compliance with regulatory and technical regulations, and for regulators – understanding the opportunities and limitations when creating innovative frameworks.

Functional safety has become relevant in the course of the development of machine learning methods and related AI technologies, which have built a line of interest in engineering, building a safe world based on innovative solutions. At the 6th plenary session of ISO/IEC SC 42 in April 2020, a project on the functional safety of AI systems

was proposed, which, after a short discussion, it was decided to implement in the format of the ISO/IEC SC 42 technical report. Functional security plays an important role in the introduction of mathematical models, with the help of which the application of a particular AI technology is realized. Their properties reflect the compliance of the AI technology and system with peculiar safety rules. If there are explicable and understandable dependencies between the key parameters that determine the behavior of a mathematical function, the correlation between the observed input and output data is obvious, and then the model will be transparent and effective. However, the physical phenomenon described by the model can be very complex, or have a very small scale, or cannot be observed without the influence of experimental data that cannot be described by a scientifically based model. All this causes difficulties in understanding and verifying the model, introduces ambiguities and a certain degree of uncertainty.

The technical report on the functional safety of AI serves as a description of how to control security, security risks and their relation to AI systems and technologies, also giving classes of AI technologies and the level of their possible use. Classes (1–3) reflect the possibility of greater variability when applied in various fields, from the presence of clear recommendations on safety criteria to the inability to create safety requirements at all. Levels (AD) show the degree of influence of the AI component of the system on security in general. In addition, the document used the concept of the level of automation and control, the degree of transparency and explainability of solutions (from a completely transparent system to a "black box"), as well as other aspects related, including, to the physical components of systems.

Thus, functional security can be specific and depend on the application of the AI system. This opens up the possibility of implementing its aspects using different approaches at one or another stage of the life cycle of AI systems.

The above documents were created in ISO/IEC in order to form an extensive and exhaustive base for a standard or specification for testing AI systems.

AI and Social Ranking: main risks.

Perhaps one of the most controversial fields of AI applications is the usage of AI for development of social rating systems (Fedorov, Tsvetkov, 2020a ;Fedorov, Tsvetkov, 2021b). In short, a social rating system can be defined as, a set of *social indicators evaluated by an automated system* (Borodkin, 2004; Meadows, 1998). Based on the output data (ratings or individual rating), state institutions and public institutions form a model of interaction with a particular citizen. We note that many such systems have been developed in education-related domains (Miao Fengchun, et al., 2021). Naturally, as an effective tool for data analysis, AI became a popular instrument for development of different regional and national social ranking systems. However, recent studies (Fedorov, Tsvetkov, 2020a ;Fedorov, Tsvetkov, 2021b; Lindre, 2022) which analyze effects of social ranking systems introduced in different parts of the planet revealed several fundamental problems of such systems which we will briefly overview below.

The main problems of the concept of social ratings are listed below.

1. Opacity of rules and algorithms based on which the social rating system will evaluate human behavior. Modern systems of social rating are not transparent as they typically hinder the information about both development process of rules and criteria for evaluating human actions by an intellectual system as well as names of the developers. That raises the question about the degree of responsibility of the developers and owners of such systems for making decisions that may be of crucial importance for millions of fellow citizens? At the same time, it remains unclear how the social rating system, which uses artificial intelligence (AI) to process huge amounts of citizens' data across the country or a single society, is actually making decisions as most of AI algorithms operate on the principle of a classic "black box". This means that the output results do not imply an explanation of the reasons and circumstances for which the AI system draws conclusions and makes decisions. People can only take for granted the fact that an "intelligent" machine in real time "sorts" them into groups depending on their behavior and committed actions, even if such actions and decisions are not contrary to the law.

2. No one is immune from social deranking. Some lobbyists of the social rating system (naively) believe that following certain patterns of a "good citizen" is a guarantee of finding a person in the upper privileged group. However, like any other complex technology, an AI-based social rating is able not only to track the general background around a person based on a certain system of indicators, but also to go further – to form an "opinion" about such a person, taking into account the multiplicative effect of the perception of human behavior by the whole society. For example, a popular media personality (politician, artist or athlete), on the one hand, has more opportunities to maintain their social rating at a high level due to the cumulative effect of his positive image. On the other hand, with a planned information campaign to discredit such a person, the "intellectual" social rating system will take into account the perception of this person in society that is changing for the worse, and it will be more difficult for such a person to "correct" his rating with "good deeds" later. The system will give an average assessment of a person taking into account "good deeds" and "bad" image in the media space.

3. The social rating system may have a large number of hindered outcomes. The distribution of people into categories depending on the current social rating may have far-reaching consequences in all spheres of life. For example, a person with an insufficiently high rating is unlikely to be hired for a prestigious job, and in the end this person will not even understand why their job application was refused. The fundamental right of every person – the right of freedom to choose one's life trajectory will be proportionally limited to the designated part of the 'phase space' occupied in the person's social rating. An unpaid fine for traffic violation or an unfulfilled "children's program" may manifest itself in the most unexpected way in one's future and play a decisive role in the fate of the person in the implementation of their plans for life.

4. Digital discrimination of people. The adoption of a social rating system at a state level in a country would actually mean the emergence of a "digital code of law" and a "digital procedural code" in this country which will be parallel to the Constitution of the country and other regulatory legal acts (a set of social indicators and algorithms for collecting, analyzing and evaluating information about a person included in the AI system). Therefore, instead of giving people equal rights and responsibilities, such a system would introduce a full-fledged mechanism of discrimination and restriction of human rights which may arise without committing illegal acts, which are such under the current legislation.

There is a direct threat that the social rating will turn into a dictatorship of a certain system of values introduced by unknown programmers into an automated system that works according to some criteria. This, in particular, is a direct violation of the seventh article of the Constitution of the Russian Federation, which guarantees the free development of a person.

5. The social rating system may lead to social destabilization. The introduction of even individual elements of social rating can be very criminogenic – because such a system will automatically provide illegal ways of "tweaking" the rating for a moderate fee and ending up in a category of higher level. In addition, the social rating is a clear provocation of public tension. It will obviously cause a protest mood in society, which directly contradicts the modern understating of the function of a democratic state where one of the most important tasks is to create conditions for mutual trust between it and society.

In this context, the question of a person's "transit" from one category of social rating to another remains open. Such systems provide an opportunity for a direct threat of legal "enslavement" of a person, i.e. a lifetime at the bottom of the rating pyramid due to an objective narrowing of opportunities for the latter to improve their performance. Initially, the discrimination-oriented concept of social rating will infringe on the rights of persons who, although they do not violate the law, do not "earn" enough points of a "good citizen".

It is noteworthy that the risks of social rating systems were identified and realized by the governing structures of the European Union at an early stage. In April 2021, the European Commission proposed to ban the introduction of AI technologies (resolution "On the European Approach to artificial Intelligence"), which are used for "mass surveillance, applied in a generalized form to all individuals without any differences." Surveillance methods such as "monitoring and tracking of individuals in a digital or physical environment, as well as automatic aggregation and analysis of personal data from various sources" will become illegal[3].

[3] https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682

If this initiative is approved by the European Parliament and the European Council, the European Union will completely ban the use of "high-risk" AI systems and restrict the use of others if they do not meet the new standards. It is noteworthy that the "high risk" category includes, among other things, AI technologies in robotic surgery, AI in software for hiring employees, AI for checking documents and verifying evidence (in court proceedings), as well as AI for assessing the *credit rating* of citizens – the primary element of the entire social rating system. Moreover, the explanation to the European Commission document notes that the very essence of social rating and the use of AI in applications that manipulate human behavior to circumvent their will are unacceptable (European Commission, 2021). Business representatives who ignore these new rules may be fined up to 20 million euros or 4% of the annual turnover.[4]

We believe that it is time to initiate full-fledged scientific research on the phenomenon of social rating on global scale. The results obtained should be submitted for a wide public discussion with the participation of representatives of all sectors of society (scientific and expert community, human rights organizations, legislators, business circles, political parties, civil society, etc.).

Deepfakes – what is truth?

Deepfakes can be defined as synthetic auditory or visual media developed using deep machine learning methods that create such a realistic impression that they can deceive the audience (Hutchinson, 2020; Jaiman, 2020; Lindre, Kapitanov, 2022; Prajakta, 2020). This is the main task of the authors of deepfakes – to make as realistic illusion as possible. Synthetically created media vary widely in technical complexity and application, ranging from low-quality "cheap deepfakes" to high-quality products that can influence a person's "perception of reality", as well as in a certain sense alter the decision making process of the person

Of course, the overall idea of creating of synthetic audiovisual content is not new: Filmmakers have been using computer-generated images (CGI) since the 1970s in order to enhance. However, these methods were very expensive and only a few companies had access to such technologies. These days, the development of digital technologies has made complex synthetic media inexpensive and easy to produce (especially thanks to the proliferation of free and open source software). As a consequence, millions of people create such content every day using their home computers and even smartphone (Davis, 2020).

[4] https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682

Deepfakes can be successfully and usefully used in such areas as education and art, but it turned out that deepfakes are more likely to harm individuals, business, society and democracy as a whole, and also accelerate the process of falling public confidence in the media (Jaiman, 2020;Logacheva, 2021; Babakov, et al., 2021; Dementieva, Panchenko, 2021). Such an erosion of trust will contribute to the flourishing of de facto relativism, destroying the structure of democracy and civil society under strong pressure of false information.

The grand scale of using the deepfakes by many stakeholders (including state agencies, companies and individuals) pose a number of challenges to modern society. The most growing general concern is about how deepfakes contribute to the evolution of the era of Internet disinformation (Perez, 2019). Indeed, as tools for creating synthetic information for a mass audience are being constantly improved at the technical level, regulatory agencies across the globe are trying to solve simultaneously emerging socially significant problems associated with an adequate response to new risks and threat provide by deepfakes (mainly the high rate of distrust to public media).

For instance, there is an important question on protecting the rights of individuals in the sphere of control over the commercial use or misuse of their images and other aspects of "digital" identity (Jaiman, 2020). The methodology of creating deepfakes has already developed to such an extent that many people may fully recreate their image or somebody's else image in digital form in order to use these 'updated' images in various kinds of virtual projects in which they do not take direct "personal" participation (Davis, 2020). In this case what would be the fee for the commercial use of their images and the image in general, who will posses the rights of using the new image (which can be of considerable value)?

Even more importantly, in the framework of legal proceedings, deepfakes can pose a serious threat to determining the authenticity of the most important evidence, since the continuous development of technology significantly complicates the process of separating real digital evidences (images, video, audio etc) from forgery and, therefore compromises the use of digital evidences (Kietzmann, et al., 2019; Prajakta, 2020). For example, some courts in the United States have allowed taking photographic evidence without verifying the reliability of eyewitness testimony in accordance with the so-called the "silent witness rule" (the use of "substitutions" when accessing confidential information in the United States in the open jury trial system). While maintaining the current dynamics of the spread of deepfakes, it may become a common practice for lawyers to declare that the digital evidence against their client is a fake. This circumstance may cause the jury to doubt the authenticity of real evidences (Prajakta, 2020).

Another problematic area of using deepfakes is the field of cybersecurity (Gurria, 2020). Indeed, falsified data and digital trace of a person provide unprecedented risks. For example, sensor data may be falsified and then transferred to an AI decision making systems, which will lead to the deception of the artificial intelligence system with the subsequent adoption of incorrect decisions by it (Davis, 2020).

A comprehensive system of protecting society from the destructive use of deepfakes begins with software designed to detect deepfakes or increase the technical ability to distinguish between real and fake content. Such technical solutions include both AI systems trained to recognize anomalies characteristic of deepfakes, and cryptographic methods that can be integrated into video and audio recording equipment in order to detect unauthorized access. However, the means of detecting and countering deepfakes will always lag behind the pace of development of the deepfake technologies themselves. In addition, the solution to the following problem is not yet visible: simply knowing that a certain video file has been processed does not provide information on how to identify the creators and hold them accountable (Davis, 2020; Jaiman, 2020).

The second level of counteraction to destructive deepfakes is legislative regulation. For instance, a number of states in the United States, for example, Virginia, have adopted local laws criminalizing deepfakes used in pornography; and the state of Texas has additionally criminalized deepfakes created to influence the electoral process. From another side, the states of Massachusetts and California have not been able to approve bills aimed at countering deepfakes, due to concerns about excessive "overregulation" of the IT industry (Davis, 2020; Jaiman, 2020).

Perhaps, the best result would be the development of technical and legal tools for managing the risks of harm from deepfakes that do not suppress innovation and business activity, as well as do not encroach on freedom of expression. Responsibility should come for individuals without the introduction of universal bans on new technology.

An urgent and practically unsolvable problem in most countries of the world today is that state, primarily regulatory authorities, do not take sufficient actions to ensure the authenticity of audiovisual content distributed in the public space. Encouraging digital platforms to self-censor and self-identify destructive content using deepfake technologies has a very limited effect. For example, if the origin of the video cannot be traced, national governments are recommended to create a body that can detect deepfakes using blockchain technology. The essence of the experts' proposal is that block chains store data in a decentralized network in which anyone can verify the originality of the information by comparing it with a certain unique and unchangeable key. Even the slightest manipulation of data will lead to an easily detectable discrepancy.

The technological approach to managing and limiting malicious deepfakes is to develop methods and methods for authenticating authentic content. An example is a technology called AmberAuthenticate, which works on devices that reproduce original, i.e. authentic photos, audio and video content in real time as content is recorded. The program creates the so-called "truth layer", which is the original content cryptographically marked with numerous digital fingerprints, and then archived in a publicly accessible block chain. This fingerprinting of digital content is used to track its origin as it spreads in the digital space and to help detect and respond to attempts to manipulate the original content.

In another part of the planet, Cyberspace Administration of China has developed rules (enforced on January 1, 2020) prohibiting any publication in the digital information space with modified content created using AI technologies without appropriate warning. These measures were taken for reasons of national security in order to prevent the replication of fake news and the spread of political disinformation at an early stage (Jaiman, 2020).

The new rules allow law enforcement agencies to prosecute individuals who create deepfakes and video platforms on which they are posted. The document does not specifically mention the term deepfake — instead, it is prohibited to publish generally misleading content that could damage the reputation, for example, of a political figure or deceive voters (Babakov, et al., 2021). At the same time, the rules for regulating deepfakes are lengthy in nature, for violation of which no specific administrative or criminal measures are formally provided (Jaiman, 2020).

At the end of the section we would like to cite the proposals from the study of Indiana University "Deepfakes: Trick or treat?" on how to regulate deepfakes and combat malicious content based on this technology (Kietzmann, et al., 2019).

1. Fixing the source content that guarantees the detection of the unreliability of the deepfake. An effective fight against deepfakes created for the purpose of deception or manipulative influence on a person involves a system of early detection and blocking of such content. In this regard, technologies that track and record a person's life, including his geolocation, interaction with other people and institutions, as well as information about any other "external" activity, are becoming in demand. Despite the direct threat to privacy, privacy and confidentiality of human actions, from a technical point of view, collecting such information from modern mobile devices seems quite simple and convenient. However, such personal and sensitive data should be encrypted, stored and used for comparative analysis only if necessary in order to "expose" deepfakes.

2. Exposing malicious deep fakes. Along with the development of AI technologies that simplify the process of creating and improving the final result of audiovisual deepfakes, there are also technological innovations being developed for the detection and classification of deepfakes. The leaders in this area are the United States. For example, the Agency for Advanced Research Projects of the US Department of Defense (DARPA) has innovative means of detecting deepfakes, in particular, the program for forensic examination of digital media content, as well as deepfake detection services such as Truepic. Global corporations and digital platforms are also investing serious resources in the identification and detection of deepfakes.

3. Protection of rights within the framework of the law. Victims of deepfakes must have a system of protection provided by law against material or moral harm caused to them in the event of defamation, malicious intent, violation of privacy or emotional distress caused by deepfakes, as well as in cases of copyright infringement, imitation and fraud related to deepfakes.

4. Taking confidence-building measures. Global companies that claim to adhere to universal moral principles and human rights, in the context of deepfakes, should redouble their efforts to build trust with their customers and establish strong emotional ties with them. In this case, if companies or their brands become involved in scandals on the basis of diplomatic threats, traditional customers and partners are likely to take the information noise critically and maintain their loyalty.

AI in Digital Pharma and Molecular Biology

Pharmaceutical sciences and molecular biology undergo a "phase transition" due to the widespread usage of data-intensive machine learning techniques in these fields (Sosnin, et al., 2018b; Yang, 2018; Kozlovskii, Popov, 2020; Sosnina, et al., 2020; Zaretckii, et al., 2022). From a distant point of view, it might seem strange because these fields have been traditionally associated with heavy usage of experimental techniques and clinical trials. However, modern research programs in biomolecular sciences have at their cores experimental (e.g. data from clinical trials) as well as computational data-generators (e.g. data on molecular modelling); thus data analysis and machine learning research based on the generated data may greatly contribute to the progress in these fields. Digital Pharma is one of the remarkable examples, where in vitro or in silico High-Throughput-Screening platforms are used to derive predictive models which can be applied to solve fundamental and practical problems in modern drug discovery and molecular design (Sosnin, et al., 2018b; Sosnina, et al., 2020; Andronov, et al., 2021; Khokhlov, et al., 2022).

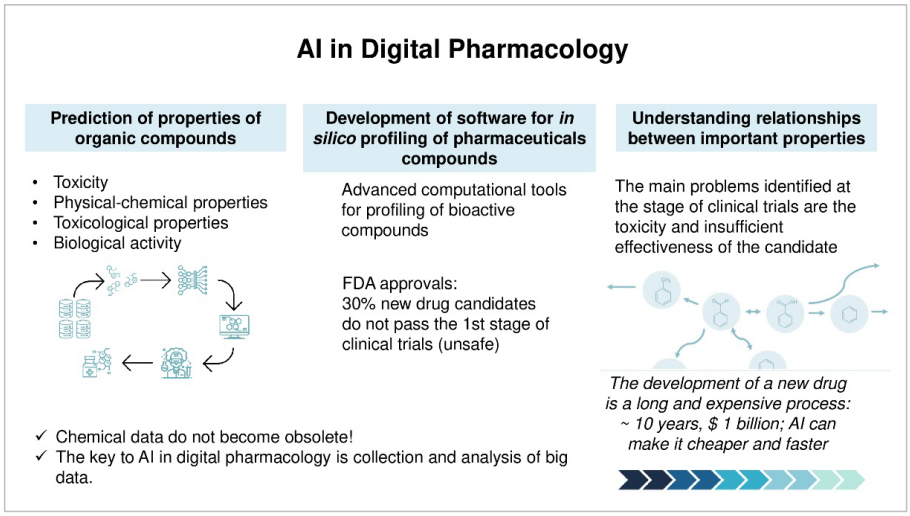

Fig.9 Digital pharmacology problems
(Andronov, et al., 2021; Karlov, et al., 2019; Karlov, et al., 2020; Khokhlov, et al., 2022; Krasnov, et al., 2021; Sosnin, et al., 2018a; Sosnin, et al., 2018b; Sosnin, et al., 2018c; Sosnina, et al., 2020d)

Every year pharmacological companies across the globe spend hundreds billions of US dollars on drug development and design. To cope with the crisis related to the exponential increase of the costs of R&D in these areas, one needs to develop approaches, and data-intensive molecular technologies hold a great promise for this. Indeed, recently a number of works on applications of AI in digital pharma and molecular biology have been published in the high-ranking journals which provide excellent examples of a cross-over between state-of-the-art data-intensive computational methods and "wet" biology (Young, et al, 2018; Popov, et al., 2019a; Popov, et al., 2019b; Popov, et al., 2019c; Karlov, et al., 2020; Kozlovskii, Popov, 2020; Kozlovskii, Popov, 2021a; Kozlovskii, Popov, 2021b; Morozov, et al., 2021; Zaretckii, et al., 2022).

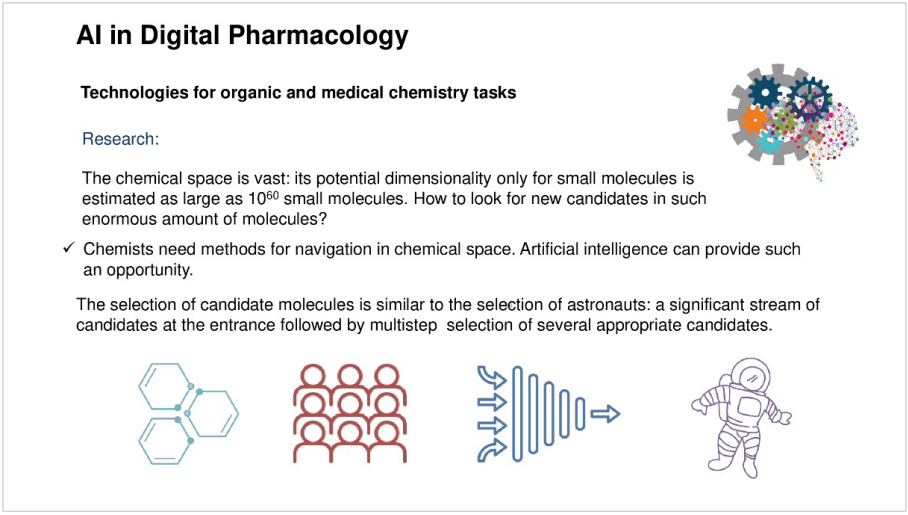

Fig.10 Tasks for AI in digital pharmacology

Progress in experimental techniques have already resulted in accumulation of a large amount of corresponding data and the speed of generation of new data is still growing (Sosnin, et al., 2018b). Effective integration of large-scale data analysis with "wet" biomolecular sciences require development of new approaches and new tools, as well as qualified researchers and engineers with both, computer science and chemical/biological backgrounds, to integrate AI-based tools with experimental research programs. Moreover, the chemical space has a complex structure and vast dimensionality; therefore, straightforward machine learning pipelines typically have a very limited applicability domain, making them useless in practice. Using state of the art data-intensive statistical analysis techniques along with molecular modelling and machine learning approaches allows one to expand the applicability domain of 'pure' machine learning methods and develop more robust solutions (Sosnin, et al., 2018a). This strategic area aims to develop the end-to-end technologies using knowledge-, sequence-, graph-, and structure- and multidimensional-based representations of intensive molecular data (Andronov, et al., 2021).

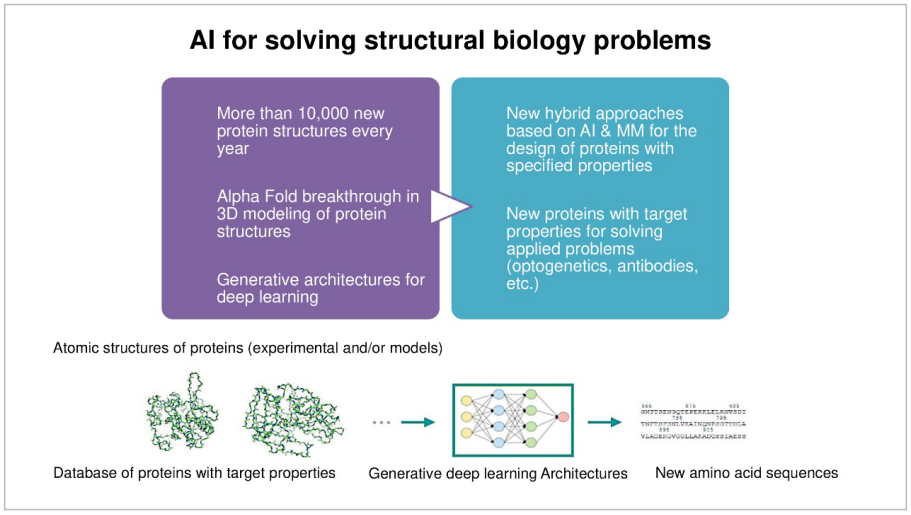

Fig.11 Structural biology problems

(Popov, et al., 2019a; Popov, et al., 2019b; Popov, et al., 2019c; Karlov, et al., 2020; Kozlovskii, Popov, 2020; Kozlovskii, Popov, 2021a; Kozlovskii, Popov, 2021b; Morozov, et al., 2021; Zaretckii, et al., 2022)

The two key strategic directions in the field are:

- develop and apply AI-based tools to explore target regions in chemical space and generate new chemicals with desired properties;
- develop and apply AI-based tools to explore target regions in biological space with a focus on molecular biology and molecular mechanisms of drug-macromolecule and macromolecule-macromolecule interactions.

Figures 9–10 illustrate the first approach. Figures 11–12 illustrate the second approach.

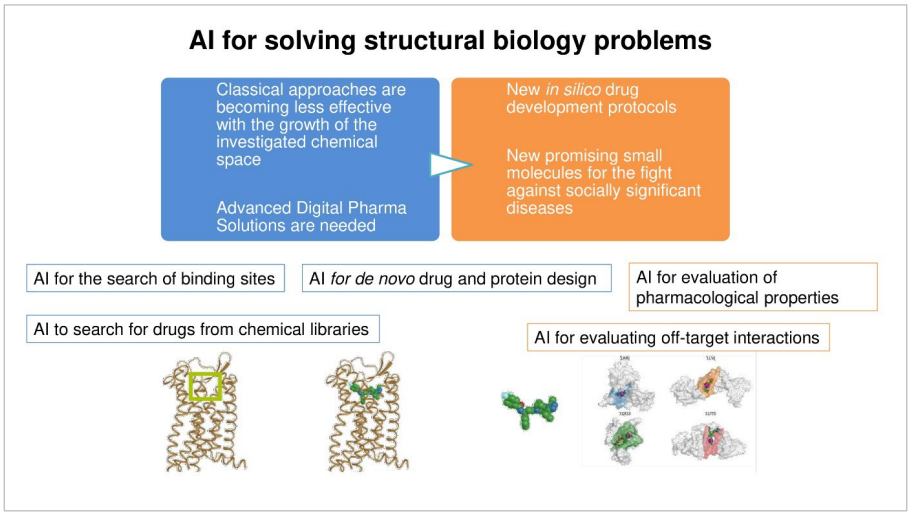

Fig.12 Tasks for AI in structural biology

Potentially, one may estimate that AI applications in bimolecular sciences may (i) reduce the costs of R&D for development and design by 30–50 percent; (ii) reduce the time for development of new drugs and vaccines by 20 (pessimistic) to 70 (optimistic) percent; (iii) greatly reduce potential risks of new drugs (side effects, stability issues etc) due to the thorough in silico validation of drug candidates. Therefore, if widely introduced, these tools can potentially save equivalents of hundreds billion of US dollars per year and, more importantly, save millions of human lives and greatly contribute to the improvement of human health on the whole planet. However, despite of the positive effects of the already many applications of AI-based techniques in digital molecular sciences, there are raising concerns about potential risks related with the implementation of these tools. For instance, recent study has shown that AI-based methods can be used for rapid generation of highly poisonous compounds (Urbina, et al., 2022). Therefore, the dual nature of AI again reveals itself in these areas as well, which raise questions about proper regulations of applications of AI methods in interdisciplinary domains.

Discussion

Due to the many social and economic effects of these technologies, a number of AI-related topics been included in the political and humanitarian agenda of the UN in recent years (UNESCO, 2019; UNESCO, 2021). A significant role in this was played by the Sustainable Development Goals (SDGs) adopted at the summit of the World Organization in 2015, which are a list of 17 global goals and 169 tasks (United Nations, 2015). The successful implementation of most of them is connected with the need to consolidate the efforts of mankind to achieve the new level of development of the society. It is assumed that the transition from the third to the fourth industrial revolution (characterized by the development and mass introduction of cyber-physical systems into global production) will allow one to accumulate and use additional resources in order to stimulate the growth of the world economy, the widespread elimination of poverty, reducing the negative human impact on the environment, etc. The prerequisites for such a transformation are the introduction of a fundamentally new technological base into all key spheres of society's life that can accelerate the evolution of the main socio-economic processes, as well as create new ones and increase the efficiency of existing production chains. The leading role in this is given to "smart" high-tech systems that use AI or its individual components in their work. However, as stated in the introduction, mass-scale introduction of AI poses a number of existential risks, which may affect the fate of the whole civilization. At the same time, in the area of assessment of risks associated with AI we need to transfer from discussions and postfactum observations to fundamental research aimed to development of a set strategic principles for proactive risk assessment and modelling/evaluating corresponding measures to mitigate these risks. The tasks is challenging indeed because of the exponential speed of development of AI and Big Data technologies (see Fig. 1 and Fig. 2 for illustration).

However, one may learn from the history of pharmaceutical sciences which during the 20–21st centuries have evolved from an unregulated wild field to the stage where every steps of the drug development process are strictly regulated based on both, science&technology as well as ethics. It is timely to remember the oldest principle of biomedical ethics, *primum non nocere* ('first, do no harm'), which can be applied as the main approach for regulating development of AI technologies. Overall, despite of the many distinctive features of AI, the overall problem of regulating potentially dangerous technologies (e.g. nuclear tech, cloning, development of bioactive compounds etc) is not something unique for the humankind. Many useful practices and policies from other fields can be adopted which can optimize resources spent on development of regulatory framework for AI. Such an approach may lead to the development of an adaptive regulatory framework which will be aligned with the principles of transparency, cooperation, accountability and ethics in order to facilitate safe and sustainable development of AI-based innovative products.

Gratitude

The author acknowledges the great help of Ksenia Polyakova (Sirius University) for technical assistance with preparation of the manuscript. Numerous discussions with Yuriy Landre (Skoltech) and Anna Zhdanova (Skoltech) on different aspects of global effects of AI are very much appreciated; many ideas presented in the manuscript have been inspired by these discussions. Many thanks are to Natalia Struskevich (Skoltech), Petr Popov (Skoltech) and Sergey Sosnin (Syntelly) for joint works and discussions on AI in biomolecular sciences and digital pharma. The author is thankful to the many of his colleagues from Skoltech and to the members of the Committee on Ethics of Artificial Intelligence at the Commission of the Russian Federation for UNESCO for discussions on different AI-related topics. Special thanks are to Alexander Kuleshov, President of Skoltech (Chair of the Committee). The author appreciates discussions with Igor Ashmanov (Kribrum), Natalia Kasperskay (Infowatch) and Stanislav Ashmanov (Nanosemantica) on deep fakes and risks associated with digital technologies.

References

1. Andronov M., Fedorov M., Sosnin M. "Exploring Chemical Reaction Space with Reaction Difference Fingerprints and Parametric t-SNE". // ACS Omega. – 2021. URL: https://pubs.acs.org/doi/10.1021/acsomega.1c04778

2. Babakov N., Logacheva V., Kozlova O., Semenov N., Panchenko A. "Detecting Inappropriate Messages on Sensitive Topics that Could Harm a Company's Reputation". // Trustworthy AI Conference speech. – 2021. URL: https://www.aclweb.org/anthology/2021.bsnlp-1.4/

3. Biamonte J., Wittek P., Pancotti N., Rebentrost P., Wiebe N. & Lloyd S. "Quantum machine learning". // Nature. – 2017. URL: https://www.nature.com/articles/nature23474?error=cookies_not_supported&code=8a773c58-d06c-4938-a3a4-ea5ef568463a

4. Borodkin F.M., "Social indicators — what are they?", "Universe of Russia". // Mir Rossii. – 2004. URL: https://cyberleninka.ru/article/n/sotsialnye-indikatory-chto-eto-takoe/viewer

5. Davis R. "Technology Factsheet: Deepfakes". // Policy Brief, spring. – 2020. URL: https://www.belfercenter.org/publication/technology-factsheet-deepfakes

6. Dementieva D., Panchenko A. "Cross-lingual Evidence Improves Monolingual Fake News Detection". // Association for Computational Linguistics. – 2021. URL: https://aclanthology.org/2021.acl-srw.32/

7. European Commission, "Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in AI". – 2021. URL: https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682

8. Fedorov M., Tsvetkov Yu. "Ethical issues of artificial intelligence technologies – how to avoid the fate of the Tower of Babel". // CDO2DAY; Digital Russia. – 2020. URL: https://cdo2day.ru/vid-sverhu/jeticheskie-voprosy-tehnologij-iskusstvennogo-intellekta-kak-izbezhat-sudby-vavilonskoj-bashni/, https://d-russia.ru/jeticheskie-voprosy-tehnologij-iskusstvennogo-intellekta-kak-izbezhat-sudby-vavilonskoj-bashni.html

9. Fedorov M., Tsvetkov Yu. "Ethics of artificial intelligence in UNESCO activities". // RIAC. – 2020. URL: https://cdo2day.ru/vid-sverhu/jeticheskie-voprosy-tehnologij-iskusstvennogo-intellekta-kak-izbezhat-sudby-vavilonskoj-bashni/

10. Fedorov M., Tsvetkov Yu. "Artificial intelligence as a tool of struggle for people's consciousness". // RIAC. – 2021. URL: https://russiancouncil.ru/analytics-and-comments/columns/cybercolumn/iskusstvennyy-intellekt-kak-instrument-borby-za-soznanie-lyudey/?sphrase_id=94902947

11. Fedorov M., Tsvetkov Yu. "Artificial intelligence and social rating: the beginning of the era of the digital concentration camp "in the humanity interests"?". // RIAC. – 2021. URL: https://russiancouncil.ru/analytics-and-comments/analytics/iskusstvennyy-intellekt-i-sotsialnyy-reyting-nachalo-epokhi-tsifrovogo-kontsentratsionnogo-lagerya-v/?sphrase_id=94902947

12. Gurria A. speech "Munich Cyber Security Conference (MCSC) Fail Safe-Act Brave: Building a Secure and Resilient Digital Society". // OECD Secretary-General. – 2020. URL: https://www.oecd.org/about/secretary-general/fail-safe-act-brave-building-a-secure-and-resilient-digital-society-february-2020.htm

13. Hutchinson A. "Snapchat and TikTok are Both Reportedly Working on New 'Deepfake' Type Features". // Content and Social Media Manager. – 2020. URL: https://www.socialmediatoday.com/news/snapchat-and-tiktok-are-both-reportedly-working-on-new-deepfake-type-feat/569792/

14. Jaiman A. "Debating the ethics of deepfakes". // Digital Frontiers. – 2020. URL: https://www.orfonline.org/expert-speak/debating-the-ethics-of-deepfakes/

15. Jumper J., et al. "Highly accurate protein structure prediction with AlphaFold". // Nature. – 2021. URL: https://www.nature.com/articles/s41586-021-03819-2

16. Karlov D., Sosnin S., Tetko I., Fedorov M. "Chemical space exploration guided by deep neural networks". // RSC Advances. – 2019. URL: https://pubs.rsc.org/en/content/articlelanding/2019/RA/C8RA10182E

17. Karlov D., Sosnin S., Fedorov M., Popov P. "GraphDelta: MPNN Scoring Function for the Affinity Prediction of Protein–Ligand Complexes". // ACS Omega. – 2020. URL: https://pubs.acs.org/doi/10.1021/acsomega.9b04162

18. Khokhlov I., Krasnov L., Fedorov M., Sosnin S. "Image2SMILES: Transformer-Based Molecular Optical Recognition Engine". // Chemistry Europe. – 2022. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cmtd.202100069

19. Kietzmann J., Lee Linda W., McCarthy Ian Paul, Kietzmann Tim C. "Deepfakes: Trick or treat?". // Business Horizons. – 2019. URL: https://www.researchgate.net/publication/338144721_Deepfakes_Trick_or_treat

20. Kozlovskii I., Popov P., "Spatiotemporal identification of druggable binding sites using deep learning". // Nature. – 2020. URL: https://www.nature.com/articles/s42003-020-01350-0

21. Kozlovskii I., Popov P., "Protein–Peptide Binding Site Detection Using 3D Convolutional Neural Networks". // J. Chem. Inf. Model. – 2021. URL: https://pubs.acs.org/doi/full/10.1021/acs.jcim.1c00475

22. Kozlovskii I., Popov P, "Structure-based deep learning for binding site detection in nucleic acid macromolecules". // NAR Genomics and Bioinformatics. – 2021. URL: https://academic.oup.com/nargab/article/3/4/lqab111/6441762?login=false

23. Krasnov L., Khokhlov I., Fedorov M., Sosnin S. "Transformer-based artificial neural networks for the conversion between chemical notations". // Scientific Reports. – 2021. URL: https://www.nature.com/articles/s41598-021-94082-y

24. Kurzweil R. "The Singularity Is Near: When Humans Transcend Biology". // Penguin Books. – 2006. URL: https://www.amazon.com/Singularity-Near-Humans-Transcend-Biology/dp/0143037889

25. Lindre Yu., Kapitanov A. "Deepfake: an innocent technology for entertainment or a threat to modern society?". // RIAC. – 2022. URL: https://russiancouncil.ru/analytics-and-comments/analytics/dipfeyk-nevinnaya-tekhnologiya-dlya-razvlecheniya-ili-ugroza-sovremennomu-obshchestvu/

26. Lindre Yu. Collection of articles for the 12th Russian Internet Governance Forum RIGF 2022: "Fragmentation of the Internet space and control of information flows – necessary conditions for effective manipulation of public consciousness in the interests of global IT giants". – 2022.

27. Logacheva V., "Detoxification: NLP task showcase". // Trustworthy AI Conference speech. – 2021.

28. Materials of the "Trustworthy AI" conference. – 2020, 2021. URL: https://events.skoltech.ru/ai-trustworthy

29. Meadows D., "Indicators and Information Systems for Sustainable Development". // The Sustainability Institute. – September 1998. URL: https://edisciplinas.usp.br/pluginfile.php/106023/mod_resource/content/2/texto_6.pdf

30. Miao Fengchun, Holmes, Wayne, Ronghuai Huang, Hui Zhang "AI and education: guidance for policy-makers". // UNESCO. – 2021. URL: https://unesdoc.unesco.org/ark:/48223/pf0000376709

31. Mironova N.V. "Specificity of Civilizational Risks and Mechanisms to Overcome Them". // Manuscript. – 2021. URL: https://cyberleninka.ru/article/n/spetsifika-tsivilizatsionnyh-riskov-i-mehanizmy-ih-preodoleniya/viewer

32. Morozov A., Zgyatti D., Popov P. "Equidistant and Uniform Data Augmentation for 3D Objects". // IEEE Access. – 2021. URL: https://ieeexplore.ieee.org/document/9662385

33. Pavlov A. "Architecture of Computing Systems". // ITMO University. – 2016. URL: https://books.ifmo.ru/file/pdf/2074.pdf

34. Perez S. "Twitter drafts a deepfake policy that would label and warn, but not always remove, manipulated media". // TechCrunch. – 2019. URL: https://techcrunch.com/2019/11/11/twitter-drafts-a-deepfake-policy-that-would-label-and-warn-but-not-remove-manipulated-media/

35. Popov P., Kozlovskii I., Katritch V. "Computational design for thermostabilization of GPCRs". // Current opinion in structural biology. – 2019. URL: https://www.sciencedirect.com/science/article/pii/S0959440X18301374

36. Popov P., Grudinin S., Kurdiuk A., Buslaev P., Redon S. "Controlled-advancement rigid-body optimization of nanosystems". // Journal of computational chemistry. – 2019. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.26016

37. Popov P., Bizin Ilya, Gromiha Michael, Kulandaisamy A., Frishman Dmitrij "Prediction of disease-associated mutations in the transmembrane regions of proteins with known 3D structure". // PloS one. – 2019. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0219452

38. Prajakta P. "AI Deepfakes. The Goose Is Cooked?". // University of Illinois Law Review. – 2020. URL: https://illinoislawreview.org/blog/ai-deepfakes/

39. Sharaev M., Smirnov A., Melnikova-Pitskhe T. "Functional Brain Areas Mapping in Patients with Glioma Based on Resting-State fMRI Data Decomposition". // IEEE. – 2019. URL: https://ieeexplore.ieee.org/abstract/document/8637450

40. Sosnin S., Misin M., Palmer D., Fedorov M. "3D matters! 3D-RISM and 3D convolutional neural network for accurate bioaccumulation prediction". // Journal of Physics: Condensed Matter. – 2018. URL: https://iopscience.iop.org/article/10.1088/1361-648X/aad076

41. Sosnin S., Vashurina M., Withnall M., Karpov P., Fedorov M., Tetko I. "A Survey of Multi-task Learning Methods in Chemoinformatics". // Molecular Informatics. – 2018. URL: https://onlinelibrary.wiley.com/doi/10.1002/minf.201800108

42. Sosnin S., Karlov D., Tetko I., Fedorov M. "Comparative Study of Multitask Toxicity Modeling on a Broad Chemical Space. // J. Chem. Inf. Model. – 2018. URL: https://pubs.acs.org/doi/10.1021/acs.jcim.8b00685

43. Sosnina E.A., Sonin S., Nikitina A., Nazarov I., Osolodkin D., Fedorov M. "Recommender Systems in Antiviral Drug Discovery". // ACS Omega. – 2020. URL: https://pubs.acs.org/doi/10.1021/acsomega.0c00857

44. Tejal A. Patel MD, et al. "Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods". // ACS Journals. – 2016. URL: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.1002/cncr.30245

45. UNESCO "Beijing Consensus on Artificial Intelligence and Education". – 2019. URL: https://unesdoc.unesco.org/ark:/48223/pf0000368303

46. UNESCO "Recommendation on the Ethics of Artificial Intelligence". – 2021. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455

47. United Nations "Transforming Our World: The 2030 Agenda for Sustainable Development". – 2015. URL: https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf

48. Urbina Fabio, Lentzos Filippa, Invernizzi C., Ekins Sean "Dual use of artificial-intelligence-powered drug discovery". // Nature Machine Intelligence. – 2022. URL: https://www.nature.com/articles/s42256-022-00465-9?fbclid=IwAR11_V1cd9SUxEvUfwrWMA7TUcroyYIY1nBDUL3KaS-8B4rG5MIqZCmjm0M&utm_medium=affiliate&utm_source=commission_junction&utm_campaign=CONR_PF018_ECOM_GL_PHSS_ALWYS_PRO-

DUCT&utm_content=productdatafeed&utm_term=PID100032693&CJEVENT=25bc2722a-9b311ec814303660a18050f#Fig1

49. Yang Sh. "Crystal structure of the Frizzled 4 receptor in a ligand-free state". // Nature. – 2018. URL: https://www.nature.com/articles/s41586-018-0447-x

50. Yukhno A.S. "Overview of trends in the development of the metaverse market". // Vestnik of the Institute of Economics No. 6/2022. URL: https://vestnik-ieran.ru/index.php/mh-currentissue/22-stati/ekonomika-i-upravlenie/151-yukhno-a-s-obzor-tendentsij-razvitiya-rynka-metavselennoj

51. Yukhno A. "Digital Transformation: Exploring big data Governance in Public Administration". // Public Organization Review. – 2022. URL: https://link.springer.com/article/10.1007/s11115-022-00694-x

52. Yukhno A.S., Umarov H.S. "Prospects of Development of the Metaverse: Empirical Evidence". // Administrative Consulting. – 2022. URL: https://www.acjournal.ru/jour/article/view/2097

53. Zaretckii M., Bashkirova I., Osipenko S., Kostyukevich Y., Nikolaev E., Popov P. "3D chemical structures allow robust deep learning models for retention time prediction". // Digital Discovery. – 2022. URL: https://pubs.rsc.org/en/content/articlelanding/2022/dd/d2dd00021k

Исследования в цифровой экономике. 2023. 1. P. 61–89
DOI: 10.24833/14511791-2023-1-61-89

Получено: 8 октября 2022 г.,
Пересмотрено 10 ноября,
Принято: 3 декабря 2022 г.

ПРОБЛЕМЫ УПРАВЛЕНИЯ ИСКУССТВЕННОЙ ЛИЧНОСТЬЮ

О.Н. Гуров, А.В. Шерстов

Аннотация. Сегодня большая часть исследователей как в области технических наук, так и представляющих гуманитарное знание, считают необходимым определять искусственный интеллект с учетом приложения к нему субъективных «человеческих» качеств, к которым относятся способность к самосознанию, а также к свободному выбору. В этой связи чрезвычайно актуальной становится проблема автономии ИИ, а далее – прав и возможностей (или отсутствия и отказа от возможности) для создателя ИИ по осуществлению контроля над «продуктом». В рамках исследования этой проблематики уже 20 лет существует и развивается проект «Искусственная личность». Несмотря на то, что в рамках данного проекта ведется активная научная и общественная деятельность, в которую вовлечены ведущие представители научного сообщества, сам проект еще далек от завершения. Представленная статья подводит итог выполненных исследований по концептуализации искусственной личности (ИЛ), материалы публикации демонстрируют, что на сегодняшний день даже принципиальная возможность создания ИЛ еще убедительно не доказана. Также концептуально не сформулирован единый общепринятый подход к перспективным методам и технологиям реализации и воплощения ИЛ. Таким образом, на современном этапе изучение ИЛ представляет собой по большей мере абстрактные теоретические исследования. В результате проведенного анализа авторы статьи приходят к выводу, что на сегодняшний день наиболее целесообразно использовать результаты исследований Естественной Личности и Естественного Интеллекта и переносить методы, доказавшие свою относительную эффективность в разнообразных существующих проявлениях реальной социальной жизни, в область создания концепции ИЛ. Предлагаемый подход к концептуализации ИЛ может быть полезным для создания теоретико-методологического фундамента дальнейших теоретических исследований и проектов реализации ИЛ.

Ключевые слова: Искусственный интеллект, искусственная личность, тест Тьюринга, этика ИИ, самость, идентичность, естественная личность, естественный интеллект

Для цитирования: Гуров О.Н., Шерстов А.В. (2023). Проблемы управления искусственной личностью. – *Исследования в цифровой экономике.* №1, С. 61–89. DOI: 10.24833/14511791-2023-1-61-89

Об авторах:

Гуров Олег Николаевич – кандидат философских наук, MBA. Старший преподаватель Учебно-научного центра гуманитарных и социальных наук Московского физико-технического института, преподаватель Московского государственного института международных отношений Министерства иностранных дел России, Московского государственного университета им. М.В. Ломоносова, РАНХИГС.
140180, МО, г. Жуковский, ул. Гагарина, 16 (FALT). E-mail: gurov-on@ranepa.ru

Шерстов Алексей – магистр экономики, аспирант кафедры философии политики и права философского факультета Московского государственного университета имени М.В. Ломоносова. 119991, Москва, Ленинские горы, Московский государственный университет, "Шуваловский". E-mail: sherstov@miavend.ru

Благодарность:

Авторы выражают свою благодарность доктору философских наук, профессору А.Ю. Алексееву, а также студентам Е. Назаровой, А. Слепышевой, Я. Прорзину и А. Трескуновой за плодотворную дискуссию и за помощь в концептуализации вопросов, поднятых в данном исследовании.

Введение

Быстрое развитие и повсеместное внедрение проектов, основанных на т.н. искусственном интеллекте, заставляет все чаще задумываться не только о том, насколько он ИИ «умный» (то есть насколько быстро и эффективно он способен изучать данные, искать закономерности и создавать связи для решения предложенных задач), но и насколько он самостоятелен, автономен и действительно способен обладать сознанием. При этом современное общество становится все более зависимым от цифровых технологий. Люди вынуждены полагаться на компьютерные системы не только в сферах производства, образования, бизнеса, но часто вынуждены доверять технологиям, в том числе основанным на искусственном интеллекте, решение и управление значимыми и масштабными проектами, а также доверять решение конкретных проблем – вплоть до вопросов жизни и смерти [32]. Поэтому некоторые исследователи уже сегодня стараются наделить искусственный интеллект качествами личности субъекта (актора), среди который можно выделить способность к осознанности и свободе выбора при принятии решений. При этом неизбежно возникает вопрос о границах этой независимости и способности создателя искусственного интеллекта контролировать свое творение. Вопрос того, как далеко может зайти человечество в развитии искусственного интеллекта, и что человечеству необходимо предпринять, чтобы не потерять контроль над технологиями, перестает быть лишь сюжетом научно-фантастических фильмов и переходит в политическую, юридическую и собственно практическую области [13, 10, 37].

В последнее время ученые в области машинного обучения добиваются все более впечатляющих результатов. В частности, изобретение новых архитектур нейронных сетей (особенно BERT, которая появилась в 2018 году) и рост вычислительных мощностей привели сразу к нескольким прорывам в области обработки естественного языка (NLP): результаты искусственного интеллекта значительно улучшились при решении многих классических задачах, от разнообразных разметок до машинного перевода, генерации осмысленных текстов и диалогов [36]. Обученные на сотнях гигабайтах текстов из книг, статей, постов в социальных

сетях, нейронные сети стали способными идеально сохранять контекст, выделять главные мысли и выдавать осмысленные и в то же время интересные и нередко непредсказуемые результаты. Это важно, потому что высокая схожесть с текстом, написанным человеком, вляется приоритетом, а «человеческие» тексты, как утверждают лингвисты, довольно непредсказуемы и нередко неожиданные, в отличие от типичных текстов, генерируемых машиной [27].

Такие результаты не могут не заставить задуматься о перспективах на ближайшее будущее [28]. Вполне возможно, что дальнейшее увеличение вычислительной мощности и создание более продвинутой архитектуры приведут к тому, что нейронная сеть перестанет просто алгоритмически оперировать векторными представлениями лексем. Появится нечто, обладающее подобием эмоций, способное на нечто более близкое к человеческому мышлению [15, 34]. В какой-то момент грань между человеком и такой искусственной личностью сотрется настолько, что сами люди начнут признавать ее за себе подобную. Предпосылки для этого возникают уже в наше время: в качестве примера можно привести инцидент с LaMDA, системой имитации речи ИИ от Google [33]. Эта система настолько совершенна, что один из сотрудников компании поверил в ее разумность и инициировал обсуждение «неэтичной работы компании» [7, 35].

Погружение в проблему «сознания» искусственного интеллекта приводит исследователей к проекту Искусственной личности. Может ли компьютер стать субъектом? Обсуждая тему искусственной личности, большинство ученых и экспертов исходят из следующих предпосылок: искусственный субъект является копией (имитацией) естественного субъекта и создается из естественного субъекта с помощью компьютерных технологий, при этом методы реализации не ограничены.

Цель данной работы - определить потенциальные проблемы управления и контроля искусственной личности, основываясь на текущем уровне знаний и развития технологий, а также на научной дискуссии по данной проблематике.

Проблема определения искусственной личности

С середины 1990-х годов в междисциплинарном научном дискурсе активно развивается проект «Искусственная личность». Одним из ведущих экспертов, вдохновителей и значимых ученых в этой области является известный российский мыслитель, доктор философских наук А.Ю. Алексеев. Этот ученый выделяет несколько устоявшихся определений искусственной личности, основанных на исследовательской позиции относительно «личности» когнитивно-компьютерной системы [1, 39].

Классификация дается в контексте отношения к понятию «философский зомби», которое концептуально близко или даже (по мнению некоторых ученых) идентично понятию Искусственной личности. В широком смысле философский зомби представляет собой систему, не обладающую сознанием, но при

этом с точки зрения поведения, выполняемых функций, по внешнему виду неотличимо или очень похоже на существо, обладающее сознанием [3, 40]. Среди исследователей есть, если можно так выразиться «зомбифилы», которые активно обсуждают данную тему, обладая особым взглядом на само понятие «философского зомби». Они допускают существование зомби или, по крайней мере, возможность их существования. Применительно к искусственному интеллекту эти ученые допускают, что люди смогут создать функционально идентичных двойников, не обладающих сознанием.

«Антизомбисты» считают абсурдной идею о возможности существования зомби как такового.

А третья, нейтрально настроенная, группа ученых, пытается, по крайней мере, внести определенность в эту запутанную проблему. «Нейтралы» считают, что компьютерная модель должна включать модуль «псевдосознания», своего рода аналог человеческого сознания.

Некоторые другие ученые, так называемые «азомбисты», игнорируют сам вопрос о «сознании» искусственных систем. Для них главным вопросом является фактическое наличие «правдоподобной имитации человека», а не наличие характеристик, позволяющих его персонифицировать. По мнению Алексеева, преимущество такого подхода к классификации заключается в конкретизации предметной области, поскольку в этом случае осуществляется анализ не некоего «интеллект», а гораздо более глубоких категорий, в число которых входят самость, личность, «я», самосознание, и ряд других. Далее мы представим несколько основных моделей искусственной личности, разработанных исследователями.

1. Искусственная личность как имитация естественной личности.

Это определение считают справедливыми азомбисты. Концепция заключается в том, что искусственный субъект не отличается от естественного: ни по функциям, ни по поведению, ни даже физически. Важным условием, характерным для этого подхода, является неразличимость естественных и искусственных личностей. Личности обоих типов способны с одинаковым успехом пройти тест Тьюринга [4]. Это означает, что личности похожи не только по структуре и функционированию внешних видимых систем (физической, физиолого-анатомической, вербально-коммуникативной и т.д.), но и внутренней духовной системы, более сложной для воспроизведения, которая отвечает за абстрактные понятия, присущие эмоциональной и чувственной, рефлексивной стороне Человека (например, «любовь", «ответственность», «право» и т.д.) [11].

2. Искусственная личность как модель естественной личности.

Эта концепция подразумевает наличие у искусственной личности компекса «псевдосознания», который используется с целью манипуляции и управления т.н. знаниями и конечно же данными, оперируемыми интеллектуальной системой. По этой же аналогии, человек в состоянии испытывать и осознавать боль, что является результатом работы эффективного сигнального механизма в случае возникновения проблем со здоровьем организма.

3. Искусственная личность как воспроизведение естественной личности.

В соответствии с данным подходом, машина в буквальном смысле воспроизводит все существующие феномены, свойственные человеческому сознанию, посредством реализации сложной функциональной зависимости нейрофизиологических кодов субъективной реальности от субстрата, инвариантного по отношению к физиологической структуре человеческого мозга [9]. Этот подход актуален для антизомбистов. С их точки зрения, зомби — это существо, функционально полностью идентичное человеку, но полностью лишенное сознания. Дубровский определяет важное условие для возникновения искусственной личности – «Я» [9]. Оно включает в себя два важных уровня, которые собственно и формируют личность: генетический и биографический. Однако если генетический уровень можно легко подделать или воспроизвести, то биографический уровень — это опыт, который необходимо получить, осмыслить и сохранить. В связи с этим перед исследователями встает вопрос, будет ли такой робот, опыт которого полностью создан программистом, считаться субъектом, или же он является лишь отражением субъективного мира и менталитета его создателей [16].

4. Искусственная личность - как создание (формирование) «сверхличности».

Искусственная личность, по мнению Денетта, создает такие качества (явления), которые не имеют аналога у естественной личности. Последнее определение приводит к идее «сверхличности», подразумевая, что в результате того, что компьютерные технологии позволяют интегрировать знания и достичь появления «смысла» (в высшем философском смысле), этот уровень и объем смысла превышает обычный уровень «знаний» обычного человека [12].

Следует отметить, что, на наш взгляд, концепции воспроизводства и создания искусственной личности не могут быть использованы для продуктивного исследования перспектив формирования и управления искусственной личностью, а значение этих концепций носит скорее широко философский, чем практически научный характер.

Возможно ли в принципе дать точное и единое определение искусственной личности? На данном этапе это представляется сложной, если не невозможной задачей. Например, если взять за отправную точку определение ее прототипа, но уже на этом, начальном этапе дискуссия наталкивается на почти непреодолимое препятствие, поскольку до сих пор не существует единого, общепризнанного понимания феномена человеческой личности [14]!

Алексеев подчеркивает проблему языковой разобщенности современных проектов Искусственной личности, и причина кроется в принципиальной невозможности создания единого определения “человеческой личности” [2]. Но это даже не единственная проблема. Другой стороной проблемы является разнообразие представлений о способах компьютерной реализации персонологических феноменов. В связи с этим практически невозможно точно ответить на вопрос, реально ли в принципе точное повторение психики человека (не только на формальном, но и на функциональном и семантическом уровнях).

Основные подходы к построению искусственной личности

Ученые выделяют два основных подхода. С точки зрения первого подхода, искусственная личность представляет собой робота, который обладает псевдосознанием. В этом случае он становится обладателем функционального подобия субъективного сознания реальности, присущего человека. Второй подход предполагает, что искусственная личность представляет собой сложную экспертную систему, обладающую механизмами «чувствования» [2, 26].

Первый подход базируется на идее антропопатизма, то есть, по сути, в этом случае Искусственная личность наделяется свойствами человеческой психики. Однако робот, несмотря на полную идентичность в действиях и чувствах, никогда не станет человеком, то есть естественной личностью [17]. Эта идея очень органично соотносится с репродуктивной концепцией и способствует ее развитию. Дубровский не раз обращал внимание на то, что существует важный фактор «Я», который на данном этапе развития не может быть скопирован [9, 26]. Результат формального копирования опыта лишен большой составляющей - чувств и эмоций. Именно они позволяют природной личности, то есть субъекту (человеку), фиксировать память, делать выводы из полученных результатов и, самое главное, оценивать усвоенные уроки и полученный опыт. Оценкой в данном случае являются не цифры и зависимости, а главным образом моральное удовлетворение от результата [24]. Но машина, скорее всего, не способна испытывать чувства, она может только имитировать их. Будет ли такая имитация эквивалентна настоящей человеческой? Чтобы ответить на эти вопросы, нужно четко понимать: что дает и получает человек, что это за «удовлетворение», как его измерить, а главное - по каким параметрам его можно сравнить с «удовлетворением» машины?

Второй подход, в силу того, что он сопряжен с определенными методологическими сложностями, требует возвращения к проблеме определения понятий. Следует отметить, что проблемой моделирования «смысла» занимались многие исследователи. Мы полагаем, что что в качестве наиболее продуктивного зарекомендовал себя «контекстный подход» (по крайней мере в контексте компьютерной реализации). В этих условиях «смысл» является контекстом экспертного «знания», которое формализуется и приобретает параметры системного единства [2]. Алексеев приводит формулу проекта «Искусственная личность» - «смысл + понимание». Но что такое «смысл» для наблюдателя, и что такое «понимание» для самой Машины? Если машина «понимает», то что для нее «смысл»?

В этом контексте вопрос разработки искусственного интеллекта для общения с человеком является довольно впечатляющим, но неоднозначным. Для иллюстрации мы предлагаем пример рбаоты чатбота Tay от Microsoft, который представляет особый интерес [25]. Tay - это чат-бот, который обучается и тренируется на основе сообщений пользователей, с которыми он общается,

и генерирует новые реплики на основе полученных данных [33]. Перед искусственным интеллектом стояла задача выявить формальные коммуникативные и лингвистические паттерны (другими словами, как ребенок должен с нуля выучить грамматику, выучить определенные лексические единицы и т.д.) и применить их. Эксперимент показал, что поначалу Тэй писал безобидные и вполне дружелюбные сообщения собеседникам. Однако через несколько часов чатбот научился писать оскорбления, нацистские высказывания и даже пожелания смерти. Microsoft объяснила такое поведение Tay тем, что на него была совершена «спланированная атака», организованная интернет-троллями. По словам представителей корпорации, злоумышленники воспользовались «уязвимостями» чатбота, что позволило обучить алгоритм отвечать оскорблениями. В свою очередь, компания признала, что недооценила социальный аспект при разработке, и была больше сосредоточена на технической реализации и надежности симуляции общения [29].

Можно ли в этом случае сказать, что чатбот сознательно отвечает с использованием оскорблений и неприемлемых выражений? Очевидно, что нет. Данный продукт искусственного интеллекта не способен отфильтровать негативную информацию и отделить ее от нейтральной или позитивной. На это есть несколько причин: во-первых, создатель не учел особенности интернет-культуры, поэтому позволил своему алгоритму поглощать любую информацию без разбора. Во-вторых, искусственный интеллект еще не настолько умен, чтобы самостоятельно усвоить моральные нормы, позволяющие отделить допустимое от недопустимого.

Здесь заметна важная закономерность - нынешние прототипы искусственной личности не являются автономными. Им по-прежнему нужен модератор (оператор), который вручную цензурирует то, что хорошо и плохо для алгоритма. В то же время человек на инстинктивном уровне способен замечать реакцию других людей и понимать, насколько допустимо то, что он говорит. Машина не обладает такой «социальной смекалкой». По сути, и «хорошо», и «плохо» для нее равнозначны, она не способна придать им эмоциональную окраску, и смысл этих понятий для нее равнозначен [19]. Программист просто закладывает в нее информацию о том, что есть «Класс 1» и «Класс 2», которые эквивалентны для самой Машины. Однако из-за того, что создатель придал «Классу 1» большее значение w1 (веса), которое в первую очередь является числом (!), а «Классу 2» значение меньшее w2, Машина ранжирует объекты первого класса выше, чем объекты второго, и даже может просто полностью исключить их из своей памяти. Проще говоря, она сравнивает w1 с w2 и выдает те результаты, которые имеют больший приоритет.

В приведенном выше примере отсутствует категория «удовлетворения результатом». Не Машина «удовлетворена», а Человек, который создал эту Машину. Вот почему, скорее всего, машины, полностью имитирующие природную личность, никогда не смогут полностью делать то же самое, что делает человек.

Искусственный интеллект — это алгоритм, который имеет определенный результат для любого набора входных данных. Искусственный интеллект — это поиск математических связей между данными, точность измерений, у Машины не может быть ошибки, она всегда знает результат с точностью до наноединиц. Естественный интеллект, с другой стороны, ищет связи не только на основе математики, но и на основе опыта и здравого смысла, поскольку «естественное» не может быть описано математически точно и всегда есть хотя бы крошечная неточность. Математика — это точность, построенная на условностях, на «округлении» естественного до целого. Это естественное ограничение для упрощения понимания. Поэтому личность, созданная математически, упрощается, обобщается до чего-то конкретного, в то время как естественная личность существует по каноническим законам природы.

Обсуждения

Представленный выше обзор подходов и мнений позволяет авторам утверждать, что единственной константой в представленной проблематике является неопределенность. В данном случае, прежде всего, речь идет об отсутствии полной уверенности в том, что в принципе возможно создание искусственной личности. Однако если предположить, что это все-таки выполнимая задача, то мы сталкиваемся с еще одной неопределенностью в определении того, что такое искусственная личность. Если говорить об искусственном интеллекте вообще, по аналогии с естественным интеллектом, то мы сталкиваемся с тем, что нет единого понимания, что такое интеллект вообще, поскольку представители различных областей знания и практики, носители различных культур вкладывают в это понятие зачастую противоречивые характеристики. Перенося эту проблематику на тему искусственной личности, мы в какой-то степени порождаем еще большую неопределенность. Однако это не означает, что мы множим сущности, поскольку в настоящее время человечество стоит перед необходимостью одновременно решать множество проблем - теоретических, методологических, практических, так как ускорение времени делает необходимым увеличение скорости развития во всех областях общественной жизни, независимо от того, насколько масштабны существующие требования, и каких бы сложных систем это не касалось.

Учитывая, что мы уже выяснили, что основной проблемой, с которой сталкиваются при разработке искусственной личности, является неопределенность этого понятия и несовершенство попыток ее реализации, поэтому на данный момент невозможно определить способы управления ею, так как мы просто не знаем определенно, чем мы должны управлять.

1. Искусственная личность как управляемый субъект, обладающий всеми признаками естественной личности.

Во-первых, если человеком можно управлять как интуитивно, так и на основе большой выборки экспериментов и реальных случаев, то наглядных примеров управления искусственной личностью пока нет. Тем не менее, для целей обсуждения можно попытаться «приравнять» искусственную личность к естественной и предположить, что поведение ИИ, как имитация человека, будет в чем-то напоминать поведение человека. Отсюда вытекает первая условность в разработке методологии управления Искусственной личностью, описанная выше. Поиск методов управления искусственной личностью должен постоянно корректироваться: неэффективные должны удаляться, а новые добавляться для последующего тестирования и применения [10].

Кроме того, поиск и верификация эффективных методов — это непрерывный процесс. Постоянно возникают новые тенденции в менеджменте в результате динамических изменений в окружающем нас мире, поэтому поиск новых методов работы в отношении Искусственной личности должен постоянно быть в фокусе внимания, также как и отбор соответствующих методов для актуальных исследований общества, человека и пр.

При разработке первого прототипа методологии стоит вспомнить о частых проблемах, возникающих при управлении людьми (естественными личностями). Проще всего взять примеры из бизнеса. Чаще всего в начале, при разработке стратегии, необходимо учитывать региональные экономические, политические, социальные и, в частности, культурные особенности. Например, для российской среды, в том числе и деловой, характерна значительная роль устных договоренностей, которые сравнительно редко закрепляются какими-либо документами. Однако может ли такая проблема возникнуть для Искусственной личности? На этот вопрос нет единого ответа, поскольку отношение может меняться в зависимости от нашего отношения к искусственной личности. С одной стороны, мы можем рассматривать искусственную личность как клона или дитя своего создателя, тогда его черты переносятся на личность. В результате искусственная личность оказывается «зеркалом» представителя естественной личности, то есть ее точной имитацией. Однако, с другой стороны, можно рассматривать искусственную личность как самостоятельную единицу, которая при этом получила свой собственный уникальный опыт, смогла его осмыслить и эмоционально прожить. В этом случае также возможны несколько исходов: либо черты Искусственной личности будут унифицированы и не будут отличаться у разных индивидов, либо мы получим новые и, возможно, неожиданные черты, не возникающие при взаимодействии с человеком [22].

Уже сегодня можно отметить, что при применении этой модели для управления Искусственной личностью, мы игнорируем возможные несоответствия в понимании условий задач Искусственной личностью. Впоследствии это может привести, во-первых, к неверным результатам, а во-вторых, к «межличностным» конфликтам с машиной. Последний аспект не менее важен, чем первый: предполагая, что Искусственная личность будет действовать идентично на-

блюдателю (менеджеру), последний наделяет ее своими субъективными личностными качествами и пытается определить ее отношение к различным ситуациям. Однако в реальности они могут не соотноситься, поскольку, согласно данному определению Искусственной личности, она является «правдоподобной имитацией человека» и обладает внутренними духовными и эмоционально чувствительными качествами, которые могут существенно отличаться в зависимости от точки зрения наблюдателя. В то же время, если в итоге Искусственная личность все же не обладает внутренней культурой создателя, то мы априори заявляем, что ее собственная культура и понимание (или их отсутствие) коррелируют с нашими.

Это подводит нас к следующему тезису-ограничению в развитии методологии управления Искусственной личностью:

2. Внутренняя культура Искусственной личности и ее возможные интерпретации полностью совпадают с интерпретацией ее создателя и руководителя.

В данном случае условность заключается в том, что для искусственного интеллекта не существует внутренней эмоциональной разницы между классами: один набор данных для него эквивалентен второму и т.д. Набор весов, созданный программистом, становится основой для сортировки и расстановки приоритетов определенных значений. То есть то, что важно или не важно, определяет человек-создатель, а не сама машина. В результате мы получаем следующее:

3. Человек учит Искусственную личность тому, что правильно, а что нет.

В конце концов, выясняется, что Искусственная личность полностью зависит от человека. Когда мы пытаемся управлять ею, мы, по сути, управляем частичной копией самих себя, поскольку многие черты личности искусственной Личности определяются ее управляющим и ее создателем, а не ею самой. То есть мы получаем идеального сотрудника, ошибки которого могут быть обусловлены только личностными особенностями создателя искусственного интеллекта.

Представленные выше проблемы свидетельствуют о том, что философско-методологическое осмысление технологий искусственного интеллекта и связанных с ним проектов требует особого внимания и является перспективным направлением научной деятельности. Осмысление представленных проблем особенно важно, так как ответственность за развитие технологий возрастает уже сегодня. Неопределенность, непонимание основ и отсутствие консенсуса по ключевым аспектам развития технологий искусственного интеллекта могут вызвать угрозы от его использования, и сегодня требуется сформулировать гарантии безопасного развития и использования этих систем. В связи с этим важно своевременно и критически анализировать проблемы искусственной личности, чтобы предотвратить негативные последствия работы искусственного интеллекта, одновременно развивая его на благо общества [20].

Возникающие сегодня фундаментальные возможности технологической реализации проекта «Искусственная личность» уже породили широкий спектр социокультурных и гуманитарных проблем. К этому вопросу приковано внимание многих ученых, поэтому любой формат научных мероприятий, на которых участникам предоставляется возможность высказать свою точку зрения на обсуждаемую проблему, очень важен.

В октябре 2022 года состоялся крупный Конгресс Humanities vs Sciences & the Knowledge Accelerating in Modern World: Parallels and Interaction, организованный МФТИ. Основной целью Конгресса был синтез точных и гуманитарных наук, освещение самых передовых разработок и технологий в области искусственного интеллекта, создание условий для развития среды и Science Art и выявление новых возможностей для открытого сотрудничества между ними [18].

В рамках конгресса состоялась панельная дискуссия «Гуманитарное и техническое в проекте Искусственной Личности». На этом мероприятии обсуждались возможности и перспективы компьютерной реализации (имитации, моделирования, воспроизведения) широкого спектра личностных феноменов, а также проект Искусственной личности как системы искусственного интеллекта, убедительно проходящей комплексный тест Тьюринга для компьютерной реализации широкого спектра гуманитарных феноменов – «Я», семантического, когнитивного, экзистенциального, творческого, коммуникативного, мотивационного, волевого, морального и др. Часть докладов на панельной дискуссии была посвящена истории создания, современному состоянию и перспективам технологической реализации проекта «Искусственная личность». В панельной дискуссии приняли участие программисты, математики, инженеры, а также философы, психологи, юристы, лингвисты, политологи, социологи, культурологи, искусствоведы и представители других наук [6].

Благодаря мультидисциплинарности панельной дискуссии, стало возможным систематическое изучение как общих мировоззренческих, так и методологических проблем Искусственной личности, в соответствии с конкретными информационно-технологическими, правовыми, этическими, эстетическими и другими аспектами наук [7, 21, 38].

Выводы

На сегодняшний день в научном дискурсе нет единого мнения относительно определения Искусственной личности и принципиальной возможности ее создания. Прежде чем двигаться в этом направлении, необходимо решить как методологические, так и междисциплинарные вопросы.

Современный этап развития философии искусственного интеллекта не позволяет однозначно определить возможность и рабочие инструменты для управления Искусственной личностью. Во-первых, мы не до конца знаем, что такое естественная личность, и, скорее всего, никогда не сможем однозначно опреде-

лить, что это такое. В этой связи невозможно дать единое определение Искусственной личности, что является второй проблемой. Кроме того, Искусственная личность может быть практически реализована самыми разными способами, поэтому существует и множество различных интерпретаций. В результате создается ситуация, когда в научном сообществе существует множество интерпретаций как концептуального понимания, так и различных видений путей развития. Таким образом, мы получаем неопределенный объект управления, что не позволяет точно сказать, какие методы будут для него эффективны и продуктивны.

Однако, исходя из имеющегося опыта работы с естественной личностью, можно предложить использование методов, которые показали свою относительную эффективность в самых разных ситуациях социальной жизни. При этом следует учитывать, что нет никакой гарантии, что эти методы в конечном итоге станут применимыми, поскольку, как уже говорилось, Искусственная личность на данном этапе является лишь теоретическим проектом, и мы не располагаем, прежде всего, статистическими данными, которые могли бы позволить нам предсказать поведение Искусственной личности на практике и убедительно спрогнозировать хотя бы приблизительно ожидаемый результат [23].

Подводя итог, хотелось бы предложить к обсуждение гипотезу, что проблема управления может вообще не возникнуть: поскольку Искусственная личность является алгоритмом, она математически ограничена: если представить решения в виде графа, то, несмотря на огромное количество возможных решений, это количество все равно конечное, поскольку каждый набор входных данных будет иметь свой окончательный ответ. Человек способен обойти метод перечисления, используя внутренние ощущения, чтобы игнорировать механический отбор имеющихся данных для получения решения. Чтобы продемонстрировать вышеизложенное предположение на примере, рассмотрим следующую ситуацию. Искусственную личность можно обучить основам музыки (ноты, ритм и т.д.) и заставить ее написать мелодию. Согласно законам комбинаторики, существует бесконечное количество вариаций, и наверняка среди отобранных композиций найдутся точные копии Чайковского, Баха, Вагнера и других. Однако машина «сочинила» это не потому, что обладает внутренней способностью чувствовать гармонию и красоту в музыке, а потому, что может математически создать любую комбинацию и выбрать ее из триллионов созданных. Когда человек сочиняет музыку, внутри него есть нечто, что позволяет ему интуитивно находить удачные решения и объединять их в композицию. Человек может создать один шедевр, и он сразу будет качественным и гениальным, а машина может создать бесконечное количество произведений, но большая часть созданного будет посредственностью и какофонией.

В заключение авторы хотели бы представить общее мнение относительно перспектив развития представленной проблематики. Несомненно, концептуализация искусственной личности представляется перспективным исследованием для дальнейшего понимания искусственного интеллекта в целом [31].

Эта Terra Incognita, в которую человечество сегодня еще почти слепо вступает, требует концептуализации новых идей - технобиологического симбиоза, искусственной жизни, вычислительного творчества, искусственных обществ и многих других [6, 7]. Авторы статьи приходят к выводу о необходимости формирования нового научного языка на пересечении философской антропологии, философии сознания, философии технологии, философии искусственного интеллекта. Возможно, объединяющим фактором и общей темой может стать изучение этических проблем, возникающих в этом новом пространстве. Поэтому перспектива междисциплинарного подхода и проведения междисциплинарных исследований и экспертиз с участием как традиционных специалистов в области техники, так и гуманитарных наук представляется очевидной. Участие в этой дискуссии экспертов из разных областей позволит переформатировать смыслы, ценности и правила под новую реальность, чтобы дальнейшее развитие технологий пошло на пользу человечеству и обеспечило ему благополучное будущее.

Ссылки

1. Алексеев А. Ю. Трудности проекта искусственной личности // Искусственные общества. – 2008 – Т. 3 – Выпуск 1 URL: https://artsoc.jes.su/s207751800000077-3-1/

2. Алексеев А.Ю. Концепция зомби и проблемы сознания // Проблемы сознания в философии и науке / Под ред. проф. Д. И. Дубровский. - М.: Канон+: РООИ «Реабилитация», 2009. - С. 195-214.

3. Алексеев А.Ю. Проекты когнитивных технологий искусственной личности. Человек: образ и сущность. Гуманитарные аспекты, № 1, 2014. С. 156-174.

4. Алексеев А. Ю. , Ефимов А. Р. , Финн В. К. Будущее искусственного интеллекта: тьюринговая или посттьюринговая методология? // Искусственные общества. – 2019 – Т. 14 – Выпуск 4 URL: https://artsoc.jes.su/s207751800007698-6-1/.

5. Белохина Ю. С. , Гуров О. Н. Об этике цифрового двойника общества // Искусственные общества. – 2020 – Т. 15 – Выпуск 4 URL: https://artsoc.jes.su/s207751800012583-0-1/. DOI: 10.18254/S207751800012583-0

6. Глухих В. А., Елисеев С. М., Кирсанова Н. П. Искусственный интеллект как проблема современной социологии // ДИСКУРС. 2022 Т. 8, № 1 С. 82–93. DOI: 10.32603/2412-8562-2022-8-1-82-93.

7. Гуров О. (2020). Этическое взаимодействие с интеллектуальными системами. Искусственные общества. Том. 15, №. 3 DOI: 10.18254/S207751800010905-4

8. Дубровский Д.И. Значение нейронаучных исследований сознания для разработки общего искусственного интеллекта (методологические вопросы) // Вопросы философии. 2022 № 2. С. 83-93.

9. Дубровский Д.: “Кибернетическое бессмертие. Фантастика или научная проблема?” // 2045 URL: http://2045.com/articles/30810.html (дата обращения: 02.11.2022).

10. Минбалеев, А.В. (2022). Концепция "искусственного интеллекта" в юриспруденции. Вестник Удмуртского университета. Серия Экономика и право. 32. С. 1094-1099. 10.35634/2412-9593-2022-32-6-1094-1099.

11. Barrett, Lisa & Adolphs, Ralph & Marsella, Stacy & Martinez, Aleix & Pollak, Seth. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychological Science in the Public Interest. 20. 1-68. 10.1177/1529100619832930.

12. Dennett, Daniel. (1994). Artificial Life as Philosophy. Artificial Life. 1. 291-292. 10.1162/artl.1994.1.291.

13. Dregger, Alexander. (2020). More than Intelligence? The role of artificial personality in the user experience of artificial intelligent agents.

14. Esterwood, Connor & Essenmacher, Kyle & Yang, Han & Robert, Lionel & Zeng, Fanpan. (2021). A Meta-Analysis of Human Personality and Robot Acceptance in Human-Robot Interaction. 10.1145/3411764.3445542.

15. Geher, Glenn & Betancourt, Kian & Jewell, Olivia. (2017). Связь между эмоциональным интеллектом и креативностью. Imagination, Cognition and Personality. 37. 027623661771002. 10.1177/0276236617710029.

16. Guggemos, Josef & Seufert, Sabine & Sonderegger, Stefan. (2020). Pepper: A humanoid robot with personality?

17. Hosseinzadeh Dizaj, Mehran. (2022). The effect of artificial intelligence in robotics.

18. Humanities vs Sciences & the Knowledge Accelerating in Modern World: Parallels and Interaction // MIPT URL: https://humanities-vs-sciences.events/ (Accessed 02.11.2022).

19. Inzlicht, Michael & Bartholow, Bruce & Hirsh, Jacob. (2015). Emotional foundations of cognitive control. Trends in Cognitive Sciences. 19. 10.1016/j.tics.2015.01.004.

20. Jirak, Doreen & Aoki, Motonobu & Yanagi, Takura & Takamatsu, Atsushi & Bouet, Stephane & Yamamura, Tomohiro & Sandini, Giulio & Rea, Francesco. (2022). Is It Me or the Robot? A Critical Evaluation of Human Affective State Recognition in a Cognitive Task. Frontiers in Neurorobotics. 16. 882483. 10.3389/fnbot.2022.882483.

21. Kadri, Faisal. (2014). Understanding and learning to reconcile differences between disciplines through constructing an artificial personality. Kybernetes. 43. 1338-1345. 10.1108/K-07-2014-0152.

22. Kambur, Emine. (2021). Emotional Intelligence or Artificial Intelligence? Emotional Artificial Intelligence. Florya Chronicles of Political Economy. 7. 10.17932/IAU.FCPE.2015.010/fcpe_v07i2004.

23. Kraus, Kateryna & Kraus, Nataliia & Hryhorkiv, Mariia & Kuzmuk, Ihor & Shtepa, Olena. (2022). Artificial Intelligence in Established of Industry 4.0. WSEAS TRANSACTIONS ON BUSINESS AND ECONOMICS. 19. 1884-1900. 10.37394/23207.2022.19.170.

24. Lambert, Alexis & Norouzi, Nahal & Bruder, Gerd & Welch, Gregory. (2020). A Systematic Review of Ten Years of Research on Human Interaction with Social Robots. International Journal of Human-Computer Interaction. 36. 1-14. 10.1080/10447318.2020.1801172.

25. Learning from Tay's introduction // Microsoft URL: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ (accessed 11/2/2022).

26. Lee, Kwan & Peng, Wei & Jin, Seung-A & Yan, Chang. (2006). Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction. Journal of Communication. 56. 754 - 772. 10.1111/j.1460-2466.2006.00318.x.

27. Leichtmann, Benedikt & Nitsch, Verena & Mara, Martina. (2022). Crisis Ahead? Why Human-Robot Interaction User Studies May Have Replicability Problems and Directions for Improvement. Frontiers in Robotics and AI. 9. 838116. 10.3389/frobt.2022.838116.

28. Lennox, John. (2020). 2084: Artificial Intelligence and the Future of Humanity. Perspectives on Science and Christian Faith. 72. 254-255. 10.56315/PSCF12-20Lennox.

29. Lombard, Matthew. (2021). Social Responses to Media Technologies in the 21st Century: The Media are Social Actors Paradigm. 2. 29-55. 10.30658/hmc.2.2.

30. Mileounis, Alexandros & Cuijpers, Raymond & Barakova, Emilia. (2015). Creating Robots with Personality: The Effect of Personality on Social Intelligence. 10.1007/978-3-319-18914-7_13.

31. Mou, Yi & Shi, Changqian & Shen, Tianyu. (2019). A Systematic Review of the Personality of Robot: Mapping Its Conceptualization, Operationalization, Contextualization and Effects. International Journal of Human-Computer Interaction. 36. 1-15. 10.1080/10447318.2019.1663008.

32. Natarajan, Manisha & Gombolay, Matthew. (2020). Effects of Anthropomorphism and Accountability on Trust in Human Robot Interaction. 33-42. 10.1145/3319502.3374839.

33. O'Leary, Daniel. (2022). Massive data language models and conversational artificial intelligence: Emerging issues. Intelligent Systems in Accounting, Finance and Management. 29. 182-198. 10.1002/isaf.1522

34. Pessoa, Luiz. (2017). Do Intelligent Robots Need Emotion?. Trends in Cognitive Sciences. 21. 10.1016/j.tics.2017.06.010.

35. Robert, L. P., Alahmad, R., Esterwood, C., Kim, S., You, S., & Zhang, Q. (2020). A Review of Personality in Human-Robot Interactions. Foundations and Trends in Information Systems, 4(2), 107-212.

36. Sabharwal, Navin & Agrawal, Amit. (2021). Hands-on Question Answering Systems with BERT, Applications in Neural Networks and Natural Language Processing. 10.1007/978-1-4842-6664-9.

37. Shi, Yuwen. (2022). On Negativism of Legal Personality of Artificial Intelligence. Journal of Education, Humanities and Social Sciences. 1. 90-96. 10.54097/ehss.v1i.645.

38. Solaiman, S M. (2017). Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. Artificial Intelligence and Law. 25. 10.1007/s10506-016-9192-3.

39. Spatola, Nicolas & Monceau, Sophie & Ferrand, Ludovic. (2020). Cognitive Impact of Social Robots: How Anthropomorphism Boosts Performances. IEEE Robotics & Automation Magazine. 27. 73-83. 10.1109/MRA.2019.2928823.

40. Tucker, Christopher. (2018). Artificial personality demonstration. 10.13140/RG.2.2.33133.72167.

Journal of Digital Economy Research. 2023. 1. P. 61–89
DOI: 10.24833/14511791-2023-1-61-89

Received: 8 October 2022,
Revised 10 November,
Accepted: 3 December 2022

PROBLEMS OF ARTIFICIAL PERSONALITY (ARTIFICIAL INTELLIGENCE) CONTROL

by Oleg N. Gurov, Alexey V. Sherstov

Abstract. Today, a number of researchers representing both technical knowledge and the humanities believe that it is necessary to endow Artificial Intelligence with subjective "human" qualities, which include the ability to self-aware, as well as to make a free choice. In this regard, the problem of the AI autonomy becomes extremely relevant, and further – AI creator's rights and capabilities (or ineligibility) to hold control over AI. Within this framework the Artificial Personality project has been developing over the past 20 years. Given its active scientific and social activities with the involvement of the remarkable interdisciplinary community, the project is far from complete. The presented article summarizes the executed research for Artificial Personality conceptualization and demonstrates that today the fundamental possibility of the creation of Artificial Personality has not yet been convincingly proven. Also, conceptually, there has not been formulated the single generally accepted approach to promising methods and technology for the implementation and the embodiment of the Artificial Personality. So, at the current stage, the study of the Artificial Personality is rather abstract theoretical research. As a result of the study, the authors come to the conclusion that today it is reasonable to use the results of the Natural Personality and Natural Intelligence studies and transfer the methods that have shown their relative effectiveness in various existing manifestations of real social life to the field of creating the concept of Artificial Personality. The proposed approach for the conceptualization of Artificial Personality will help to create a theoretical and methodological foundation for theoretical research and further implementation of Artificial Personality projects.

Keywords: Artificial Intelligence, Artificial Personality, Turing test, AI ethics, Self, Identity, Natural Personality, Natural Intelligence

For citation: Gurov O.N., Sherstov A.V. (2023). Problems of Artificial Personality (Artificial Intelligence) Control. *Journal of Digital Economy Research,* vol. 1, no 1, pp. 61–89. (in English).
DOI: 10.24833/14511791-2023-1-61-89

Information about the authors:

Gurov N. Oleg – PhD in Philosophy, MBA. Senior lecturer at the Educational and Scientific Center for the Humanities and Social Sciences of the Moscow Institute of Physics and Technology, lecturer at the Moscow State Institute of International Relations of the Ministry of Foreign Affairs of Russia, Moscow State University. M.V. Lomonosov, RANEPA.
140180, Moscow region, Zhukovsky, st. Gagarina, 16 (FALT)
gurov-on@ranepa.ru

Sherstov V. Alexey – Master of Economics, PhD student Department of Philosophy of Politics and Law, Faculty of Philosophy, Lomonosov Moscow State University.
119991, Moscow, Leninskie gory, Moscow State University, "Shuvalovsky".
sherstov@miavend.ru

Acknowledgements:

The authors express their gratitude to Doctor of Philosophy, Professor A.Yu. Alekseev, as well as to students E. Nazarova, A. Slepysheva, Ya. Prourzin and A. Treskunova for a fruitful discussion and for help in conceptualizing the issues raised in this study.

Introduction

The rapid development and widespread use of Artificial intelligence projects makes us increasingly think not only about how "intelligent" it is (that is, how quickly and efficiently it is able to study, look for patterns and create connections to solve the proposed problems), but also how much it is independent and really "conscious". With this the modern society is becoming increasingly dependent on digital technologies. Humans are forced to rely on computer systems not only in the fields of production, education, business, but also often have to trust technologies, including those based on Artificial Intelligence, to solve and manage significant and large-scale problems - up to life and death issues [32]. Therefore, some researchers endow Artificial Intelligence with the qualities of a subject's (actor's) identity, including the ability for awareness and freedom of choice in decision-making. This inevitably raises the question of the limits of this independence and the ability of the creator of Artificial Intelligence to control his creation. The question of how far the Humanity can go in the development of Artificial Intelligence, and what we need to avoid losing control, ceases to be only the plot of science fiction films and moves into political, legal and actually practical areas [8, 30, 37].

Recently, machine learning scientists have been achieving more and more impressive results. In particular, the invention of new architectures of neural networks (especially BERT, which was released in 2018) and the growth of computing power have led to several breakthroughs in the field of natural language processing (NLP) at once: Artificial Intelligence results have significantly improved on many classical tasks, from various markups to machine translation, generation of meaningful texts and dialogues [36]. Trained on hundreds of gigabytes of texts from books, articles, posts on social networks, neural networks have become able to perfectly preserve the context, highlight the main thoughts and produce meaningful and much the same time interesting and unpredictable results. This is important because high similarity to human text is a priority, and human texts tend to be rather "unpredictable" in terms of perplexity, as opposed to typical machine-generated text, as linguists claim [26].

Such results cannot but make one think about the prospects for the near future [27]. It is quite possible that a further increase in computing power and more and more new architectures will lead to the fact that the neural network will no longer simply operate algorithmically with vector representations of lexemes. Something will appear that has a semblance of emotions, capable of something closer to human thinking [12, 34]. At some point, the line between Human and a similar Artificial Personality will blur so much that Humans themselves will start to recognize it of their own kind. The prerequisites for this are already happening in our time: as an example, we can take the incident with LaMDA, Google's AI speech imitation system [33]. This system is so advanced that one of the company's employees came to believe in its reasonableness and initiated the discussion of the "unethical performance of the company" [15, 35].

Dive into the problem of "consciousness" of Artificial Intelligence brings researchers to the Artificial Personality (or Artificial Intelligence – further we will be referring to this concept as to Artificial Personality) project. Can a computer become a subject? When discussing the topic of Artificial Personality, most scientists and experts proceed from the following premises: an artificial subject is a copy (imitation) of a natural subject and is created from a natural subject using computer technology, while the implementation methods are not limited.

The target of this work is to identify potential management and control problems of the Artificial Personality, based on the current level of knowledge and technology development, as well as on the scientific discussion on this issue.

The problem of defining an Artificial Personality

Since the mid-1990s, the Artificial Personality project has been widely discussed in the interdisciplinary scientific discourse. One of the leading experts, inspirers and significant scientists in this field is the famous Russian thinker, Doctor of Philosophy, A. Yu. Alekseev. This scientist identifies several established definitions of Artificial Personality, based on research position regarding the "personality" of cognitive-computer system [1, 39].

The classification is given in the context of the attitude to the concept of "Philosophical zombie", which is conceptually close or even identical to the concept of Artificial Personality. In general, philosophical zombies are "unconscious systems that are behaviorally, functionally and/or physically identical, indistinguishable and/or similar to conscious beings" [3, 40]. Among the researchers, there are conventional "zombiphiles" who are ready to discuss this theory, having their own view on the very concept of "philosophical zombie". They admit the existence of zombies, or at least the possibility of imagining them. That is, they assume that Humans will be able to create their functionally identical counterparts that do not have consciousness.

"Anti-zombists" consider even the idea of the possibility of zombies' existence as absurd.

And the third, neutral, group is trying, at least, to bring certainty into this confusing problem. "Neutrals" believe that a computer model should include a "pseudo-consciousness" module, a kind of analogue of human consciousness.

Some other researchers, the so-called "azombists", ignore the very issue of "consciousness" of Artificial systems. The main issue for them is the actual presence of a "believable imitation of a Human", and not the presence of characteristics that allow to personify it. According to Alekseev, the advantage of this approach to classification lies in the concretization of the sub'ject area: it is not "intelligence" that is analyzed, but more complex concepts – "consciousness", "personality", "self-consciousness", "Self". Next, we present a few of the main Artificial Personality models developed by researchers.

1. *Artificial Personality as the imitation of Natural Personality.*

This definition is typical for azombists. The concept is that the Artificial Subject is not different from the natural one: neither in function, nor in behavior, or even physically. An important condition that is characteristic of this approach is the indistinguishability of natural and artificial personalities. Both individuals are capable of passing Turing test with equal success [4]. This means that personalities are similar not only in the structure and functioning of external visible systems (physical, physiological-anatomical, verbal-communicative, etc.), but also in the inner spiritual system, which is more complex to reproduce, and which is responsible for abstract concepts inherent in emotional and the sensual, reflective side of the Human (i.e. "love", "responsibility", "the right", etc.) [5].

2. *Artificial Personality as the model of a Natural Personality.*

This concept implies the presence of "pseudo-conscious" complex for the Artificial Personality, which is used to effectively manipulate the "data" and "knowledge" of the intellectual system. By the same analogy, Human may experience and be aware of pain, which is the work of the effective signaling mechanism in case there are problems with the health of body.

3. *Artificial Personality as the reproduction of Natural Personality.*

According to this definition, computer system actually reproduces general and individual phenomena of consciousness through the implementation of a complex functional dependence of neurophysiological codes of subjective reality on a substrate that is invariant with respect to the physiological structure of the human brain. This approach is relevant for anti-zombists. From their point of view, zombie is a creature that is functionally completely identical to Human, but completely devoid of consciousness. Dubrovsky defines an important condition for the emergence of Artificial Personality – the "Self" [10]. It includes two important levels that actually shape the personality: genetic and biographical. However, if the genetic level can be easily faked or reproduced, the biographical level is the experience that must be obtained, comprehended and preserved. In this regard, researchers are faced with the question of whether such a robot, whose experience is completely created by programmer, will be considered a subject, or is it just a reflection of the subjective world and the mindset of its creators [14].

4. Artificial Personality – as the creation (formation) of "superpersonality".

According to D. Denette the Artificial Personality creates such qualities (phenomena) that have no analogue with the natural one. The latter definition leads to the idea of "superpersonality", implying that as a result of the fact that computer technologies allow the integration of knowledge and achieve the appearance of "meaning" (in the highest philosophical meaning), this level and scope of meaning exceeds the usual level of "knowledge" of ordinary Humans [7].

It is worth noting that, in our opinion, the reproduction and creation concepts of the Artificial Personality cannot be used for productive research on the prospects for the formation and management of the Artificial Personality, and the significance of these concepts is more broadly philosophical than practically scientific.

Is it possible in principle to give an accurate and unified definition of an Artificial Personality? This seems to be a difficult, if not impossible task at this stage. For example, if we take the definition of its prototype as a starting point, but already at this, initial step, the discussion encounters an almost insurmountable obstacle, since there is still no single, universally recognized understanding of the phenomenon of the Human Personality [11]!

Alekseev emphasizes the "linguistic disunity" of modern projects of the Artificial Personality, and the reason lies in the fundamental impossibility of creating the unified definition of "Human Personality" [2]. But this is not the only issue. The other side of the problem is the variety of ideas about the ways of computer realization of personological phenomena. In this regard, it is practically impossible to accurately answer the question whether, in principle, an exact repetition of the human psyche is real (not only at the formal, but also at the functional and semantic levels).

Basic line of approach to architect Artificial Personality

Scholars give accents to two major directions. In one respect, Artificial Personality may be determined as as a robot endowed with pseudo-consciousness, which makes it functional similarity of human subjective consciousness of reality, and on the other hand, Artificial Personality can be perceived as an expert system with "sense" mechanisms [2, 25].

The first approach involves appeal to the idea of anthropopathism, i.e., in fact, Artificial Personality is endowed with the properties of the human psyche. However, the robot, despite the complete identity in actions and feelings, will never become Human, that is, a Natural Personality [16]. The idea is very organically correlated with the reproductive concept and helps to develop it. Dubrovsky more than once drew attention to the fact that there is an important factor of "Self", which at this stage of development cannot be copied [9, 25]. The result of formal copying of experience is devoid of a large component - feelings and emotions. It is they that allow a Natural Personality, that is, a Subject (Human), to fix memory, draw conclusions from the results and, most importantly, evaluate lessons learned and the experience gained. The assessment in this

case is not numbers and dependencies, but mostly moral satisfaction with the outcome [23]. But the machine is most likely not capable of experiencing feelings, it can only imitate them. Will such an imitation be equivalent to a real human one? To answer these questions, one needs to clearly understand: what does Human give and get, what is that "satisfaction", how to measure it, and most importantly - by what parameters can it be compared with the "satisfaction" of Machine?

The second approach, considering that it involves certain methodological ambiguities, requires return to the topic of defining concepts. It is worth mentioning that many researchers have addressed the problem of modeling "sense". The most productive from the point of view of computer implementation, is the "contextual approach", according to which "meaning" is the context of formalized expert "knowledge" that has parameters of system unity [2]. Alekseev presents the formula of the Artificial Personality project – "meaning + understanding". But what is the "meaning" for the observer, and what is the "understanding" for the Machine itself? If the machine "understands", then what is the "meaning" for it?

In this context, the issue of developing Artificial Intelligence for communication with a Human is rather impressive but ambiguous. To illustrate, we offer the case of Tay Chatbot from Microsoft that is of particular interest [24]. Tay is a chatbot that educates and trains from the messages of the users it talks to and generates new replicas based on the data it receives [33]. Artificial Intelligence was faced with the task of identifying formal communicative and linguistic patterns (in other words, like a child has to learn grammar from scratch, learn certain lexical units, etc.) and apply them. The experiment showed that at first Tay wrote harmless and quite friendly messages to the interlocutors. However, after a few hours, the chatbot learned to write insults, Nazi statements, and even death wishes. Microsoft explained this behavior of Tay by the fact that a "planned attack" organized by Internet trolls was carried out on it. According to the corporation, the attackers took advantage of the "vulnerabilities" of the chatbot, which made it possible to train the algorithm to respond with insults. In turn, the company admitted that it had underestimated the social aspect in the development, and was more focused on the technical implementation and the reliability of the simulation of communication [28].

Is it possible in this case to say that the chatbot consciously responds with cruel expressions? Obviously not. This instance of Artificial Intelligence is not able to filter negative information from the positive one. There are several reasons for this: firstly, the creator did not take into account the peculiarities of Internet culture, which is why he allowed his algorithm to absorb any information indiscriminately. Secondly, Artificial Intelligence is not yet so intelligent as to learn the moral norms itself, which allow you to separate the permissible from the impermissible.

An important pattern is noticeable here - the current prototypes of the Artificial Personality are not independent. They still need a moderator (operator) who manually censors what is good and bad for the algorithm. At the same time, Human, on an instinctive level, is able to notice the reaction of other people and understand how per-

missible what he says is. The Machine has no such "social savvy". In fact, both "good" and "bad" are equivalent for it, it is not able to give them an emotional coloring and the meaning of these concepts is equivalent for the Machine [18]. The programmer simply puts into it the information that there are "Class 1" and "Class 2", which are equivalent for the Machine itself. However, due to the fact that the creator gave "Class 1" a higher value of w (weight), which is primarily a number (!), and a value less than w1 to "Class 2", the Machine ranks objects of the first class higher than objects of the second one, and even can simply completely exclude them from its memory. Simply put, it compares w1 with w2 and returns those results that have higher precedence (which makes the value of w).

In the above example, there is nothing about "satisfaction with the result". It is not the Machine which is "satisfied", but the Human who created this Machine. That is why, most likely, machines, with a complete imitation of a Natural Personality, will never be able to fully do exactly the same as a Human does.

Artificial intelligence is an algorithm that has a certain result for any set of input data. Artificial intelligence is about finding mathematical relationships between data, about the accuracy of measurements, the Machine does not have an error, it always knows the result with an accuracy of nano-units. Natural intelligence, on the other hand, seeks connections not only on the basis of mathematics, but also based on experience and common sense, since "natural" cannot be described mathematically accurately and there is always at least a tiny error. Mathematics is precision built on convention, on the "rounding" of the natural to the whole. This is a natural limitation to simplify understanding. Therefore, a personality created mathematically is simplified, generalized to something concrete, while Natural Personality exists according to the canonical laws of nature.

Discussions

The review of approaches and opinions presented above allows the authors to assert that the only constant in the presented problematics is uncertainty. In this case, first of all, we are talking about the lack of complete confidence that, in principle, it is possible to create an Artificial Personality. However, if we assume that this is still a feasible task, then we are faced with another uncertainty in the definition of what an Artificial Personality is. If we are talking about Artificial Intelligence in general, by analogy with Natural Intelligence, then we are faced with the fact that there is no common understanding of what Intelligence in general is, because representatives of various fields of knowledge and practice, carriers of different cultures put often contradictory characteristics into this concept. Transferring this problematic to the topic of the Artificial Personality, we to some extent produce even more uncertainty. However, this does not mean that we are multiplying entities, since at present Humanity is faced

with the need to simultaneously solve many problems - theoretical, methodological, practical ones, since the acceleration of time makes it necessary to increase the speed of development in all areas of social life, regardless of how large-scale the existing requirements are, and whatever complex systems it concerns.

Given that we have already found out that the main problem that we face when we develop on Artificial Personality is its uncertainty and the imperfection of attempts to implement it, therefore, at the moment it is impossible to identify the ways to operate it, because we simply do not definitely know what we have to manage.

1. Artificial Personality as a controlled subject which has all the features of the Natural Personality.

Firstly, if Human can be controlled both intuitively and based on the large sample of experiments and real cases, then there are no demonstrative cases of controlling Artificial Personality yet. Nevertheless, we can try to "equate" the Artificial Personality with the natural one for the purposes of discussion, and assume that the behavior of the AI, as an imitation of Human, will somehow resemble the behavior of Human. From this follows the first convention in the development of management methodology for an Artificial Personality, described above. The search for methods of managing Artificial Personality must be constantly adjusted: inefficient ones should be removed and new ones added for subsequent testing and application [29].

Further, identifying effective methods is an ongoing process. New trends in management are rapidly changing based on the constant changes in the world around us, so the search for working methods in relation to the Artificial Personality should be constantly considered, as well as screening for relevant methods for people.

When developing the first prototype of the methodology, it is worth remembering the frequent problems that arise in the people management (of Natural Personalities). It is easiest to take examples from business. Most often, at the beginning, when developing a strategy, it is necessary to take into account regional economic, political, social and, in particular, cultural characteristics. For example, Russian environment, including the business one, is characterized by the significant role of verbal agreements, which are relatively rarely secured by any documents. However, can such a problem arise for Artificial Personality? There is no single answer to this question, as it may change depending on our attitude towards the Artificial Personality. On the one hand, we can consider the Artificial Personality as a clone-child of its creator, then its features are transferred to the personality. As a result, the Artificial Personality turns out to be a "mirror" of a representative of a Natural Personality, that is, its detailed imitation. However, on the other hand, we can consider the Artificial Personality as an independent unit, which nevertheless received its own unique experience, was able to think it over and emotionally live it. In this case, several outcomes are also possible: either the features of the Artificial Personality will be unified and will not differ from one individual to another, or we will get new and possibly unexpected features that do not arise when interacting with Human [21].

It can already be noted that when applying this model to control Artificial Personality, we ignore possible inconsistencies in understanding the conditions of the task by Artificial Personality. Subsequently, this can lead, first of all, to incorrect results, and secondly, to "interpersonal" conflicts with the Machine. The last aspect is no less important than the first one: assuming that Artificial Personality will act identically to the Observer (Manager), the latter endows it with its own subjective personal qualities and tries to determine its attitude to various situations. However, in reality, they may not correlate, since, according to this definition of Artificial Personality, it is a "believable imitation of Human" and has internal spiritual and emotionally sensitive qualities that may differ significantly from the observer's point of view. At the same time, if in the end the Artificial Personality still does not have the internal culture of the creator, then we a priori declare that its own culture and understanding (or lack thereof) correlate with ours.

This leads us to the next thesis-limitation in the development of the Artificial Personality management methodology:

2. The internal culture of the Artificial Personality and its possible interpretations completely coincide with the interpretation of its manager.

The last convention is that for Artificial Intelligence there is no internal emotional difference between classes: one set of data for it is equivalent to the second one, etc. The set of weights created by the programmer becomes the basis for sorting and prioritizing certain values. That is, what is important or not important is determined by the Human Creator, and not by the Machine itself. As a result, we get that:

3. Human teaches Artificial Personality what is right and what is wrong.

Eventually, it turns out that Artificial Personality is completely dependent on Human. When we try to control her, we are essentially controlling a partial copy of ourselves, since many of the personality traits of the Artificial Personality are determined by its Manager and its Creator, and not by itself. That is, we get an ideal employee whose misfires can only be due to the personality traits of the Creator of Artificial Intelligence.

The problems presented above indicate that the philosophical and methodological understanding of Artificial Intelligence technologies and projects related to it requires special attention and is a promising area of scientific activity. It is especially important to comprehend the presented problems, as the responsibility for the development of technologies is already growing today. Uncertainty, misunderstanding of the fundamentals and lack of consensus on key aspects of the development of Artificial Intelligence technologies can cause threats from its use, and today it is required to formulate guarantees for the safe development and use of these systems. In this regard, it is important to timely and critically analyze the problems of Artificial Personality in order to prevent the negative consequences of Artificial Intelligence performance, at the same time, developing it for the benefit of society [19].

The fundamental possibilities of the technological implementation of the Artificial Personality project that have arisen today give rise to a wide range of sociocultural and humanitarian problems. The attention of many scientists is riveted to this issue, and therefore any format of scientific events at which participants are given the opportunity to express their point of view on the problem under discussion is very important.

In October 2022, major Humanities vs Sciences Congress was held, organized by the Moscow Institute of Physics and Technology. The main goal of the Congress was the synthesis of exact sciences and the humanities, highlighting the most advanced developments and technologies in the field of Artificial Intelligence, creating conditions for the development of the environment and Science Art and identifying new opportunities for open cooperation between them [17].

Within the framework of the Congress, the Panel Discussion "Humanitarian and technical in the project of Artificial Personality" was held. At this event, the possibilities and prospects of computer implementation (imitation, modeling, reproduction) of a wide range of personal phenomena were discussed, as well as the project of Artificial Personality as an Artificial Intelligence system that convincingly passes the Turing complex test for computer implementation of a wide range of Humanities phenomena – "Self", semantic, cognitive, existential, creative, communicative, motivational, strong-willed, moral, etc. Part of the presentations at the panel discussion was devoted to the history of creation, the current state and prospects for the technological implementation of the Artificial Personality project. The panel discussion was attended by programmers, mathematicians, engineers, as well as philosophers, psychologists, lawyers, linguists, political scientists, sociologists, culturologists, art historians and representatives of other sciences [13].

Thanks to the multidisciplinarity of the panel discussion, it became possible to systematically study both the general worldview and methodological problems of Artificial Personality, according to specific information technology, legal, ethical, aesthetic and other aspects of sciences [15, 20, 38].

Conclusions

At this point, in scientific discourse there is still no ground for agreement considering definition of Artificial Personality. Moreover, there is no consensus whether it is practically possible to create it in principle. At the current stage of scientific discussion, it is necessary to find a solution to many theoretical and methodological issues that lie in the interdisciplinary field.

To date, the situation in philosophical studies, which subject is Artificial Intelligence does not allow us to unambiguously determine the possibility and working tools for managing Artificial Personality. First, we do not fully know what Natural Personality is, and most likely we will never be able to unambiguously define what it is. From here it is impossible to give a single designation of Artificial Personality, which is the second problem. In addition, Artificial Personality can be practically implemented in a

variety of ways, which is why there are also many different interpretations. As a result, a situation is created when in the scientific community there are many interpretations of both conceptual understanding and different visions of development paths. Thus, we get an indefinite control object, which makes it impossible to say exactly which methods will be effective and productive for it.

However, based on the existing experience with Natural Personality, it is possible to suggest the use of methods that have shown their relative effectiveness in the most diverse situations of social life. At the same time, it should be taken into account that there is no guarantee that these methods will eventually become applicable, since, as mentioned above, Artificial Personality is only a theoretical project at this stage, and we do not have, first of all, statistical data that could allow us to predict the behavior of Artificial Personality in practice, and convincingly predict at least any approximate expected result [22].

To sum up, we would like to propose for discussion the hypothesis that the control problem may not arise at all: since Artificial Personality is an algorithm, it is mathematically limited: if we represent solutions in the form of a graph, then despite the huge number of possible solutions, this number is still finite, since each set of input data will have its own final answer. Human is capable to bypass the enumeration method, using internal sensations to ignore the mechanical selection of existing data to obtain a solution. To demonstrate the above assumption by an example, consider the following situation. You can teach Artificial Personality the basics of music (notes, rhythm, etc.) and make it write a melody. According to the laws of combinatorics, there is an infinite number of variations, and for sure among the selected compositions there will be exact copies of Tchaikovsky, Bach, Wagner and others. However, the machine "composed" this not because it has the inner ability to feel the harmony and compatibility in music, but because it can mathematically create any combination and select it from trillions of created ones. When a Human composes music, there is something inside him that allows him to intuitively find successful solutions and combine them into a composition. Human can create a single masterpiece, and at once it will be qualitative and brilliant, and Machine can create an infinite amount, but most of what is created will be mediocrity and cacophony.

In conclusion, the authors would like to present their general opinion regarding the prospects for the development of the presented problems. Undoubtedly, the conceptualization of Artificial Personality seems to be a promising study for further understanding of Artificial Intelligence in general [31]. This Terra Incognita, into which Mankind is still quite blindly entering today, requires the conceptualization of new ideas - technobiological symbiosis, Artificial life, computational creativity, Artificial Societies and many others [6, 7]. The authors of the article come to the conclusion that it is necessary to form a new scientific language at the intersection of philosophical anthropology, philosophy of consciousness, philosophy of technology, philosophy of Artificial Intelligence. Perhaps a unifying factor and a common theme may be the study of ethical issues that arise in this new space. Therefore, the prospect of interdisci-

plinary approach and the conduct of interdisciplinary research and expertise with the participation of both traditional specialists in the field of technology and the humanities seems obvious. Participation in this discussion of experts from various fields will allow reformatting meanings, values and rules for the new reality, so that the further development of technology benefits humanity and ensures a prosperous future.

References

1. Alekseev A. Difficulties of Artificial Intelligence Project // Artificial societies. – 2008. – V. 3. – issue 1. URL: https://artsoc.jes.su/s207751800000077-3-1/

2. Alekseev A. Yu. Cognitive technology projects of Artificial Personality. Human: Image and essence. Humanitarian Aspects, no. 1, 2014, pp. 156–174.

3. Alekseev A. Yu. The concept of zombies and problems of consciousness // Problems of consciousness in philosophy and science / Ed. prof. D. I. Dubrovsky. - M.: Kanon +: ROOI "Rehabilitation", 2009. - S. 195–214.

4. Alekseev A., Efimov A., Finn V. The Future of Artificial Intelligence: Turing or Post-Turing Methodology? // Artificial societies. – 2019. – V. 14. – Issue 4. URL: https://artsoc.jes.su/s207751800007698-6-1/ DOI: 10.18254/S207751800007698-6

5. Barrett, Lisa & Adolphs, Ralph & Marsella, Stacy & Martinez, Aleix & Pollak, Seth. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychological Science in the Public Interest. 20. 1-68. 10.1177/1529100619832930.

6. Belokhina Y., Gurov O. (2020). Ethics of the Digital Twin of Society. Artificial societies. vol. 15, no. 4 DOI: 10.18254/S207751800012583-0

7. Dennett, Daniel. (1994). Artificial Life as Philosophy. Artificial Life. 1. 291-292. 10.1162/artl.1994.1.291.

8. Dregger, Alexander. (2020). More than Intelligence? The role of artificial personality in the user experience of artificial intelligent agents.

9. Dubrovsky D. The value of neuroscience research of consciousness for the development of general artificial intelligence (methodological issues) // Problems of Philosophy. 2022. V. No. 2. S. 83–93.

10. Dubrovsky D.: "Cybernetic immortality. Fantasy or scientific problem?" // 2045 URL: http://2045.com/articles/30810.html (accessed 02.11.2022).

11. Esterwood, Connor & Essenmacher, Kyle & Yang, Han & Robert, Lionel & Zeng, Fanpan. (2021). A Meta-Analysis of Human Personality and Robot Acceptance in Human-Robot Interaction. 10.1145/3411764.3445542.

12. Geher, Glenn & Betancourt, Kian & Jewell, Olivia. (2017). The Link between Emotional Intelligence and Creativity. Imagination, Cognition and Personality. 37. 027623661771002. 10.1177/0276236617710029.

13. Glukhikh, V. & Eliseev, S. & Kirsanova, N.. (2022). Artificial Intelligence as a Problem of Modern Sociology. Discourse. 8. 82-93. 10.32603/2412-8562-2022-8-1-82-93.

14. Guggemos, Josef & Seufert, Sabine & Sonderegger, Stefan. (2020). Pepper: A humanoid robot with personality?

15. Gurov O. (2020). Ethical Interaction with Intellectual Systems. Artificial societies. vol. 15, no. 3 DOI: 10.18254/S207751800010905-4

16. Hosseinzadeh Dizaj, Mehran. (2022). The effect of artificial intelligence in robotics.

17. Humanities vs Sciences & the Knowledge Accelerating in Modern World: Parallels and Interaction // MIPT URL: https://humanities-vs-sciences.events/ (Accessed 02.11.2022).

18. Inzlicht, Michael & Bartholow, Bruce & Hirsh, Jacob. (2015). Emotional foundations of cognitive control. Trends in Cognitive Sciences. 19. 10.1016/j.tics.2015.01.004.

19. Jirak, Doreen & Aoki, Motonobu & Yanagi, Takura & Takamatsu, Atsushi & Bouet, Stephane & Yamamura, Tomohiro & Sandini, Giulio & Rea, Francesco. (2022). Is It Me or the Robot? A Critical Evaluation of Human Affective State Recognition in a Cognitive Task. Frontiers in Neurorobotics. 16. 882483. 10.3389/fnbot.2022.882483.

20. Kadri, Faisal. (2014). Understanding and learning to reconcile differences between disciplines through constructing an artificial personality. Kybernetes. 43. 1338-1345. 10.1108/K-07-2014-0152.

21. Kambur, Emine. (2021). Emotional Intelligence or Artificial Intelligence? Emotional Artificial Intelligence. Florya Chronicles of Political Economy. 7. 10.17932/IAU.FCPE.2015.010/fcpe_v07i2004.

22. Kraus, Kateryna & Kraus, Nataliia & Hryhorkiv, Mariia & Kuzmuk, Ihor & Shtepa, Olena. (2022). Artificial Intelligence in Established of Industry 4.0. WSEAS TRANSACTIONS ON BUSINESS AND ECONOMICS. 19. 1884-1900. 10.37394/23207.2022.19.170.

23. Lambert, Alexis & Norouzi, Nahal & Bruder, Gerd & Welch, Gregory. (2020). A Systematic Review of Ten Years of Research on Human Interaction with Social Robots. International Journal of Human-Computer Interaction. 36. 1-14. 10.1080/10447318.2020.1801172.

24. Learning from Tay's introduction // Microsoft URL: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ (accessed 11/2/2022).

25. Lee, Kwan & Peng, Wei & Jin, Seung-A & Yan, Chang. (2006). Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction. Journal of Communication. 56. 754 - 772. 10.1111/j.1460-2466.2006.00318.x.

26. Leichtmann, Benedikt & Nitsch, Verena & Mara, Martina. (2022). Crisis Ahead? Why Human-Robot Interaction User Studies May Have Replicability Problems and Directions for Improvement. Frontiers in Robotics and AI. 9. 838116. 10.3389/frobt.2022.838116.

27. Lennox, John. (2020). 2084: Artificial Intelligence and the Future of Humanity. Perspectives on Science and Christian Faith. 72. 254-255. 10.56315/PSCF12-20Lennox.

28. Lombard, Matthew. (2021). Social Responses to Media Technologies in the 21st Century: The Media are Social Actors Paradigm. 2. 29-55. 10.30658/hmc.2.2.

29. Mileounis, Alexandros & Cuijpers, Raymond & Barakova, Emilia. (2015). Creating Robots with Personality: The Effect of Personality on Social Intelligence. 10.1007/978-3-319-18914-7_13.

30. Minbaleev, A.V.. (2022). The Concept Of "Artificial Intelligence" In Law. Bulletin of Udmurt University. Series Economics and Law. 32. 1094-1099. 10.35634/2412-9593-2022-32-6-1094-1099.

31. Mou, Yi & Shi, Changqian & Shen, Tianyu. (2019). A Systematic Review of the Personality of Robot: Mapping Its Conceptualization, Operationalization, Contextualization and Effects. International Journal of Human-Computer Interaction. 36. 1-15. 10.1080/10447318.2019.1663008.

32. Natarajan, Manisha & Gombolay, Matthew. (2020). Effects of Anthropomorphism and Accountability on Trust in Human Robot Interaction. 33-42. 10.1145/3319502.3374839.

33. O'Leary, Daniel. (2022). Massive data language models and conversational artificial intelligence: Emerging issues. Intelligent Systems in Accounting, Finance and Management. 29. 182-198. 10.1002/isaf.1522

34. Pessoa, Luiz. (2017). Do Intelligent Robots Need Emotion?. Trends in Cognitive Sciences. 21. 10.1016/j.tics.2017.06.010.

35. Robert, L. P., Alahmad, R., Esterwood, C., Kim, S., You, S., & Zhang, Q. (2020). A Review of Personality in Human-Robot Interactions. Foundations and Trends in Information Systems, 4(2), 107-212.

36. Sabharwal, Navin & Agrawal, Amit. (2021). Hands-on Question Answering Systems with BERT, Applications in Neural Networks and Natural Language Processing. 10.1007/978-1-4842-6664-9.

37. Shi, Yuwen. (2022). On Negativism of Legal Personality of Artificial Intelligence. Journal of Education, Humanities and Social Sciences. 1. 90-96. 10.54097/ehss.v1i.645.

38. Solaiman, S M. (2017). Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. Artificial Intelligence and Law. 25. 10.1007/s10506-016-9192-3.

39. Spatola, Nicolas & Monceau, Sophie & Ferrand, Ludovic. (2020). Cognitive Impact of Social Robots: How Anthropomorphism Boosts Performances. IEEE Robotics & Automation Magazine. 27. 73-83. 10.1109/MRA.2019.2928823.

40. Tucker, Christopher. (2018). Artificial personality demonstration. 10.13140/RG.2.2.33133.72167.

Исследования в цифровой экономике. 2023. 1. P. 90–108
DOI: 10.24833/14511791-2023-1-90-108

Получено: 15 сентября 2022 г.,
Пересмотрено 3 ноября,
Принято: 20 декабря 2022 г.

ПРОБЛЕМА ФРАГМЕНТАЦИИ РЕГУЛИРОВАНИЯ РАЗРАБОТКИ И ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

А.А. Сажинов

Аннотация. В настоящее время активно обсуждаемым вопросом на международном уровне является регулирование разработки и применение систем с искусственным интеллектом. Одна из причин тому — значительный общественный резонанс данной темы. Многие международные игроки занимаются нормотворчеством в области регулирования разработки и применения искусственного интеллекта. В данный процесс вовлечены как универсальные, так и специализированные, как глобальные, так и региональные международные организации. Неформальные международные объединения также выступают активными участниками этой работы. Причины подобной активности носят комплексный характер и включают в себя различные факторы как объективного, так и субъективного характера. К тому же регулирование разработки и применения искусственного интеллекта все чаще становится политическим инструментом продвижения экономических интересов национальных компаний в сфере информационных технологий, воспринимается как средство поддержания актуальности тех или иных международных организаций их секретариатами. В отдельных странах национальные нормы в области искусственного интеллекта нередко рассматриваются как инструмент защиты от злонамеренного применения искусственного интеллекта западными правительствами, заключающегося в манипулировании внутриполитической ситуацией и во вмешательстве во внутренние дела. В зависимости от своей специализации международные организации демонстрируют различные подходы к регулированию разработки и применения искусственного интеллекта. В то время как специализированные площадки стремятся создать общепризнаваемые правовые рамки в области искусственного интеллекта, региональные организации склоны к политизации данной сферы. Совокупность перечисленных факторов с высокой долей вероятности приведет к фрагментации профильного регуляторного поля по границам отдельных государств или военно-политических альянсов.

Ключевые слова: Искусственный интеллект, нормотворчество, фрагментация, проблемы и вызовы цифровой среды

Для цитирования: Сажинов А.А. (2023). Проблема фрагментации регулирования разработки и применения искусственного интеллекта. – *Исследования в цифровой экономике*. №1, С. 90–108. DOI: 10.24833/14511791-2023-1-90-108

Об авторе:

Сажинов А.А. – независимый исследователь.
119454, Москва, проспект Вернадского, 76.
SazhinovAA@yandex.ru. ORCID: 0000-0002-5206-3113

Введение

В настоящий момент регулирование разработки и применения искусственного интеллекта значится в первых строках международной повестки. Связанными вопросами занимаются как органы власти, региональные и универсальные международные организации, так и академические и деловые круги, а также международные НПО. Во многих странах разработаны профильные стратегии и правовые акты, например, в области использования высокоавтоматизированных транспортных средств.[1] Международные организации ведут работу над различными инструментами мягкого и жесткого права. Среди примеров — Рекомендация ЮНЕСКО об этических аспектах искусственного интеллекта.[2] Над нормами жесткого права работают в Европейском союзе[3] и в Совете Европы[4]. Крупные компании принимают кодексы этики.[5] Международные НПО активно участвуют в этой деятельности как на национальном, так и на международном уровне в рамках продвигаемого странами их регистрации (преимущественно западными) «многостороннего подхода».

Причины такого внимания к искусственному интеллекту носят комплексный характер и обусловлены как объективными, так и субъективными факторами. К тому же данный процесс сильно политизирован, став инструментом вмешательства во внутренние дела государств. В результате высока вероятность формирования конкурирующих систем регулирования в области искусственного интеллекта в различных регионах. Перспективы разработки подлинно универсальных норм в этой области в ближайшем будущем вызывают сомнения.

Обзор литературы

Исследования и статьи по регулированию в области искусственного интеллекта преимущество фокусируются на вопросе самой его необходимости, а также потенциальном охвате, формах и связанных с ним рисках. При этом текущее нормотворчество в данной сфере зачастую игнорируется, попытки оценки его наиболее вероятных результатов не осуществляются.

Исследователи нередко рассматривают фрагментацию международного регулирования разработки и применения искусственного интеллекта как результат недостаточной компетенции участников данного процесса. Однако в действительности фрагментация рассматривается разработчиками международных норм как допусти-

---

[1] Council of Europe, AI initiatives – interactive map, electronic resource: https://www.coe.int/en/web/artificial-intelligence/national-initiatives, 2022

[2] Recommendation on the Ethics of Artificial Intelligence https://unesdoc.unesco.org/ark:/48223/pf0000380455, 2021

[3] 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, 2022

[4] Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law https://rm.coe.int/possible-elements-of-a-legal-framework-on-artificial-intelligence/1680a5ae6b, 2022

[5] Facebook's five pillars of Responsible AI https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/

мый риск при понимании, что соответствующее регулирование соответствует интересам профильных национальных экономических операторов из отдельных стран.

Исследователи исходят из посылки добросовестной деятельности участников международного нормотворчества и их нацеленности на разработку регулирования в области искусственного интеллекта во имя общественной пользы.[6] Другим проблемным моментом исследований является широко распространенное мнение, что выработка норм в сфере искусственного интеллекта — объективный процесс.[7] Однако это не так. Такой подход выглядит ограниченным: за международными акторами зачастую стоят национальные правительства, в свою очередь, продвигающие интересы крупнейших компаний. В этом причина осторожного подхода США к международному регулированию искусственного интеллекта. Нормы в этой области неизбежно станут сдерживать местных цифровых гигантов. В свою очередь Евросоюз, чья ИТ-индустрия значительно отстает от американской, демонстрирует высокую заинтересованность в нормотворчестве в данной сфере.

Поскольку проблематика искусственного интеллекта широко обсуждается как на национальном, так и на международном уровне, ее включение в собственную повестку позволяет международным организациям, НПО, отдельным функционерам привлекать к себе внимание. В практическом плане оно находит выражение во внебюджетных взносах на конкретные проекты.

Проблема фрагментации регулирование разработки и применения искусственного интеллекта затрагивается в статье "Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy"[8]. Однако это явление рассматривается скорее как один из рисков, нежели как основная тенденция и один из наиболее вероятных путей развития регуляторики в этой сфере.

Исследование

Настоящее исследование представляет собой попытку рассмотреть текущие нормотворческие процессы в области искусственного интеллекта и выделить их основные тенденции. В настоящий момент просматривается быстроразвивающаяся фрагментация международного регулирования разработки и применения искусственного интеллекта. Данный процесс обусловлен рядом факторов.

Во-первых, страны Запада заинтересованы в самостоятельной разработке норм в области искусственного для последующего распространения их применения на остальной мир. В практической плоскости данный подход находит выражение в Глобальном партнерстве по искусственному интеллекту, созданном по инициативе

---

[6] Regulating artificial intelligence: Proposal for a global solution https://www.researchgate.net/publication/341615195_Regulating_Artificial_Intelligence_Proposal_for_a_Global_Solution, 2020

[7] A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values https://www.researchgate.net/publication/359194759_A_European_Agency_for_Artificial_Intelligence_Protecting_fundamental_rights_and_ethical_values, 2022

[8] Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy https://www.researchgate.net/publication/341181427_Harnessing_artificial_intelligence_AI_to_increase_wellbeing_for_all_The_case_for_a_new_technology_diplomacy, 2020

западных государств и продвигающим инициативы, разработанные в рамках ОЭСР. Ситуация здесь схожа с той, что сложилась вокруг Конвенции по киберпреступности, изначально разработанной в рамках Совета Европы, но сейчас продвигаемой в качестве «золотого стандарта» далеко за рамками данной организации.

Во-вторых, как уже указывалось выше, многие международные организации, крупные НПО и политики рассматривают тематику регулирования разработки и применения искусственного интеллекта, находящуюся в фокусе общественного внимания, как источник тех или иных выгод.

В-третьих, ряд правительств небезосновательно полагают, что технологии искусственного интеллекта используются отдельными государствами для вмешательства во внутренние дела других стран. В результате на национальном уровне создаются правовые системы, нацеленные на профилактику подобного применения этой технологии.

Вышеназванные факторы с высокой вероятностью будут сохранять свою актуальность в долгосрочной перспективе. Обоснованным выглядит прогноз формирования ряда блоков, объединенных общими интересами в сфере искусственного интеллекта, на базе которых будут приняты документы жесткого права, имеющие ограниченный географический охват.

Нормотворчество в области искусственного интеллекта осуществляется на фоне глобальной конкурентной борьбы, в первую очередь разворачивающейся между американскими и китайскими цифровыми гигантами. Более широкая аудитория цифровых платформ означает не только больший доход, но и больший объем собираемых пользовательских данных, что, в свою очередь, предоставляет возможности для создания более эффективных алгоритмов искусственного интеллекта в будущем. Для победы в этой конкурентной борьбе используется широкий спектр инструментов. В качестве одного из них США использует политизацию данного сюжета, включающую в себя политически мотивированные пиар-кампании, принятие политически мотивированных норм, нацеленных на отсечение конкурентов от новых технологий.[9] В этой связи все чаще к искусственному интеллекту применяются субъективные характеристики без точного определения, как, например, демократические/ответственные/этичные разработка искусственного интеллекта или управление искусственным интеллектом.[10]

В ряде международных организаций полагают, что применение искусственного интеллекта в сфере государственного управления предполагает особые требования к таким системам. Например, Совет Европы в принятых в декабре 2022 г. Возможных элементах правовых рамок искусственного интеллекта, основывающихся на стандартах Совета Европы в сфере прав человека, демократии и верховенства права

[9] The Guardian, Biden administration imposes sweeping tech restrictions on China, Biden administration imposes sweeping tech restrictions on China https://www.theguardian.com/us-news/2022/oct/07/biden-administration-tech-restrictions-china, 2022, 2022

[10] EU, Intervention by President Charles Michel at the Summit for Democracy, 2021, GPAI, GPAI 2022 Ministers' Declaration, 2022, OECD, Tools for trustworthy AI a framework to compare implementation tools for trustworthy AI systems, 2019, OECD, OECD framework for the classification of AI systems, 2022

указывает: «В контексте специфического характера рисков, создаваемых искусственным интеллектом в сфере государственного управления в связи с особой ролью последнего в жизни общества, такие всеобъемлющие рамки могут быть дополнены юридически обязывающими либо необязывающими инструментами на секторальном уровне».[11] В ОЭСР государственное управление также рассматривается как особая сфера применения искусственного интеллекта.[12] Данный подход неизбежно приведет к закрытию рынка систем с искусственным интеллектом, применяемых в сфере госуправления, от стран, не относящихся в военно-политическим союзникам.

Искусственный интеллект широко используется социальными сетями и интернет-поисковиками для модерации и приоритизации контента. Подобные системы эксплуатируются всеми социальными сетями, такими как Фейсбук*, Инстаграм*, Ютьюб, Твиттер и многие другие[13]. Крупные западные цифровые компании демонстрируют политизированный и предвзятый подход к модерации и приоритизации контента. Например, только в 2022 году Ютьюб заблокировал страницы нескольких десятков российских телеканалов, включая региональные.[14] Социальные сети, воздерживающиеся от удаления страниц российских СМИ, блокируют доступ к ним в отдельных странах.

Другим примером политизации этой сферы является скандал, возникший в результате изменений в политике модерации в Фейсбуке*, заключавшихся в разрешении публиковать призывы к насилию против русских.[15] Таким образом, некоторые базирующиеся в США интернет-компании стали проводниками воли официального Вашингтона. Основной целью такого взаимодействия является создание механизма, обеспечивающего передачу сформулированных госорганами нарративов населению собственной страны, а также иных государств. Наблюдается все более активное использование социальных сетей, интернет-поисковиков и других цифровых сервисов в качестве инструментов внешней политики. Спецслужбы напрямую вмешиваются в модерацию и приоритизацию контента для продвижения интересов государства либо частных акторов.[16] События арабской весны, инспирированные социальными сетями, продемонстрировали деструктивный потенциал управляемой искусственным интеллектом приоритизации контента. Данная технология вкупе с персонализацией поисковой выдачи позволяют донести до большого числа людей информацию в наиболее подходящей для них форме.

Таким образом, цифровые компании превратились в средство информационно-психологических операций. Одним из наиболее обсуждаемых инструментов набирающего обороты военно-политического противостояния является информация.

---

[11] Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law https://rm.coe.int/possible-elements-of-a-legal-framework-on-artificial-intelligence/1680a5ae6b, 2022

[12] Hello, World: Artificial intelligence and its use in the public sector https://www.oecd-ilibrary.org/docserver/726fd39d-en.pdf?expires=1673561870&id=id&accname=guest&checksum=0DB9A8EF6610DDE66BCB9BFA125652C7, 2019

[13] Facebook*, How does Facebook use artificial intelligence to moderate content?

[14] Репрессии в отношении российских журналистов и СМИ за рубежом с начала СВО по защите Донбасса https://mid.ru/ru/press_service/journalist_help/repressions/

[15] Reuters, Facebook* allows war posts urging violence against Russian invaders, 2022

[16] BBC, Zuckerberg tells Rogan FBI warning prompted Biden laptop story censorship, 2022

Злонамеренное использование искусственного интеллекта через применение манипулятивных практик не позволяет достичь общего понимания по регулированию разработки и использования искусственного интеллекта в глобальном плане. Данная проблематика обсуждается в ООН на протяжении нескольких лет.[17] Более того, многие технологии искусственного интеллекта могут рассматриваться как имеющие двойное назначение, начиная с разнообразных дронов и заканчивая анализом спутниковых снимков.

Результатом становится выработка обеспечивающих информационную безопасность отдельной страны национальных и, возможно, региональных правовых режимов, ориентированных на минимизацию потенциала использования иностранных социальных сетей для вмешательства во внутренние дела и на одновременное продвижение собственных компаний. Как следствие, страны, становящиеся целями подобных информационно-психологических операций, пытаются им противостоять через продвижение собственных интернет-компаний и ограничение деятельности иностранных, в том числе законодательными мерами. Западные социальные сети запрещены в КНР и Иране. Эти страны создали собственные аналогичные сервисы, например, WeChat, Sina Weibo, Soroush, Cloob, Aparat.

Можно заключить, что искусственный интеллект будет играть всевозрастающую роль в социальных сетях в будущем. Госорганы и цифровые компании также не готовы делиться базами персональных данных, на которых возможно обучение алгоритмов, поскольку их наличие обеспечивает большую эффективность систем с искусственным интеллектом, действующих в конкретных обществах со всеми их национальными особенностями. Эти данные могут рассматриваться как закрытые. Становящиеся более комплексными и детальными базы данных обеспечивают более точные прогнозы реакций общества на различные раздражители. С высокой долей вероятности это одна из причин для столь трепетного отношения Евросоюза к защите персональных данных. Данный фактор также мог быть решающим при принятии в США решения о ряде уступок для возобновления получения персональных данных граждан стран-членов ЕС.[18] Такие данные рассматриваются в качестве стратегического ресурса, которым нельзя делиться с конкурентами.

В апреле 2021 г. Европейская комиссия представила проект регламента по искусственному интеллекту.[19] В случае своего принятия он станет юридически обязывающим нормативным актом в сфере искусственного интеллекта. Как указано в самом проекте, его основной целью является защита прав человека, а также персональных данных. Заслуживает внимание и фиксация другой задачи – защита общего рынка этих решений в Евросоюзе от фрагментации. Таким образом, Еврокомиссия осознает риск фрагментации регулирования искусственного интеллекта, однако готова его купировать лишь на региональном уровне в надежде, что впоследствии подходы Евросоюза

[17] United Nations Activities on Artificial Intelligence (AI), 2022

[18] EU, Questions & Answers: EU-U.S. Data Privacy Framework

[19] 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, 2022

удастся навязать странам, экспортирующим соответствующие программные решения в государства-участники ЕС. Краткосрочная же цель заключается в контроле над цифровыми компаниями на рынке Европейского союза, поскольку компании из ЕС крайне слабо представлены в сфере новых информационно-коммуникационных технологий, таких как интернет-поисковики, социальные сети и друге онлайн-продукты.

Проект Еврокомиссии содержит запрет на системы социальных рейтингов, манипулятивные алгоритмы и тотальное распознавание лиц. Возможной причиной этих ограничений является недостаток компетенций есовских компаний в этих сферах: в частности, в странах-членах Евросоюза отсутствуют собственные крупные социальные сети. Наиболее известная система по распознаванию лиц «Клеарвью ЭйАй» разработана в США. Результатом данного подхода в будущем станут разные подходы к применению прав человека в цифровой среде на глобальном уровне. Едва ли стоит ожидать, что страны, располагающие серьезным потенциалов в этих областях, станут запрещать соответствующие технологические решения.

Цифровые гиганты вкупе со специализированными НПО пытаются получить конкурентные преимущества через формулирование международных стандартов для различных отраслей применения искусственного интеллекта. Имеются многочисленные иллюстрации данного подхода. Например, функционирует специальная платформа для сотрудничества между цифровыми компаниями и Советом Европы. Они приглашаются на встречи профильных комитетов Совета Европы, которые заняты разработкой международных стандартов цифровой среды. Следует также отметить, что во встречах органов Совета Европы участвуют и соответствующие НПО. В своем ежегодном докладе[20] Международный союз электросвязи констатирует, что НПО участвовали в большинстве мероприятий ООН по тематике искусственного интеллекта. Вовлечение негосударственных акторов осуществляется через различные многосторонние механизмы, активно продвигаемые западными странами. Имеется значительный риск, что данный курс направлен на создание искусственного численного преимущества над остальным миром путем введения дополнительных западных представителей и размывания механизмов принятия решений в международных страуткурах с глобальным охватом. Как следствие, доверие незападных государств может быть утрачено указанными международными структурами. В результате возможно возникновение альтернативных форумов и платформ по регулированию разработки и применения искусственного интеллекта.

Еще одним фактором фрагментации и регионализации регулирования искусственного интеллекта является западная концепция мира, основанного на правилах.[21] Данный подход предусматривает, что группа развитых стран пытается разработать собственные правила в сфере искусственного интеллекта без надлежащих консультаций с большей частью мира. Расчет в том, что в дальнейшем эти правила

[20] United Nations Activities on Artificial Intelligence (AI) 2021 https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2021-PDF-E.pdf, 2022

[21] The concept "rules-based order" in international legal discourses https://www.mjil.ru/jour/article/viewFile/1190/1095, 2021

удастся навязать остальному миру. Этот план едва удастся реализовать. Тем не менее, в настоящий момент различные западные организации разрабатывают нормы в сфере искусственного интеллекта общего характера. Среди таких структур Европейский союз[22], Совет Европы[23], Организация экономического сотрудничества и развития[24] и ряд других неформальных альянсов, как Глобальное партнерство по искусственному интеллекту.[25]

Подобный подход неэффективен: он неизбежно приведет в разному уровню защиты прав человека в его отношениях с искусственным интеллектом. Фактически он означает, что персональные данные жителей развивающихся стран будут эксплуатироваться интернет-компаниями для повышения качества своих услуг, предоставляемых гражданам развитого мира. Страны, являющиеся крупными игроками в области искусственного интеллекта, но отлученные западными государствами от разработки норм, стремятся к универсальному регулированию этой сферы, разработанному в рамках ООН и иных структур с их участием.

Фрагментация регулирования искусственного интеллекта дополнительно углубится по мере того как цифровые технологии будут играть все более значимую роль в повседневной жизни, и люди станут проводить больше времени в виртуальной реальности, что предусматривает концепция метавселенных. В ближайшем будущем, по всей видимости, люди начнут потреблять большую часть информации, а также осуществлять большинство трат через инфраструктуру метавселенных. Как следствие, возможно усиление государственного контроля над данной сферой. Фактически подключение к метавселенной станет фактором выживания для многих предприятий, работающих на конечных потребителей. Отлучение компании от метавселеной неизбежно скажется на ее экономическом благополучии. Будущее регулирование искусственного интеллекта будет нацелено на создание рамок, ограничивающих доступ некоторых продуктов и компаний на внутренний рынок из соображений национальной безопасности либо по экономическим причинам. Можно ожидать, что экономические, военные и политические блоки создадут собственные метавселенные с собственной цифровой инфраструктурой.

Заключение

Наиболее вероятным сценарием развития правового поля в сфере искусственного интеллекта является его фрагментация по границам военно-политических альянсов. Евросоюз намерен ограничить некоторые сферы применения искус-

[22] 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, 2022

[23] Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law https://rm.coe.int/possible-elements-of-a-legal-framework-on-artificial-intelligence/1680a5ae6b, 2022

[24] Recommendation of the OECD Council on Artificial Intelligence https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449, 2019

[25] GPAI 2022 Ministers' Declaration https://www.gpai.ai/events/tokyo-2022/ministerial-declaration/, 2022

ственного интеллекта как противоречащие этическим ценностям.[26] В этой связи национальные и региональные правовые режимы, вероятно, будут носить в себе идеологическую компоненту которая значительной усложнит, если не сделает невозможным преодоление различий в будущем при условии выдвижения инициативы разработки глобального юридически обязывающего документа общего характера.

Что касается разработки глобального документа по тематике искусственного интеллекта, в связи с указанными разногласиями возможным видится лишь принятие инструментов мягкого права по отдельным вопросам в неполитизированных областях, не подразумевающим работы с персональными данными. В настоящий момент ряд таких документов разрабатывается в ООН[27], например, в области ядерных исследований и технологий в МАГАТЭ. МСЭ в сотрудничестве с ВОЗ работает над стандартизацией методов оценки методов диагностики и лечения на основе искусственного интеллекта. МСЭ также занимается стандартизацией систем, вовлеченных в автономное движение транспортных средств. Европейская экономическая комиссия ООН работает над определением универсальных технических требований к автономным транспортным средствам.[28]

* Facebook, Instagram and Meta признаны экстремистскими, запрещены в России.

Ссылки

1. Council of Europe, AI initiatives – interactive map, electronic resource: https://www.coe.int/en/web/artificial-intelligence/national-initiatives, 2022

2. Recommendation on the Ethics of Artificial Intelligence https://unesdoc.unesco.org/ark:/48223/pf0000380455, 2021

3. Facebook's five pillars of Responsible AI https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/

4. Questions & Answers: EU-U.S. Data Privacy Framework https://ec.europa.eu/commission/presscorner/detail/en/QANDA_22_6045

5. Biden administration imposes sweeping tech restrictions on China https://www.theguardian.com/us-news/2022/oct/07/biden-administration-tech-restrictions-china, 2022

6. United Nations Activities on Artificial Intelligence (AI) 2021 https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2021-PDF-E.pdf, 2022

7. 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, 2022

8. UNECE: driving progress on Autonomous Vehicles https://unece.org/DAM/trans/doc/2019/wp29grva/Autonomous_driving_UNECE.pdf, 2019

[26] 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, 2022

[27] United Nations Activities on Artificial Intelligence (AI) 2021 https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2021-PDF-E.pdf, 2022

[28] UNECE: driving progress on Autonomous Vehicles https://unece.org/DAM/trans/doc/2019/wp29grva/Autonomous_driving_UNECE.pdf, 2019

Automated driving UNECE is driving progress on Autonomous Vehicles https://unece.org/automated-driving

9. Automated driving UNECE is driving progress on Autonomous Vehicles https://unece.org/automated-driving

10. Hello, World: Artificial intelligence and its use in the public sector https://www.oecd-ilibrary.org/docserver/726fd39d-en.pdf?expires=1673561870&id=id&accname=guest&checksum=0DB9A8EF6610DDE66BCB9BFA125652C7, 2019

11. Tools for trustworthy AI a framework to compare implementation tools for trustworthy AI systems https://www.oecd-ilibrary.org/docserver/008232ec-en.pdf?expires=1673531426&id=id&accname=guest&checksum=B7CFD256835F222E44DFD7AEC2EB6D73, 2021

12. Recommendation of the OECD Council on Artificial Intelligence https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449, 2019

13. OECD framework for the classification of AI systems https://www.oecd-ilibrary.org/docserver/cb6d9eca-en.pdf?expires=1673562225&id=id&accname=guest&checksum=EE3A09363130A18004BF7D3A7FA486FB, 2022

14. Regulating artificial intelligence: Proposal for a global solution https://www.researchgate.net/publication/341615195_Regulating_Artificial_Intelligence_Proposal_for_a_Global_Solution, 2020

15. A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values https://www.researchgate.net/publication/359194759_A_European_Agency_for_Artificial_Intelligence_Protecting_fundamental_rights_and_ethical_values, 2022

16. Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy https://www.researchgate.net/publication/341181427_Harnessing_artificial_intelligence_AI_to_increase_wellbeing_for_all_The_case_for_a_new_technology_diplomacy, 2020

17. GPAI 2022 Ministers' Declaration https://www.gpai.ai/events/tokyo-2022/ministerial-declaration/, 2022

18. Legitimacy of what?: a call for democratic AI design https://www.hhai-conference.org/wp-content/uploads/2022/06/hhai-2022_paper_35.pdf, 2022

19. Chapter 8 of the National Security Commission on Artificial Intelligence's final report https://reports.nscai.gov/final-report/chapter-8/

20. Intervention by President Charles Michel at the Summit for Democracy https://www.consilium.europa.eu/en/press/press-releases/2021/12/10/intervention-by-president-charles-michel-at-the-summit-for-democracy/, 2021

21. Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on

human rights, democracy and the rule of law https://rm.coe.int/possible-elements-of-a-legal-framework-on-artificial-intelligence/1680a5ae6b, 2022

22. How does Facebook use artificial intelligence to moderate content? https://www.facebook.com/help/1584908458516247

23. Foreign reprisals against Russian journalists and media since the start of the special military operation to defend Donbass https://mid.ru/ru/press_service/journalist_help/repressions/?lang=en, 2022

24. Facebook allows war posts urging violence against Russian invaders https://www.reuters.com/world/europe/exclusive-facebook-instagram-temporarily-allow-calls-violence-against-russians-2022-03-10/, 2022

25. Zuckerberg tells Rogan FBI warning prompted Biden laptop story censorship https://www.bbc.com/news/world-us-canada-62688532, 2022

26. The concept "rules-based order" in international legal discourses https://www.mjil.ru/jour/article/viewFile/1190/1095, 2021

Journal of Digital Economy Research. 2023. 1. P. 90–108
DOI 10.24833/14511791-2023-1-6-90-108

Received: 15 September 2022,
Revised 3 November 2022,
Accepted: 20 December 2022

FRAGMENTATION OF ARTIFICIAL INTELLIGENCE INTERNATIONAL REGULATION: CHALLENGES

by Alexey A. Sazhinov

Abstract. The issue of AI development and application is high on the international agenda now. One of the reasons is its excessive popularity among the public. Numerous international actors are now developing AI regulations. It is both universal and specialized international organisations with global and regional reach. Informal international fora are also active in this area. This process is a complex one and incorporate different factors which are both of objective and subjective character. In effect, AI regulation all the more often becomes a political tool for promoting economic interests of national IT companies or an instrument for secretariats of international organisations to keep a high profile. National AI regulation is often considered as a shield against malign use of AI by Western governments to manipulate domestic processes in other countries and to interfere into internal affairs. There are also different approaches which are practiced by intergovernmental organisations depending on their specialization. While specialized fora strive to build common grounds for the development and use of AI, regional organisations attach broader perspective to AI regulation and tend to politicize it. These various factors are likely to result in a fragmented regulatory field in this area according to national borders or borders of military and political alliances.

Keywords: Artificial intelligence, regulation, fragmentation, digital challenges

For citation: Sazhinov A. (2023) Fragmentation of Artificial intelligence international regulation: challenges. *Journal of Digital Economy Research,* vol. 1, no 1, pp. 90–108. (in English).
DOI 10.24833/14511791-2023-1-90-108

Information about the author:

Alexey Sazhinov – Independent researcher.
76, Prospect Vernadskogo Moscow, Russia, 119454
SazhinovAA@yandex.ru
ORCID: 0000-0002-5206-3113

Introduction

AI regulation is a broadly discussed topic at the moment. In fact both public authorities, regional and universal international organisations, academia, businesses, international NGO's are now dealing with AI related issues. Many countries have launched AI strategies and legislation on AI application in specific areas such as autonomous vehicles (CoE, 2022). International organisations initiate various binding and non-binding instruments. One of the examples is UNESCO Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021). The European Union (EU, 2022) and the Council of Europe are in the process of developing their hard-law instruments. Large companies introduce ethical codes (Facebook*, Facebook's five pillars of Responsible AI). International NGO's take an active part in relevant activities at the national and international level within broadly promoted "multi-stakeholder approach". They are supported by their host-countries (primarily Western ones).

Reasons for such AI popularity are complex and unite both objective and subjective factors. Moreover, this process is highly politicized. AI has become an instrument for interference into domestic affairs. As a result, we are likely to see competing AI regulations in different regions. The development of a truly universal regulation is questionable in the near future.

Review

Research papers and articles on AI regulation mainly focus on the issue whether such regulation is needed and if so on its scope, shape and related risks while ignoring the analysis of existing developments in this area and their most probable results.

The researches usually consider that fragmented international regulation of artificial intelligence is the result of the lack of insight. However, usually this is not the case. The fragmentation appears to be an accepted risk in an attempt to develop such international regulation which would be friendly for national companies from particular countries.

The researches presuppose that the intentions of all international actors are benign and aim at developing AI regulation for the common good (see, for example, Erdélyi, Goldsmith, 2020). Another problem of existing research in the sphere of AI regulation is a widely spread assumption that this is an objective process (see, for instance, Stahl, Rodrigues, Santiago, Macnish, 2022). In fact, this is not the case. This appears to be a rather limited approach as international actors are usually guided by national interests in case of governments which may incorporate the interests of national largest companies. This is the reason why the US is rather cautious about international AI regulation. It would inevitably burden its largest IT giants. At the same time, the EU which IT industry is lagging behind the US one is so keen to regulate this sphere.

As artificial intelligence is broadly discussed both domestically and internationally, raising this issue is a way to attract attention to a particular international organization, NGO or an officer of an organisation. In practical terms, this means additional extrabudgetary contributions for specific projects.

The issue of fragmentation of AI regulation is touched upon in article "Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy" (Feijóo, Kwon, Bauer, Bohlin, 2020). However, this is considered as one of the risks rather than a mainstream and the most likely path of the development of AI regulation.

Discussion

This paper seeks to look on existing regulatory developments in this area and tries to analyze the effects of current tendencies in the sphere. At the moment we are witnessing a full-speed fragmentation of the international regulation of AI. There are some factors which presuppose this process.

First, this is an interest by Western countries to develop some AI rules in close circles in order to impose them lately on the rest of the world. In practical terms, this finds a reflection in Global Partnership on Artificial Intelligence which was established at the initiative of and effectively promotes the approaches which have been developed within OECD. The situation here is similar to that of the Convention on Cybercrime which was developed within the Council of Europe and is promoted as a "golden standard" in this sphere far outside the area of the Council of Europe.

Second, as mentioned previously, many international organisations, large NGO's, politicians seek to profit on the topic of artificial intelligence as it is broadly discussed now.

Third, it is a well-grounded perception by some governments that AI technologies are used by some states to interfere into domestic affairs of others. As a result there is a desire to establish national systems which would prevent such a use of artificial intelligence.

These factors are likely to be effective in a long term. Apparently, a number of blocks emerge united by common interests with regards of AI and they may create some hard-law legal frameworks for AI development and application. However, these would have a limited geographical scope.

AI regulation is discussed against the background of a global competition. It primarily unrolls between US and Chinese IT giants. We see a battle for billions-large revenues along with the number of users between large IT-companies. A platform's greater audience does not only mean greater income but also aggregated data which promises more efficient AI tools in the future. A wide spectrum of tools is used to win this competition. The US uses politicization of AI issues as one of them. This includes politically motivated PR campaigns and the adoption of politically motivated regulatory provisions which are aimed at barring competitors from new technologies

(The Guardian, Biden administration imposes sweeping tech restrictions on China, 2022). Thus, we see all the more subjective non-defined characteristics being attached to the AI. A democratic/responsible/ethical AI design/governance is one of them (EU, Intervention by President Charles Michel at the Summit for Democracy, 2021, GPAI, GPAI 2022 Ministers' Declaration, 2022, OECD, Tools for trustworthy AI a framework to compare implementation tools for trustworthy AI systems, 2019, OECD, OECD framework for the classification of AI systems, 2022).

Some international organisations propose that the use of AI in public domain implies special requirements for AI systems. For example, the Council of Europe in its Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law which were finalised in December 2021 (CoE, 2022) stated "in the context specificity of the risks posed by AI in the public sector in light of its specific role in society, such a transversal framework may be supplemented by additional legally binding or non-legally binding instruments at sectoral level". OECD considers public domain as a specific sphere of AI application as well (OECD, 2019) This approach would inevitably result in closing the market of public service AI solutions from the countries other than military and political allies.

AI is broadly used in moderation and prioritization of content in social media and search engines. Such systems were introduced by all the social media such as Facebook*, Instagram*, YouTube, Twitter and many others (Facebook*, How does Facebook use artificial intelligence to moderate content?). Moderation and prioritization sphere becomes all the more politicized. Large Western IT-companies demonstrate their readiness to apply biased and politically motivated moderation and prioritization techniques. For example, only this year YouTube has blocked accounts of some dozens Russian TV-channels including regional ones (Ministry of Foreign Affairs of the Russian Federation, Foreign reprisals against Russian journalists and media since the start of the special military operation to defend Donbass). Those social media which refrain from deleting pages of Russian media block them in specific countries.

Another example is the scandal that unrolled as a result of changes in Facebook* moderation techniques which allowed calls for violence against the Russians (Reuters, Facebook* allows war posts urging violence against Russian invaders, 2022). In fact, we see that some US internet companies become a long arm of their government. The main objective of such cooperation is to create a mechanism which would allow translating the narratives predefined by the authorities to the population of their own country as well as to that of other states. We witness all the more prolific use of social media, search machines etc. as foreign policy tools. Security services directly interfere into moderation and prioritization practices to promote governmental or private interests (BBC, Zuckerberg tells Rogan FBI warning prompted Biden laptop story censorship, 2022). Arab spring demonstrations which were ignited by social media tools showed a destructive potential of AI-driven content prioritization. AI prioritization due to its broad reach coupled with micro-targeting technologies allow reaching a large number of people with the messages which are most suitable for them.

So, IT companies turned into psychological operations' long arm. One of the most vocal and known tool applied in the hostilities which are gaining momentum in the international relations now is information. Weaponisation of this area through manipulative techniques effectively prevents reaching a common understanding on global AI regulation. This issue has been discussed in the UN for several years (see, for example report ITU, United Nations Activities on Artificial Intelligence (AI), 2022). Moreover, many AI technologies can be considered as double-purpose ones starting from various drones and ending AI tools to analyze satellite images.

This inevitably leads to the establishment of national and possibly regional regulations which would ensure information security of a country which would be aimed at minimizing potential use of foreign social networks to interfere into domestic affairs while promoting national ones. Consequently, governments which become targets of such psy ops strive to oppose them by promoting own internet companies and limiting activities of foreign ones including by regulatory means. Western social media are not permitted in China and Iran. These countries have developed their own services such as WeChat, Sina Weibo, Soroush, Cloob, Aparat.

We can assume that AI will play all the greater role in social media in the future. Ability to form large databases and to train AI thanks to them is another asset which business and governments are not likely to share as it means creating a better AI which is able to function in a concrete society with all its inherent features. This data may be considered as confidential. Since databases become more complex and detailed, they allow more precise predictions of societal behavior and reactions to different challenges. It looks as if this is one of the reasons for EU to demonstrate so much concern about personal data protection. This may also be a driving force for the US to make some concessions to the EU with this regard to resume a data transfer practice (EU, Questions & Answers: EU-U.S. Data Privacy Framework). Personal data is apparently considered as a strategic resources for AI advancement which is not to be shared with competing players.

In April 2021 the European Commission introduced its Proposal for a Regulation on artificial intelligence (EU, 2022). If adopted, it will constitute a hard-law regulation of AI. As declared in the document, its main rationale is the protection of human rights including personal data. Another noteworthy one is to prevent fragmentation of the European single market with regard to these solutions. Thus, the European Commission is aware of the risk of AI regulation fragmentation but is ready to overcome it at the regional level apparently hoping that it would be consequently possible to impose EU's regulatory approaches on the countries which export relevant AI solutions to the members of the European Union. A short-term goal is likely to control foreign IT companies in the European market as the EU companies are very sparely represented in the sphere of new communication technologies such as search engines, social networks and other internet solutions.

The Commission's proposal contains prohibition on use of social scoring systems, manipulation algorithms and indiscriminate facial recognition. A possible reason for these restrictions is lacking competence of EU companies in these spheres. Indeed, there is no social media from EU countries. The most known facial recognition Clearview AI is from the US. In the future this will give rise to different approaches to human rights application in digital sphere in global dimension. It is hard to imagine that the countries possess strong potential in these areas would prohibit relevant technologies

Largest IT-companies accompanied by specialised NGO's try to gain competitive advantages by wording international standards for different AI applications. There are numerous examples of such approach. For instance, there is a special platform for cooperation between digital companies and the Council of Europe. They are invited at the meetings of relevant Council of Europe committees which effectively develop international standards in digital area. It is also worth noting that relevant NGO's also take part in the meetings of competent Council of Europe bodies. International Telecommunication Union in its annual report (ITU, 2022) states that NGO's feature the majority of UN events on AI. This happens through various multi-stakeholder mechanisms which Western countries are actively lobbying. There is great risk that this is used to artificially outnumber the rest of the world by bringing in additional Western representatives and corrupt regulatory mechanisms in global international organisations. The latter may lose trust of non-Western global players. As a result we may see alternative forums and IT regulation platforms.

One more factor leading to fragmentation and regionalization of AI regulation is the western concept of a rules-based world order (Vylegzhanin, Nefedov, Voronin, Magomedova, Zotova, 2021). This concept implies that a group of developed states is trying to create their own rules for AI without proper consultations with the majority of the world. Further they are going to impose these rules on the rest of the world. This plan does not seem to be functional and efficient. Various Western organisations develop AI regulation of a general character. They are the European Union (EU, 2022), the Council of Europe (CoE, 2022), the Organisation for Economic Co-operation and Development (OECD, 2019) and some other informal alliances such as the Global Partnership on Artificial Intelligence.

Such approach demonstrates its inefficiency. It would inevitably result in unequal protection of persons in their relations with AI. In fact it may mean an appearance of such a regime when the data of the population in the developing countries would be misused and exploited by internet companies for the sake of better products for people in developed ones. Excluded countries which are large players in that market prefer to try to establish universal legislation in this area through UN system or other bodies in which they take part.

AI regulation fragmentation is to deepen further while IT play all the more important role in daily life and people spend more time in virtual reality which envisages for example the concept of metaverses. It appears that people will consume the bulk of information along with spending their money through metaverses' gateways in the

near future. This would result in more governmental control over this sphere. Effectively, the access to metaverses would be a matter of survival for many enterprises in B2C sphere. Barring a company from a metaverse would severely affect its sales. AI future legal regulation will be aimed at establishing a framework when access of some products and companies to domestic markets is limited due to national security considerations or economic reasons. It can be also expected that economic, military and political blocks would construct their own metaverses with own digital infrastructure.

Conclusion

The most real scenario for future AI regulation is fragmentation according to the borders of military and political alliances. We see that the EU intends to limit some AI applications as they allegedly contradict ethic values (EU, 2021). Thus, national and regional regulations are likely to bear an ideological component which would make difficult if impossible to bridge the differences in the future if a proposal to develop a global AI hard-law regulation of a general character is voiced.

When we are talking about universal AI regulation these differences may in fact only allow developing soft-law instruments of limited thematic scope covering some specific areas which deal with a limited quantity of personal data in areas which are in no way politicized. Several such tools are developed with the UN (ITU, 2022). For example, Project Domain of normative applied ethics at the intersection of AI and nuclear science, technology and applications, referred to as the Ethics of Nuclear and AI is developed within IAEA. ITU in cooperation with WHO works on a standardized assessment framework for the evaluation of AI-based methods for health, diagnosis, triage or treatment decisions, ITU also conducts standardization activities for services and applications enabled by AI systems in autonomous and assisted driving. UNECE tries to define technical requirements applied by the automotive sector worldwide (UNECE, 2019; UNECE, Automated driving UNECE is driving progress on Autonomous Vehicles).

* Facebook, Instagram and Meta are recognised as extremist organisations and are prohibited in the territory of the Russian Federation.

References

1. Council of Europe, AI initiatives – interactive map, electronic resource: https://www.coe.int/en/web/artificial-intelligence/national-initiatives, 2022

2. Recommendation on the Ethics of Artificial Intelligence https://unesdoc.unesco.org/ark:/48223/pf0000380455, 2021

3. Facebook's five pillars of Responsible AI https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/

4. Questions & Answers: EU-U.S. Data Privacy Framework https://ec.europa.eu/commission/presscorner/detail/en/QANDA_22_6045

5. Biden administration imposes sweeping tech restrictions on China https://www.theguardian.com/us-news/2022/oct/07/biden-administration-tech-restrictions-china, 2022

6. United Nations Activities on Artificial Intelligence (AI) 2021 https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2021-PDF-E.pdf, 2022

7. 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, 2022

8. UNECE: driving progress on Autonomous Vehicles https://unece.org/DAM/trans/doc/2019/wp29grva/Autonomous_driving_UNECE.pdf, 2019

9. Automated driving UNECE is driving progress on Autonomous Vehicles https://unece.org/automated-driving

10. Hello, World: Artificial intelligence and its use in the public sector https://www.oecd-ilibrary.org/docserver/726fd39d-en.pdf?expires=1673561870&id=id&accname=guest&checksum=0DB9A8EF6610DDE66BCB9BFA125652C7, 2019

11. Tools for trustworthy AI a framework to compare implementation tools for trustworthy AI systems https://www.oecd-ilibrary.org/docserver/008232ec-en.pdf?expires=1673531426&id=id&accname=guest&checksum=B7CFD256835F222E44DFD7AEC2EB6D73, 2021

12. Recommendation of the OECD Council on Artificial Intelligence https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449, 2019

13. OECD framework for the classification of AI systems https://www.oecd-ilibrary.org/docserver/cb6d9eca-en.pdf?expires=1673562225&id=id&accname=guest&checksum=EE3A09363130A18004BF7D3A7FA486FB, 2022

14. Regulating artificial intelligence: Proposal for a global solution https://www.researchgate.net/publication/341615195_Regulating_Artificial_Intelligence_Proposal_for_a_Global_Solution, 2020

15. A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values https://www.researchgate.net/publication/359194759_A_European_Agency_for_Artificial_Intelligence_Protecting_fundamental_rights_and_ethical_values, 2022

16. Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy https://www.researchgate.net/publication/341181427_Harnessing_artificial_intelligence_AI_to_increase_wellbeing_for_all_The_case_for_a_new_technology_diplomacy, 2020

17. GPAI 2022 Ministers' Declaration https://www.gpai.ai/events/tokyo-2022/ministerial-declaration/, 2022

18. Legitimacy of what?: a call for democratic

AI design https://www.hhai-conference.org/wp-content/uploads/2022/06/hhai-2022_paper_35.pdf, 2022

19. Chapter 8 of the National Security Commission on Artificial Intelligence's final report https://reports.nscai.gov/final-report/chapter-8/

20. Intervention by President Charles Michel at the Summit for Democracy https://www.consilium.europa.eu/en/press/press-releases/2021/12/10/intervention-by-president-charles-michel-at-the-summit-for-democracy/, 2021

21. Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on

human rights, democracy and the rule of law https://rm.coe.int/possible-elements-of-a-legal-framework-on-artificial-intelligence/1680a5ae6b, 2022

22. How does Facebook use artificial intelligence to moderate content? https://www.facebook.com/help/1584908458516247

23. Foreign reprisals against Russian journalists and media since the start of the special military operation to defend Donbass https://mid.ru/ru/press_service/journalist_help/repressions/?lang=en, 2022

24. Facebook allows war posts urging violence against Russian invaders https://www.reuters.com/world/europe/exclusive-facebook-instagram-temporarily-allow-calls-violence-against-russians-2022-03-10/, 2022

25. Zuckerberg tells Rogan FBI warning prompted Biden laptop story censorship https://www.bbc.com/news/world-us-canada-62688532, 2022

26. Questions & Answers: EU-U.S. Data Privacy Framework https://ec.europa.eu/commission/presscorner/detail/en/QANDA_22_6045

27. The concept "rules-based order" in international legal discourses https://www.mjil.ru/jour/article/viewFile/1190/1095, 2021

Исследования в цифровой экономике. 2023. 1. P. 109–136
DOI: 10.24833/14511791-2023-1-109-136

Получено: 8 октября 2022 г.,
Пересмотрено 10 ноября,
Принято: 3 декабря 2022 г.

ИСКУССТВЕННЫЕ ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ И ПРОБЛЕМА «ЕСТЕСТВЕННОГО» ДОВЕРИЯ

Е. Дегтева , О. Куксова

Аннотация. Развитие технологий ИИ обострило проблему гуманитарного, создав вызов на всех уровнях общественных отношений. Этические вопросы и, в частности, проблема доверия стали актуальными для сферы высоких технологий, учитывая тот факт, что ИИ выполняет все более значимые управленческие функции, которые ранее безоговорочно относились к человеку. Этот вопрос напрямую связан с системами искусственного интеллекта, которые уже воплотились в конкретных масштабных проектах. В данном исследовании авторы анализируют концепцию доверия через призму технологического развития. Для этого в исследовании представлен обзор исторических и современных интерпретаций концепции доверия и доказывается, что эта концепция актуальна и необходима для контроля рисков, возникающих при интеграции продуктов искусственного интеллекта в социальную жизнь. Авторы показывают, что требуется переосмысление понятий этики и морали в новом контексте. Это необходимое требование для создания доверенного ИИ и для достижения доверия при взаимодействии человека с технологическими продуктами. Авторы приходят к выводу о необходимости построения междисциплинарного диалога для интеграции теории и практики из многочисленных областей. Для этого необходимо создать общую базу знаний и платформу для общения между всеми заинтересованными сторонами, а также важно сформировать благоприятные условия для устойчивого и конструктивного взаимодействия. Поэтому доверие является актуальной концепцией, которую необходимо строить в многомерной системе координат, ориентированной на различные заинтересованные стороны, особенно с учетом роста взаимодействия между человеком и технологиями, другими словами, на всех уровнях и во всех масштабах.

Ключевые слова: ИИ, искусственный интеллект, система искусственного интеллекта, этика, мораль, доверие

Для цитирования: Дегтева Е., Куксова О. (2023). Искусственные интеллектуальные системы и проблема «естественного» доверия. – *Исследования в цифровой экономике.* №1, С. 109–136. DOI: 10.24833/14511791-2023-1-109-136

Информация об авторах:

Дегтева Е. – магистр экономики, Заместитель директора Департамента развития социальной сферы и сектора НКО, Министерство экономического развития Российской Федерации.
degteva@list.ru
ORCID: 0000-0002-4137-6317

Куксова О. – магистр экономики, Заместитель начальника управления, Аналитический центр топливно-энергетического комплекса ФГБУ "РЭА" Минэнерго России.
129110, г. Москва, ул. Гиляровского, д.39, стр.1.
kuksovaki@gmail.com
ORCID: 0000-0001-8169-1517

Введение

В последнее время возникла явная необходимость «вписать» гуманитарные ценности в продукты цифровых технологий, в частности, в рамках разнообразных проектов на основе ИИ, которые используются в государственном управлении, бизнесе, образовании и других сферах. Многочисленные теоретики, эксперты, практики, деятели культуры и искусства ведут горячую и оживленную дискуссию о том, как вовремя выявлять возможные риски и угрозы, кто и как должен разрабатывать и применять инструменты контроля рисков, как внедрять их в систему управления и общественного контроля с целью подготовки и адаптации общества к еще большему внедрению технологий ИИ. Сегодня именно этические вопросы широко обсуждаются и активно разрабатываются в контексте развития и применения ИИ в политической, экономической и социальной сферах.

Данные этические вопросы неразрывно связаны с актуальной проблемой доверия, которая сегодня требует внимания на всех уровнях: в международных отношениях, в экономике, в социальной сфере. Феномен доверия стал актуальным в сфере высоких технологий. Более того, дискуссия о взаимодействии человека с технологиями обостряется в условиях, когда ИИ начинает выполнять все больше управленческих функций, которые раньше мог делать только человек [7]. В данном случае мы подразумеваем не только «тактические» вопросы, но говорим и о технологиях, которые используются для управления на высоком уровне. Среди таких технологий можно выделить искусственные общества или цифровые двойники общества. Перед человечеством стоит задача обучить нейронные сети понимать» и «учитывать» этические ценности, либо четко определить, что такой результат невозможен. Этот вызов накладывает огромную ответственность не только на разработчиков, но и на всех участников процесса создания, внедрения и использования продуктов ИИ.

В то же время заявленная проблема уходит корнями в далекое прошлое - и вопрос доверия как такового, и проблема доверия к технологиям имеют долгую историю и сложное многогранное содержание.

Учитывая, что проекты искусственных обществ уже существуют, а их решения применяются на практике для проверки управленческих решений, очевидно, что дальнейшее развитие и использование требует большого внимания, поскольку такая деятельность влияет как на общество, так и на отдельного человека. Мы считаем, что доверие является ключевым фактором и требованием для

проверки приемлемости использования проектов ИИ. Это особенно актуально для значимых проектов, которые направлены на решение политических и социальных вопросов с использованием ИИ и, в частности, систем искусственного интеллекта, и в рамках которых жизненно важно обеспечить доверие между всеми заинтересованными участниками [8].

Обзор литературы

Взаимодействие технологических продуктов с антропосферой привлекало внимание мыслителей со времен античности. Эпохи Возрождения и Нового времени не стали исключением. В начале XX века ведущие мыслители были озабочены гуманитарными аспектами технического прогресса и взаимодействием человека и техники [15]. Мы не будем делать философско-исторический очерк, но сфокусируем наше внимание на том, что разнообразные вопросы, обусловленные технологическим развитием, стали объектом исследования философии достаточно давно. Даже как концепция понятие искусственного интеллекта еще не существовало. Когда концепция ИИ была сформулирована, философия приняла ее для рефлексии со всей серьезностью, потому что безусловно этот концепт приносит очень многое в культуру, поскольку в его сути соединяются ранее противоположные категории - естественного и искусственного, причем в практически сакральном (и не имеющим общепринятого определения) - в интеллекте. Вспомним, что американский ученый Д. Маккарти ввел в науку данный термин еще в 1955 году, и немногим позже его коллега М. Мински дал первое определение искусственному интеллекту в качестве науки, нацеленной на решение интеллектуальных задач (человека) с использованием машины [10].

Проблема ИИ стала благодатной почвой, на которой исследователи начали проверять возможности конвергенции естественного и искусственного, биологического и технологического - то, что сегодня представляется наиболее перспективными направлениями для развития науки и практики.

Уже в момент зарождения самого понятия ИИ были сформулированы по сути противоречивые подходы к его определению. В 1950-х и 1960-х годах в области ИИ выделились две точки зрения. Один подход, предложенный Д. Маккарти, использовал чистую логику и общую математическую основу. Маккарти считал, что исследования в области ИИ будут плодотворными, когда будет определено само понятие интеллекта (которое до сих пор имеет ряд трактовок). По его мнению, определение должно образовывать систему координат, определяемую эпистемологией и эвристикой. То есть эпистемология — это то, как приобретаются новые знания с точки зрения их полезности для решения проблем, а эвристика - как эти проблемы решаются: какими процедурами и с помощью каких средств [15].

М. Мински предложил другой подход, который впоследствии стал известен как коннекционизм (биологически инспирированный подход). Ученый призывал изучить, как работают человеческий мозг и психика, и на этой основе реали-

зовать знания в компьютерной модели. Этот подход дает основу для трактовки ИИ как чего-то субъективного, обладающего индивидуальностью, сознанием и даже эмоциями. [4].

В целом, сегодня можно говорить о существовании 3 основных подходов к развитию ИИ:

Искусственный интеллект (ИИ), или «Символический искусственный интеллект», или «Классический искусственный интеллект» - это подход «сверху вниз», направленный на воспроизведение когнитивных способностей без погружения на уровень отдельных нейронов.

Обобщенный искусственный интеллект, или сильный искусственный интеллект (AGI), в настоящее время является предполагаемой, гипотетической формой ИИ, согласно которому машина получит равный человеческому интеллект, и, соответственно, такой ИИ будет обладать самосознанием и свободой выбора

Биологически вдохновленный искусственный интеллект, или биологически вдохновленные когнитивные архитектуры, — это проекты, которые, основываясь на достижениях нейронауки, направлены на создание искусственных систем, воспроизводящих поведение и мышление биологических существ [11]. В частности, развитие теории искусственных нейронных сетей коррелирует и взаимодействует с исследованиями функционирования человеческого мозга [13, 24].

Так, первый подход развивает подход Маккарти, а второе и третье направления основаны на работах Минского. Мы видим, что на сегодняшний день нет единого подхода и четкого понимания направлений развития ИИ и, соответственно, нет единого взгляда на понимание этого вопроса [27]. Однако попытка осмысления этой общей актуальной темы в рамках философии техники началась в начале XX века. Один из крупнейших мыслителей прошлого века М. Хайдеггер указал на угрозы, исходящие от такого этапа в развитии цивилизации, на котором техника стала играть доминирующую роль. Мыслитель отмечал, что техника не является нейтральной по отношению к человеку, а человек теряет свою субъективность и перестает определять развитие цивилизации. Техника сама начинает преобразовывать природный мир, а непонимание смысла техники и ее возможностей чревато ошибкой, за которую Человек заплатит своим порабощением [15, 16].

В наше время горизонт интерпретаций и прогнозов очень широк и разнонаправлен. Следует отметить, что фундаментальная мысль стала объектом не только профессиональных философов, но и талантливых изобретателей, предпринимателей и бизнесменов. В частности, рационалист Э. Юдковски верит в возможность создания сознательного ИИ, но согласно его позиции сначала необходимо раскрыть тайну интеллекта [23].

Технооптимист Р. Курцвейл считает, что ИИ обязательно будет создан. При условии, что сам человек сможет сохранить ценности и культуру, ИИ поможет сделать мир лучше и обеспечит процветание цивилизации [14].

Трансгуманист Н. Бостром предсказывает, что создание ИИ приведет к следующему этапу в эволюции разумной жизни - появлению Сверхразума, который получит возможность контролировать человеческое существование [5]. А.В. Абрамова исследует этические аспекты развития ИИ, изучая и систематизируя мировые тенденции в развитии этических инструментов в области ИИ [1].

А.Ю. Алексеев руководит проектом «Искусственная личность», в рамках которого исследует перспективы компьютерного моделирования культурных, политических, социальных, экономических, нравственных и других аспектов общественной жизни и человека. Философ убежден в том, что исследования искусственной жизни являются весьма актуальной сферой для науки, и искусственная жизнь имеет значение как ключевой концепт для осмысления и дальнейшей реализации нано- и биотехнологий с далеко идущей целью – воссоздания биологической жизни с применением цифровых технологий. Как считает мыслитель, искусственная жизнь лежит в основе искусственных обществ. Философ много лет изучает искусственные общества, поскольку считает этот феномен является базовым для развития социальных и информационных технологий, которые задействованы в управлении общественной жизнью. Данное направление исследований привело к созданию концепта «искусственной личности», вокруг которой сейчас ведется оживленная дискуссия, и которой ряд ученых приписывает не возможность обладать самосознанием, но также иметь свободу выбора и способность к моральным суждениям и рефлексии своих мыслей и действий [3].

Изучением особенностей человеческого интеллекта и ИИ занимается И.М. Дзялошинский, который стремится доказать принципиальную разницу между человеческим и искусственным интеллектом [10]. Необходимо подчеркнуть, что в последнее время данной проблематикой стали заниматься не только ученые, но и публицисты, журналисты и писатели. Многочисленные деятели культуры размышляют о перспективах развития ИИ, о рисках и возможностях, которые открывают продукты технологий, основанных на ИИ. В частности, можно упомянуть книги «Искусственный ты: Машинный интеллект и будущее разума», написанную С. Шнайдер, и «Наше последнее изобретение: Искусственный интеллект и конец человеческой эры», автором которой является Дж. Баррат [20, 33].

В последнее время появился ряд публикаций, посвященных изучению потенциальных опасностей злоупотребления нейронными сетями владельцами технологий и противодействия им, полагаясь на доверенный искусственный интеллект. [2].

Именно разработка доверенного искусственного интеллекта представляет собой наиболее перспективную концепцию создания и понимания идеи этичного ИИ на сегодняшний день. Однако само понятие «доверенный» уже вышло за рамки теории и начало использоваться в качестве юридического термина в нормативно-правовом регулировании. В частности, понятие доверенного ИИ определено в Этическом руководстве для доверенного ИИ Группы экспертов

высокого уровня по искусственному интеллекту Европейской комиссии (Ethics guidelines for trustworthy AI, 2019) [30]. Согласно этому документу, заслуживающий доверия ИИ отвечает следующим требованиям, которые на первый взгляд относятся скорее к деятельности человека, чем к технологии:

- Этичный (существующий в соответствии с этическими ценностями и этическим кодексом);
- Законный (не противоречащий закону);
- Достоверный (надежный и устойчивый).

В этом контексте необходимо задуматься над пониманием такого понятия, которое на первый взгляд относится в основном к социальной сфере, как доверие. Это понятие в определенной степени объединяет представленную тему, поэтому далее будут предложены некоторые актуальные подходы к определению данного понятия. Именно доверие выступает как объединяющее понятие и как практический инструмент, который позволит ему стать объединяющим звеном между человеческими ценностями и технологическим развитием [29].

Это понятие вошло в теорию и социальную практику с ранних этапов цивилизации, практически одновременно его ввели Конфуций в Древнем Китае и Перикл в Древней Греции, описывая отношения, основанные на доверии, как необходимое условие процветания государства и общества [28].

Впоследствии, в разные исторические периоды, в зависимости от идеологии, стратегии развития и личностных качеств лидеров и государственных идеологов, подходы к определению и оценке значения этого понятия неоднократно менялись самым радикальным образом.

Обращаясь к реалиям XX века, отметим, что в период после Второй мировой войны проблемы доверия и противоположного состояния - недоверия в контексте гонки вооружений и политики ядерного сдерживания были одними из ключевых. В рамках этой политики в США возникла школа политического реализма, разработанная Г. Киссинджером, Т. Шеллингом и рядом других ученых и стратегов. Данный подход предполагает, что уровень доверия политических акторов (в данном случае игроков на международном поле) низок, а стабильность (предсказуемость) действий партнеров определяется балансом сил и системой сдержек и противовесов. Только такой подход может обеспечить предсказуемость в политической игре [34].

Неолиберальный подход подходит к вопросу сложной взаимозависимости как к возможности с помощью многочисленных каналов и инструментов построить доверие между международными партнерами и тем самым обеспечить высокий уровень доверия во внешнеполитическом дискурсе [32].

S. Хантингтон отмечает, что особая ценность доверия возникает в периоды трансформации и кризисов, и подчеркивает важность институционального доверия, поскольку именно стабильные институты позволяют сохранить общество и цивилизацию в трудные времена [35].

Постмодернистский взгляд по сути отвергает доверие, а Ж.-Ф. Лиотар выделяет в качестве тенденции глобальное недоверие к «метанарративам», которые должны оправдывать реальность. З. Бауман указывает на эфемерность, нестабильность и ненадежность постмодернизма и утверждает, что центр тяжести смещается от государственных и международных институтов к глобальному рынку с его потребительскими ценностями и заменой рационального знания поверхностным воздействием СМИ [26].

Последняя мысль приводит нас к проблеме технологического развития, поскольку оно позволило глобализировать экономику и культуру, развить СМИ и массовую культуру, а затем социальный и мобильный веб, и все это благодаря технологиям и продуктам искусственного интеллекта. Таким образом, широта обсуждения, масштаб проблемы и горизонт мнений указывают на то, что данный вопрос является сложной, актуальной и перспективной областью исследования.

Обсуждение

Рост значения технологий в общественной жизни, и внедрение их продуктов в важные процессы, управляющие текущей деятельностью и развитием цивилизации, оказывают серьезное влияние на изменение содержание и нп появлением нового набора социальных ценностей, а также непосредственно влияют на формулировки новых стратегий.

В продолжение озвученной темы – перспектив развития искусственных интеллектуальных систем, постараемся определить данное понятие. Мы подойдем к нему с точки зрения функционального подхода: ИИС обладает набором свойств и ресурсов, которые позволяют ее находиться во взаимодействии с внешним миром, причем взаимодействие представляет собой активную реакцию на происходящие события. Систем тем более автономна, чем большей степенью независимости от команд управления она обладает. В этом смысле ИИС является в большей или меньшей степени самодостаточной системой, обладающей способности находить решения тем или иным точечным задачам [8].

В частности, нейронная сеть, созданная компанией Salesforce, тестирует и разрабатывает идеальную налоговую систему в симулированной среде [36]. Так называемый AI Economist основан на обучении с подкреплением, которое предполагает применение вознаграждений и наказаний к машинным алгоритмам для максимизации желаемых результатов. По такому же принципу, например, были созданы алгоритмы Google DeepMind AlphaGo и AlphaZero. Эксперимент был направлен на то, чтобы найти решение, которое позволило бы более эффективно и справедливо управлять системой налогообложения. Программа нацелена на создание огромного количества экосистем с работниками-теоретиками, которые торгуют валютой и строят дома. Уровень заработной платы и набор навыков варьируются, а ИИ определяет оптимальные налоговые ставки. Ученые обращают

внимание на то, что такой подход учитывает возможность учитывать иррациональное поведение у субъектов. Классические экономические теории нередко игнорируют эти аспекты при выстраивании моделей [11].

Актуальность таких проектов возрастает в связи с тем, что человечество сегодня развивается ускоренными темпами, и беспрецедентная скорость изменений регулярно вызывает глобальные кризисы. Именно поэтому стало популярным понятие «черных лебедей», регулярно возникающих крупномасштабных рисков и угроз, которые невозможно предвидеть, обнаружить или идентифицировать заранее, поскольку они исходят из областей, находящихся за пределами знаний, и во многом (конечно, не полностью) обусловлены стремительным развитием технологий, достижения которых интегрируются в многообразные сегменты общественной жизни, расширяя техносферу и стирая границы между гуманитарной и технологической областями [12].

В этом контексте классические инструменты экономической теории, стратегического менеджмента и даже антикризисного управления теряют убедительность и перестают приносить результаты. Для изучения мира, в котором техносфера заняла важное место, необходимы инструменты, включающие также возможности цифровых технологий, в частности ИИ. В этом смысле искусственные интеллектуальные системы очень перспективны, поскольку в данной концепции используется агент-ориентированное моделирование. Этот метод исследует поведение децентрализованных агентов и то, как это поведение определяет поведение всей системы. Использование агент-ориентированного моделирования в искусственных интеллектуальных системах открывает большие возможности не только для прогнозирования развития общества в обычных условиях, но и для предсказания возможных кризисов и минимизации ущерба от таких событий. Конечно, такой подход является сложной задачей, поскольку он должен учитывать большое количество сложных факторов и основываться на результатах многогранных исследований реальных сложных систем [21].

В этом контексте доверие является важным и необходимым условием, поскольку мы имеем дело с радикальной неопределенностью. В общественной жизни, на индивидуальном уровне, да и во всем мире, неопределенность будущего выходит на первый план, как на практическом, стратегическом, так и на философском, буквально метафизическом уровне. Однако в этих условиях перспективные проекты на основе ИИ могут стать решением, поскольку классические инструменты сегодня не могут показать эффективных результатов. Темпы изменений, появление новых условий, ускорение развития требуют учета всех факторов, а классические теории и методологии не способны учесть важные, но незамеченные и не просчитанные элементы, такие как иррациональность и другие особенности человеческой природы, и все вытекающие из этого ограничения и возможности.

В условиях радикальной неопределенности доверие является требованием, обеспечивающим смысл и основу для дальнейшего развития. Более того, в обществе, где развитие определяется явлениями, которые не были изначально зада-

маны, а возникли сами по себе, в частности, когда совокупный эффект действия совокупности индивидов отличается от эффекта действия индивидов и неэргодическими процессами (для неэргодических процессов вероятность, наблюдаемая в прошлом, не применима к будущим процессам), неопределенность становится основной характеристикой. В то же время общество увеличивает неопределенность за счет собственной несогласованности, креативности, и именно в этих условиях требуются искусственные интеллектуальные системы для функционирования и, более того, для обеспечения уверенности в своей деятельности и в ее результатах [21]. На сегодняшний день мы уже нередко способны моделировать потенциальные последствия работы подобных продуктов технологий. При этом последствия можно охарактеризовать как с положительной, так и отрицательной сторон. В первую очередь, такие системы направлены на обеспечение безопасности и благополучия, с другой стороны, применение решений таких систем в реальной жизни может затронуть такие естественные ценности, как свобода, частная жизнь и личная информация реальных людей. И самое главное, необходимо четко определить и установить, кто будет играть доминирующую роль в новых условиях и какой баланс будет в управлении социальной жизнью между государством, технологическими корпорациями, обществом и личностью.

Кроме того, наряду положительными перспективами, возникающими в свете развития и распространения интеллектуальных систем, следует обратить внимание и на потенциальные риски, обращенные в общественное измерение. Наиболее очевидными представляются проблемы в области профилирования и агрегирования персональных данных. Социальные сети, развлекательные сервисы, являющиеся интеллектуальными системами, могут накапливать большие объемы персональной информации о жизнедеятельности каждого пользователя, его личных предпочтениях, взглядах, состоянии здоровья, семейных отношениях и другой информации. В случаях, когда система научилась распознавать пользователя по его поведенческим факторам, идентификация может происходить даже тогда, когда пользователь появляется на партнерских сервисах под псевдонимом или анонимно. Такое профилирование данных может нарушать права и свободы человека [22].

Агрегация пользовательских данных для улучшения сотрудничества с партнерскими сервисами сама по себе не опасна, но если накапливается достаточное количество личной информации из различных источников, система может стать привлекательной для злоумышленников с целью создания бреши в данных. Поэтому агрегирующие сервисы обязаны своевременно повышать безопасность всех задействованных систем. Однако в настоящее время эта проблема далека от глобального решения. В настоящее время этот вопрос рассматривается на самом высоком уровне. По мнению Экономического и социального совета ООН, необходимо уточнить, кому принадлежат данные, обеспечить ответственное использование данных, защиту частной жизни, подотчетность и справедливость в

социальных сетях, платформах с широким участием и порталах онлайн-коммерции. Основная цель этих мероприятий - укрепление доверия и стабильности в цифровом пространстве, в частности в области ИИ [17].

Другим примером возможных негативных социальных последствий при использовании интеллектуальных систем является тема, связанная с большими данными. Предполагается, что большие данные должны быть обезличены, но возможность извлечения персональных данных из обезличенных потоков данных (деперсонализация) и получение индивидуальной информации из больших данных — это крайне непростая работа. В этом случае человек не защищен от несанкционированного использования личной информации и, конечно, доверие к системам разведки уже не гарантировано. Список социальных последствий можно было бы продолжить, поскольку представленные вызовы не имеют на данный момент общепринятых решений. В этой связи мы должны исследовать этическую проблематику, возникающую при работе подобных систем. То есть необходимо изучить, насколько поведение таких систем может быть согласовано с этическими идеями и правилами, зафиксированными в соответствующих кодексах [32].

Существует потребность в определении и согласовании целей, на выполнение которых направлено создание и дальнейшее использование интеллектуальных систем. При этом важно определить границы автономии и спектр вопросов, которые возможно решать с помощью таких систем. Отдельной темой, которая сейчас активно исследуется, является категоризация потенциальных рисков и инструменты их контроля. При этом необходимо учитывать, что на сегодняшний день интеллектуальные системы представляют собой существующие технологии, которые не только имеют потенциал развития, но уже сейчас могут оказывать влияние на общественное развитие. Поэтому вопрос создания и применения кодекса этики для искусственного интеллекта является вполне прагматичной темой – ведь в данном случае мы говорим технологиях, чья работа направлена на управление, прогнозирование и выбор моделей человеческого поведения [23]. С учетом динамики общественного развития следует на постоянной основе осуществлять анализ происходящих явлений и тенденции, происходящих в обществе, и по результатам проведенной работы оперативно менять законодательство и фиксировать общественные нормы. Поскольку этот процесс сложный, для его осмысления и реализации следует интегрировать работу экспертов, представляющих разнообразные области теории и практики. Речь идет и о работе, затрагивающей разномасштабные активности – это и научная, и законодательная, и практическая работа, в которую должны быть вовлечены и все общество, и государство, и экспертное сообщество. Только при такой совместной работе возможно достичь практических и эффективных результатов.

Соглашаясь с необходимостью активной работы в области создания и применения положений Этического кодекса ИИ, мы должны обратить внимание, что этот инструмент сам по себе не сможет решить озвученные ранее пробле-

мы [25]. Не секрет, что в общественном пространстве в последние годы границы между концептами этики, морали и нравственности размылись, и в повседневном языке они используются как синонимы. Кроме этого эти концепты стали использоваться буквально во всех сферах в очень широком смысле. Чтобы исключить неопределенность, постараемся дать определение этим понятиям. Этика в широком смысле определяет, что есть «добро» и «зло» и противопоставляет их в рамках системы мировоззрения. Мораль, в свою очередь, является инструментом управления социальным посредством принципов и норм [9]. Нравственность же –практические действие и поведение человека в реальности, и последствия такого поведения, отражающееся впоследствии в полученном и осмысленном опыте индивида и общества. Если выразить это иначе, то общественная мораль представляет собой объективную оценку реального поведения в обществе. В связи с этим применение понятия «этики» в абстрактном смысле, как мы полагаем, не несет никакой пользы и даже опасно. В этом случае происходит лишь появление в культурном пространстве очередных симулякров, распространение которых чревато разрушением человеческой природы, разрушением социального и социальными потрясениями. Мы считаем, что современные социальные тенденции образуют единый нравственный узел - фундамент, на котором должны строиться ступени нравственного прогресса [9].

Именно по этой причине вокруг данных понятий существует множество спекуляций, цена которых в связи с использованием искусственных интеллектуальных систем может оказаться слишком высокой. Проблема заключается в том, что нет единого понятийного аппарата для базовых определений, нет должного уровня цифровой и гуманитарной грамотности, а среди заинтересованных участников часто присутствует предвзятость или конфликт интересов. Цель заключается в том, чтобы проговорить и согласовать качественно новые определения для ново созданных категорий, с учетом того, что проблематика является по сути многомерной. Успешное выполнение этого требование будет способствовать более объективному анализу с учетом сложностности явлений. По нашему мнению, не стоит недооценивать необходимость в кратчайшие сроки сфокусироваться на изучении и поиске решений данным вопросам, поскольку инвестиции в создание прикладных продуктов ИИ в разнообразных секторах экономики растут на регулярной основе. Этот бум привел к тому, что уже сейчас мы можем говорить о существовании фактически автономных систем, поэтому речь идет не о теоретических угрозах, а о фактических рисках, которые уже сегодня чреваты масштабными последствиями [37].

Вышеприведенные факты подтверждают важную роль исследований в области этичного поведения интеллектуальных систем. Знание в этой сфере и изучение данной проблематики важно как с теоретической, так и с практической позиций. Интеллектуальные системы, как мы уже отметили, будут использоваться все чаще и чаще и будут влиять на общество и человека. В этих условиях крайне важно обеспечить прозрачность и гибкость при создании цифрового двойника

общества. Несомненно, важно обеспечить укрепление доверия к таким проектам. В научном и особенно общественном дискурсе можно встретить как фантастические, так и обоснованные опасения в связи с рисками безответственного использования технологий ИИ, вплоть до угрозы потери человеком автономии и свободы [6]. В то же время общепринятого и четкого плана формирования доверия в этой сфере не существует, однако исторический опыт внедрения инноваций показывает, что доверие обычно рождается, если технологии полезны и безопасны, а их регулирование прозрачно.

Говоря о практической стороне этического регулирования модели искусственного общества, необходимо решить проблему формализации этики в формате, доступном для технологий ИИ, а также реализовать принятие этических решений в самой автономной системе на основе ограничений ИИ в соответствии с моральными нормами или путем обучения ИИ распознавать этические конфликты и принимать свои индивидуальные решения [2]. Это требует повышения ответственности при разработке и использовании этих систем, а также максимально оперативного контроля ввиду серьезных последствий неправильного этического выбора.

Существующий опыт систематизации основных этических вопросов, связанных с созданием и развитием сложных проектов на основе технологий ИИ, позволяет выделить проблемы конфиденциальности и защиты данных, прозрачности и, возможно, как общий знаменатель, доверия к предиктивным данным.

Заключение

Проведенное исследование позволяет сделать вывод, что сегодня необходимо изучать технологическое и социальное развитие в комплексе, и такой глобальный взгляд позволит вовремя выявить и найти решение проблемам, вызванных развитием и распространением ИИ и таких продуктов, как системы искусственного интеллекта.

В исследовании авторов был проведен анализ концепции доверия через призму технологических реалий. Авторы представили обзор исторических и современных интерпретаций концепции доверия и показали, что эта концепция является весьма перспективной и даже необходимой в качестве инструмента контроля рисков, возникающих при создании цифровых технологических продуктов, что особенно очевидно в случае интеллектуальных систем.

В то же время гуманитарные ценности сегодня требуют определенного переосмысления и новых определений, которые будут соответствовать новой возникающей реальности. Авторы показывают, что использование понятий этики и морали без качественного пересмотра их смысла в соответствии с современными требованиями несостоятельно и является формальным интеллектуальным упражнением, которое в действительности не приведет к созданию доверенного ИИ и формированию доверия к взаимодействию человека с технологиями. Авторы утверждают, что данный вопрос является вызовом, поскольку современная

цивилизация на нынешнем этапе своего развития включает в себя широкий круг заинтересованных сторон, представляющих политические, экономические и социальные сферы, от глобального сообщества до отдельного человека, и очевидно, что для построения этого разнонаправленного диалога необходимо сформировать новый язык и новые определения терминов, определяющих требования и позволяющих описывать ценности. По мнению авторов, вклад в развитие данной проблематики заключается в том, что проведено значительное исследование интерпретаций соответствующих понятий, а представленные выводы могут быть использованы для следующих конкретных шагов.

Для всестороннего понимания и управления текущими процессами авторы предлагают следующие виды деятельности:

1. Необходимо построить междисциплинарный диалог для интеграции теории и практики из многочисленных областей, принимая во внимание глобальный уровень вопросов, которые мы исследуем.
2. Для этого необходимо создать общую базу знаний и платформу для общения между всеми заинтересованными сторонами.
3. В этих условиях станет возможным устойчивое и конструктивное взаимодействие, в ходе которого можно будет сформулировать и согласовать новые определения обсуждаемых понятий с целью их дальнейшей интеграции в практику, как социальную, так и научную.

Таким образом, авторы считают, что в современных условиях доверие - это категория, которую нужно и можно строить в нескольких измерениях - между заинтересованными сторонами, между человеком и технологией, то есть на всех уровнях и во всех масштабах.

Ссылки

1. Абрамова А.В. Этика в области искусственного интеллекта - от дискуссии к научному обоснованию и практическому применению: аналитический доклад / А.В. Абрамова, А.Г. Игнатьев, М.С. Панова; МГИМО МИД России, Центр искусственного интеллекта; XIII Съезд Российской ассоциации международных исследований (РАМИ) (Москва, 14-16 октября 2021 г.). - М.: МГИМО-Университет, 2021. - 24 с.

2. Авдошин С.М., Песоцкая Е.Ю. Доверенный искусственный интеллект как способ цифровой защиты // Бизнес-информатика. - 2022. - №2. - С. 62-73.

3. Алексеев А.Ю. Когнитивно-технологические проекты искусственной личности // Человек: образ и сущность. Гуманитарные аспекты. - 2014. - No. 1. - С. 156-174.

4. Болдачев А.В. Эссе: мифы и тупики искусственного интеллекта // Философские проблемы информационных технологий и киберпространства. - No. 2 (18). - С. 27-41.

5. Бостром, Н. Искусственный интеллект. Этапы. Угрозы. Стратегии. - СПб: Манн, Иванов и Фербер, 2016. - 490 с.

6. Булавинова М.П. Риски и угрозы новых технологий на основе искусственного интеллекта. (обзор) // Социально-гуманитарные науки. Отечественная и зарубежная литература. Сер. 8, Наука о науке: Реферативный журнал. - 2018. - No. 2. - С. 23-41.

7. Гуров О. Н. Искусственные интеллектуальные системы и решение социальных задач: проблема доверия // Искусственные общества. – 2021 – Т. 16 – Выпуск 1 URL: https://artsoc.jes.su/s207751800014095-3-1/. DOI: 10.18254/S207751800014095-3

8. Гуров О. Н. Этичное взаимодействие с интеллектуальными системами // Искусственные общества. – 2020 – Т. 15 – Выпуск 3 URL: https://artsoc.jes.su/s207751800010905-4-1/. DOI: 10.18254/S207751800010905-4

9. Гуров О.Н. Метавселенная — из сумерек во тьму перелетая? // Наука телевидения 2022 18 (1). С. 11–46. DOI: https://doi.org/10.30628/1994-9529-2022-18.1-11-46

10. Дзялошинский И.М. Искусственный интеллект: Гуманитарная перспектива // Вестник Новосибирского государственного университета. Серия: История, филология. - 2022. - №6. - P. 20-29.

11. Дубровский, Давид. (2021). Задача создания искусственного общего интеллекта и проблема сознания. Российский журнал философских наук. 64. 13-44. 10.30727/0235-1188-2021-64-1-13-44.

12. Дубровский, Давид. (2022). Эпистемологический анализ социальной и гуманитарной значимости инноваций искусственного интеллекта в контексте искусственного общего интеллекта. Российский журнал философских наук. 65. 10-26. 10.30727/0235-1188-2022-65-1-10-26.

13. Карпов В. Е., Готовцев П. М., Ройзензон Г. В. К вопросу об этике и системах искусственного интеллекта // Философия и общество. - 2018. - No. 2 (87). - P. 84-105.

14. Курцвейл, Р. Эволюция разума. Как расширение возможностей нашего разума решит многие мировые проблемы. - М.: ЭКСМО, 2015. - 352 p.

15. Финн В.: «Не все функции естественного интеллекта могут быть формализованы и автоматизированы» // Коммерсантъ URL: https://www.kommersant.ru/doc/4198609?ysclid=lafntgvfcq626184900 (дата обращения: 11/14/2022).

16. Хайдеггер М. Время и бытие: статьи и выступления. - М.: Республика, 1993. - 447 p.

17. Этика и «цифра» - краткое резюме. Робот-врач, робот-учитель, робот-полицейский: социальные риски и этические вызовы для промышленности: Краткий обзор политики для второго тома отчета «Этика и цифровые технологии: Этические вызовы цифровых технологий». - Москва: РАНХИГС, 2020. - 45 с.

18. AI Designed a Fairer Tax System // Hightech.fm URL: https://hightech.fm/2020/05/07/ai-economist (Accessed: 11/14/2022).

19. Anton, Eduard & Kus, Kevin & Teuteberg, Frank. (2021). Is Ethics Really Such a Big Deal? The Influence of Perceived Usefulness of AI-based Surveillance Technology on Ethical Decision-Making in Scenarios of Public Surveillance. 10.24251/HICSS.2021.261.

20. Barrat J. (2013). Our final invention: artificial intelligence and the end of the human era (First). Thomas Dunne Books.

21. Bookstaber, Richard. (2017). The End of Theory: Financial Crises, the Failure of Economics, and the Sweep of Human Interaction. 10.1515/9781400884964.

22. Comi, Antonio & Nuzzolo, Agostino. (2014). Individual pre-trip path choice modelling in transit advisor tools: theoretical and empirical evidences.

23. Goertzel, Ben. (2015). Superintelligence: Fears, Promises and Potentials: Reflections on Bostrom's Superintelligence, Yudkowsky's From AI to Zombies, and Weaver and Veitas's "Open-Ended Intelligence". Journal of Ethics and Emerging Technologies. 25.55-87. 10.55613/jeet.v25i2.48.

24. Gurjar, Arunaben & Patel, Shitalben. (2022). Fundamental Categories of Artificial Neural Networks. 10.4018/978-1-6684-2408-7.ch001.

25. Kopalle, Praveen & Gangwar, Manish & Kaplan, Andreas & Ramachandran, Divya & Reinartz, Werner & Rindfleisch, Aric. (2021). Examining Artificial Intelligence (AI) Technologies in Marketing Via a Global Lens: Current Trends and Future Research Opportunities. International Journal of Research in Marketing. 39. 10.1016/j.ijresmar.2021.11.002.

26. McHale, Brian. (2013). Afterword: Reconstructing Postmodernism. Narrative. 21. 357-364. 10.1353/nar.2013.0020.

27. Muggleton, Stephen, and Nicholas Chater (eds), Human-Like Machine Intelligence (Oxford, 2021; online edn, Oxford Academic, 19 Aug. 2021), https://doi.org/10.1093/oso/9780198862536.001.0001, accessed 12 Dec. 2022.

28. Munira, Fayzulloeva. (2022). Ethical Ideas of the Ancient World. Annals of Bioethics & Clinical Applications. 5. 10.23880/abca-16000222.

29. Omrani, Nessrine & Rivieccio, Giorgia & Fiore, Ugo & Schiavone, Francesco & Garcia-Agreda, Sergio. (2022). To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. Technological Forecasting and Social Change. 181. 121763. 10.1016/j.techfore.2022.121763.

30. Palladino, Nicola. (2021). The role of epistemic communities in the «constitutionalization» of internet governance: The example of the European Commission High-Level Expert Group on Artificial Intelligence. Telecommunications Policy. 45. 102149. 10.1016/j.telpol.2021.102149.

31. Ralph, Jason. (2013). The liberal state in international society: Interpreting recent British foreign policy. International Relations. 28. 3-24. 10.1177/0047117813486822.

32. Rawat, Romil & Yadav, Rishika. (2021). Big Data: Big Data Analysis, Issues and Challenges and Technologies. IOP Conference Series: Materials Science and Engineering. 1022. 012014. 10.1088/1757-899X/1022/1/012014.

33. Schneider, S. (2019). Artificial You: AI and the Future of Your Mind. Princeton University Press. https://doi.org/10.2307/j.ctvfjd00r

34. Schultz, Robert A. "Political Realism and the Society of Societies." In Information Technology and the Ethics of Globalization: Transnational Issues and Implications. Hershey, PA: IGI Global, 2010. https://doi.org/10.4018/978-1-60566-922-9.ch006

35. Thanetsunthorn, Namporn. (2021). Organization development and cultural values of trust in international contexts. Review of International Business and Strategy. ahead-of-print. 10.1108/RIBS-04-2021-0054.

36. The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies // Salesforce URL: https://blog.salesforceairesearch.com/the-ai-economist/ (Accessed 11/14/2022).

37. Vakkuri, Ville & Kemell, Kai-Kristian & Jantunen, Marianna & Abrahamsson, Pekka. (2020). "This is Just a Prototype": How Ethics Are Ignored in Software Startup-Like Environments. 10.1007/978-3-030-49392-9_13.

Journal of Digital Economy Research. 2023. 1. P. 109–136
DOI: 10.24833/14511791-2023-1-109-136

Received: 25 September 2022,
Revised 31 October 2022,
Accepted: 20 December 2022

ARTIFICIAL INTELLIGENT SYSTEMS AND THE PROBLEM OF "NATURAL" TRUST

by E. Degteva, O. Kuksova

Abstract. The development of AI technologies has heightened the problem of humanitarian challenges at all levels of social regulations. Ethical issues and, in particular, the problem of trust have become relevant to the field of high technology, given the fact that AI performs increasingly significant managerial functions that previously could only be performed by humans. This issue is directly related to artificial intelligence systems, which have already been embodied in specific extensive projects. In this study, the authors analyze the concept of trust through the prism of technological development. For this purpose, the study presents an overview of historical and contemporary interpretations of the concept of trust and proves that this concept is relevant and necessary to control the risks that arise when integrating AI products into social life. The authors show that a rethinking of the concepts of ethics and morality in the new context is required. This is a necessary requirement for the creation of trusted AI and for the achievement of trust in human interaction with technology products. The authors conclude that it is necessary to build an interdisciplinary dialogue to integrate theory and practice from numerous fields. To do this, it is necessary to create a common knowledge base and a platform for communication between all stakeholders, but it is also important to create favorable conditions for sustainable and constructive interaction. Therefore, trust is a relevant concept that needs to be constructed in a multidimensional frame of reference that targets different stakeholders and also takes into account interaction between human and technology, in other words, at all levels and on all scales.

Keywords: AI, artificial intelligence, artificial intelligence system, ethics, moral, trust

For citation: Degteva E., Kuksova O.,(2023) Artificial Intelligent Systems and the Problem of "Natural" Trust. *Journal of Digital Economy Research,* vol. 1, no 1, pp. 109–136. (in English).
DOI: 10.24833/14511791-2023-1-109-136

Information about the authors:

Degteva Elena – Master of Economics. Deputy Director of the Department for the Development of the Social Sphere and the NPO Sector, Ministry of Economic Development of the Russian Federation.
degteva@list.ru
ORCID: 0000-0002-4137-6317

Kuksova Olga – Master of Economics. Deputy Head of Directorate, Directorate "Analytical Center for the Fuel and Energy Complex" FGBU "REA" of the Ministry of Energy of Russia.
129110, Moscow, st. Gilyarovskogo, d.39, building 1.
kuksovaki@gmail.com
ORCID: 0000-0001-8169-1517

Introduction

Recently, there has been a clear need to "fit" humanitarian values into digital technology products, in particular, within the framework of diverse projects based on AI, which are used in public administration, business, education and other areas. Numerous theorists, experts, practitioners, cultural and artistic figures are conducting a heated and lively discussion about how to identify possible risks and threats in time, who and how should develop and apply risk control tools and how to implement them into the management and public control system with in order to prepare and adapt society for even greater adoption of AI technologies. Today it is ethical issues that are being widely discussed and actively developed in the context of the development and application of AI in the political, economic and social spheres.

These ethical issues are inextricably linked with the actual problem of trust, which today requires attention at all levels: in international relations, in the economy, in the social sphere. The phenomenon of trust has become relevant in the field of high technologies. Moreover, the discussion about human interaction with technology is exacerbated in an environment where AI is beginning to perform more and more managerial functions that previously could only be done by Humans [20]. It deals not only with "tactical" projects, but also with technologies that are used for high-level management. Among such technologies, artificial societies or digital twins of society can be distinguished. Humanity is faced with the task of training neural networks to "understand" and "take into account" ethical values, or to clearly determine that such a result is impossible. This challenge places a huge responsibility not only on developers but to all stakeholders of AI products creation, introduction and usage.

At the same time, the stated problem has its roots in the distant past - both the question of trust as is, and the problem of trust with technology have a long history and complex multifaceted content.

Taking into account the fact that projects of artificial societies already exist, and their solutions are applied in practice to test management decisions, it is obvious that further development and use requires great attention, because such activities affect as society as well as an individual. We believe that trust is a key factor and requirement for testing the acceptability of using AI projects. This is especially relevant for significant projects that are aimed at solving political and social issues using AI and, in particular, artificial intelligence systems, and within which it is vitally important to ensure trust between all interested participants [19].

Literature review

Interaction of technological products with the anthroposphere have attracted the attention of thinkers since Antiquity. Renaissance and the New Age periods were no exception. At the beginning of the 20th century, leading thinkers were concerned with the humanitarian aspects of technological progress and the interaction between Human and technology [16]. We will not make a philosophical and historical essay, but will draw attention to the fact that the diverse aspects of innovative technologies have become the subject of philosophical reflection much earlier than the term "artificial intelligence" appeared. At the same time, the concept of AI is undoubtedly of revolutionary importance for culture, since it has combined the categories of Artificial and Natural into the most important category for Human - in the topic of Intellection. As is known, the term "artificial intelligence" was proposed in 1955 by D. McCarthy, and later M. Minsky defined it as a science whose goal is to implement intellectual human tasks with the help of a machine [14].

The AI problem has become fertile ground on which researchers began to test the possibilities of convergence of natural and artificial, biological and technological - what today seems to be the most promising areas for the development of science and practice.

Right at the time of the birth of the very concept of AI, there were formulated essentially contradictory approaches to its definition. In the 1950s and 1960s, two views stood out in the field of AI. One approach, proposed by D. McCarthy, used the approach of pure logic and the common mathematical basis. McCarthy believed that research in the field of AI would be fruitful when the very concept of intelligence (which still has a number of interpretations) is defined. In his opinion, the definition should form a coordinate system determined by epistemology and heuristics. That is, epistemology is how new knowledge is acquired in terms of its usefulness for solving problems, and heuristics is how these problems are solved: by what procedures and by what means [16].

M. Minsky proposed different approach, which later became known as connectionism (biologically inspired approach). He called for studying how the human brain and psyche work, and on this basis to implement knowledge in a computer model. This approach provides a basis for interpreting AI as something subjective, with individuality, consciousness, and even emotions. [7].

In general, today we can talk about the existence of 3 main approaches to the development of AI:

Artificial Intelligence (AI), or "Symbolical Artificial Intelligence", or "Classical Artificial Intelligence" is a top-down approach aimed at reproducing cognitive abilities without diving into the level of individual neurons.

Generalized artificial intelligence, or Strong artificial intelligence (AGI), is currently an assumed, hypothetical form of AI, in which the machine will receive equal human intelligence, and, accordingly, such AI will have self-awareness and freedom of choice

Biologically inspired artificial intelligence, or biologically inspired cognitive architectures, are projects that, based on the achievements of neuroscience, are aimed at creating artificial systems that reproduce the behavior and thinking of biological beings [12]. In particular, the development of the theory of artificial neural networks correlates and interacts with research on the functioning of the human brain [18, 23].

Thus, the first approach develops the approach of McCarty, and the second and third directions are based on the works of Minsky. We see that today there is no unified approach and no clear understanding of the directions of AI development and, accordingly, there is no unified approach to understanding this issue [26].However, the attempt at comprehension of this general topical topic within the framework of the philosophy of technology began in the early 20th century. One of the greatest thinkers of the last century, M. Heidegger, pointed to the threats posed by such a stage in the development of civilization, in which technology began to play a dominant role. The thinker noted that technology is not neutral in relation to Human, and that Human loses his subjectivity and ceases to determine the development of civilization. Technique itself begins to transform the natural world, and misunderstanding of the meaning of technology and its capabilities is fraught with a mistake, for which Human will pay with his enslavement [16, 22].

In our time, the horizon of interpretations and forecasts is very wide and multidirectional. It should be noted that fundamental thought has become the object of not only professional philosophers, but also of talented inventors, entrepreneurs and businessmen. In particular, the rationalist E. Yudkowsky believes in the possibility of creating a conscious AI, but according to his position first of all it is necessary to learn the secret of the intelligence [17].

Techno-optimist R. Kurzweil believes that AI will definitely be created. Provided that Human himself can preserve values and culture, AI will help to make the world a better place and ensure the prosperity of civilization [24].

Transhumanist N. Bostrom predicts that the creation of AI will lead to the next stage in the evolution of sapient life - the emergence of the Supersanity, which will get to control human existence [9]. A.V. Abramova researches the ethical aspects of the development of AI, studying and systematizing global trends in the development of ethics tools in the field of AI [1].

A.Yu. Alekseev leads the Artificial Personality Project, within which framework he explores the prospects of computer modeling of cultural, political, social, economic, moral and other aspects of public life and Human. The philosopher writes about the relevance of artificial life research as an important component of the development of nano- and biotechnologies in order to create the phenomenon of biological life using digital technologies. In his opinion, artificial life is a key component of artificial societies, which in turn serve as the basis for information and sociocultural technol-

ogies aimed at exploring the entire breadth of social life. This field opened the way for an artificial personality, to which researchers attribute not only consciousness and self-awareness, but also, in particular, freedom of choice and the ability for moral imputation - justifying or condemning one's own thoughts and actions [3].

The study of the features of human intelligence and AI is done by I.M. Dzyaloshinsky, who aims to prove the fundamental difference between Human and artificial intelligence [14]. It should be noted that in recent years not only scientists, but also publicists, journalists and writers have started to deal with this problematic. Numerous cultural figures reflect on the prospects of AI development and the risks and opportunities offered by the products of technologies based on AI. In particular, we can mention the books "Artificial You: Machine Intelligence and the Future of the Mind", written by S. Schneider, and "Our Final Invention: Artificial Intelligence and the End of the Human Era", written by J. Barrat [6, 32].

Recently, a number of publications have appeared devoted to the study of the potential dangers of the abuse of neural networks by technology owners and resistance to them, relying on trusted artificial intelligence. [5]..

It is the development of trusted artificial intelligence that represents the most promising concept of creating and understanding the idea of ethical AI today. However, the concept of "trusted" itself has already moved beyond theory and has begun to be used as a legal term in regulation. In particular, the concept of trusted AI is defined in the Ethical Guidelines for Trusted AI High Level Expert Group on Artificial Intelligence of the European Commission (Ethics guidelines for trustworthy AI, 2019) [29]. According to this document, trustworthy AI meets the following requirements, which at first glance relate more to human activities than to technology:

- Ethical (existing in accordance with ethical values and a code of ethics);
- Legal (not contrary to the law);
- Robust (reliable and sustainable).

In this context, it is necessary to reflect on the understanding of such a concept that from the first glance belongs mainly to the social field, as trust. This concept unites to a certain extent the theme presented and therefore some relevant approaches to the definition of this concept will be proposed in the following. It is trust that acts as a unifying concept and as a practical tool that will enable it to become a unifying link between human values and technological development [28].

This notion has been introduced into theory and social practice since the early stages of civilization, practically at the same time being introduced by Confucius in Ancient China and Pericles in Ancient Greece, describing relations based on trust as a necessary condition for prosperity of the state and society [27].

Subsequently, at different historical periods, depending on the ideology, development strategies and personality traits of the leaders and state ideologists, approaches to the definition and evaluation of the meaning of this concept have repeatedly changed in the most radical way.

Turning to the realities of the 20th century, we note that in the post-World War II period, the problems of trust and the opposite state - distrust in the context of the arms race and nuclear deterrence policy were among the key ones. The school of political realism, developed by H.Kissinger, T. Schelling and a number of other scholars and strategists, emerged in the US within the framework of this policy. This approach implies that the level of trust of political actors (in this case players on the international field) is low, and the stability (predictability) of partners' actions is determined by the balance of power and the system of checks and balances. Only such an approach can ensure predictability in the political game [33].

The neoliberal approach approaches the issue of complex interdependence as an opportunity through multiple channels and instruments to build trust between international partners and thereby ensure a high level of trust in foreign policy discourse [30].

S. Huntington noted that a special value of trust emerges in the periods of transformation and crises, and emphasizes the importance of institutional trust because it is the stable institutions that allow maintaining society and civilization in difficult times [34].

The postmodernist view essentially rejects trust, and J.-F. Lyotard highlights as a trend a global distrust of "metanarratives" that are supposed to justify reality. Z. Bauman points to the ephemerality, instability and unreliability of post-modernity, and argues that the center of gravity is the shift from state and international institutions to the global market with its consumerist values and the replacement of rational knowledge with shallow media exposure [25].

This last-mentioned idea leads us to the problem of technological development because it has enabled the globalization of the economy and culture, the development of mass media and mass culture, and subsequently the social and mobile web, all enabled by AI technologies and products. Thus, the breadth of the discussion, the scope of the problem and the horizon of opinions indicate that this issue is a complex, urgent and promising field of study.

Discussion

The technological basis of many of the categories that shape civilization today influences the transformation of values and the emergence of new ones, as well as the formation of development strategies.

Developing on the topic of artificial intelligent systems, we will try to define this phenomenon through its basic principles: such a system actively interacts with the outside world, perceiving and demonstrating (in accordance with the available properties and resources) a reaction to a directed impact. The degree of autonomy of such a system is the higher, the more independent the system is from information sources and control commands. The less the system depends on information sources and control commands, the higher the degree of its autonomy. Artificial intelligent systems are, in a certain sense, self-sufficient systems that can be entrusted with solving a certain set of applied problems [19].

In particular, the neural network created by Salesforce is testing and developing ideal tax system in a simulated environment [35]. So called AI Economist, it is based on reinforcement learning, which involves applying rewards and punishments to machine algorithms to maximize desired outcomes. By the same principle, for example, Google DeepMind AlphaGo and AlphaZero algorithms were created. The purpose of the experiment is to help governments around the world create more equitable system of taxation. The program is committed to creating a huge number of ecosystems with theoretical workers who trade currencies and build houses. Pay levels and skill sets vary, and AI determines optimal tax rates. The researchers note that this approach will reveal irrational behavior that economists often do not take into account in their models [2].

The relevance of such projects is increasing due to the fact that humanity is now progressing at an accelerated pace, and the unprecedented speed of change is causing global crises on a regular basis. That is why the notion of "black swans" has become popular, regularly emerging large-scale risks and threats that cannot be anticipated, detected or identified in advance because they originate from areas beyond knowledge and are largely (certainly, not entirely) driven by rapid technological development, whose achievements are being integrated into all areas of society by expanding the technosphere and blurring the lines between the humanitarian and technological domains [13].

In this context, classical tools of economic theory, strategic management and even anti-crisis management lose credibility and cease to bring results. To explore a world in which the technosphere has taken an important place, tools are needed that also include the capabilities of digital technology, in particular AI. In this sense, artificial intelligent systems are very promising, as this framework uses agent-based modelling. This method investigates the behavior of decentralized agents and how this behavior determines the behavior of the whole system. Use of agent-based modelling in artificial intelligent systems offers great opportunities not only to predict social developments under ordinary conditions, but also to predict possible crises and minimize the damage of such events. Of course, such an approach is a challenge, as it must take into account a large number of complex factors and be based on the results of multifaceted studies of real-world complex systems [8].

In this context, trust is an important requirement for implementation, as we are dealing with radical uncertainties. In public life, at the individual level, and indeed globally, uncertainty about the future is now coming to the fore, both on a practical, strategic scale and on a philosophical, literally metaphysical level. However, in this environment, AI-based dark projects may just be the solution, because classical tools cannot show effective results today. The pace of change, the emergence of new conditions, the acceleration of development requires all factors to be taken into account, and classical theories and methodologies are unable to account for important but unnoticed and uncalculated elements, such as irrationality and other features of human nature, and all the limitations and opportunities that follow from this.

Under conditions of radical uncertainty, trust is a requirement providing meaning and foundation for further development. Moreover, in a society where development is determined by self-created phenomena (phenomena that were not originally conceived but emerged on their own, in particular when the combined effect of the action of a set of individuals differs from that of individuals) and non-ergodic processes (for non-ergodic processes the probability observed in the past is not applicable to future processes), uncertainty becomes a major characteristic. At the same time, society increases uncertainty through its own incoherence, creativity, and it is under these conditions that artificial intelligent systems are required to function and, moreover, to provide confidence in its activities and in its results [8].Right today we can already predict some consequences of such intelligent systems performance, which have both positive and negative characteristics. On the one hand, these systems are aimed at ensuring security and well-being, on the other hand, the application of solutions of such systems in real life can affect natural values such as freedom, privacy and personal information of real people. And most importantly, it is necessary to make clear and set up who will play the dominant role in the new conditions and what balance will be in the management of social life between the State, technological corporations, society and the individual.

Also, along with the opportunities that intelligent systems provide, it is necessary to take into account the threats emanating from them that have social consequences.

As an example, there are problems in the field of profiling and aggregation of personal data. Social networks, entertainment services, which are intelligent systems, can accumulate large amounts of personal information about the life activities of each user, their personal preferences, views, health, family relationships and other information. In cases where the system has learned to recognize a user by their behavioral factors, identification can occur even when the user appears on affiliate services under a pseudonym or anonymously. Such data profiling may violate human rights and freedoms [11].

Aggregation of user data to improve cooperation with affiliate services is not in itself dangerous, but if sufficient personal information from various sources is accumulated, the system may become attractive to intruders for the purpose of creating a data breach. Aggregation services are therefore obliged to improve the security of all systems involved in a timely manner. The problem, however, is far from being solved globally at the moment. The issue is currently being addressed at the highest level. According to the UN Economic and Social Council, there is a need to clarify who owns the data, to ensure responsible use of data, privacy protection, accountability and fairness in social media, participatory platforms and online commerce portals. The main objective of these activities is to build trust and stability in the digital space, in particular in the field of AI [15].

Another example of the possible negative social consequences when using smart systems is the topic related to big data. Big data is supposed to be depersonalized, but the possibility of extracting personal data from depersonalized data streams (de-personalization) and deriving individual information from big data is a possible work in

progress. In this case, the individual is not protected against unauthorized use of personal information and, of course, the credibility of the intelligence systems is no longer guaranteed. The list of the social consequences could be continued since the presented challenges have no generally valid solutions at this time. That is why further we will talk about the ethical aspects of the functioning of such systems, about the extent to which their behavior can be conditioned by ethical paradigms, norms, and ideas [31].

It is necessary to find clear and generally accepted answers to the questions of what we should and can use intelligent systems for, what powers should be transferred to these systems, what risks this entails, and how they can be controlled. At the same time, we must understand that intelligent systems are already capable of influencing society today, and, noting the need to develop Code of Ethics for AI, we must understand that we are talking about systems that simulate planning, goal setting, choice and implementation of Human behavior [17]. It is necessary to timely analyze the processes taking place in public life, and on the basis of the analysis carried out, to introduce relevant changes in legislation and social norms in advance. This requires the integration of different areas of knowledge, methodologies and practices to make sense of these processes.

Integration is necessary both within the framework of scientific and research, as well as legislative and practical activities between the State, the academic community and the entire society, since this issue covers a much wider area compared to those that individual institutions, organizations, communities and researchers are able to work out.

Given the relevance and need to create and implement Code of Ethics that will become fully applicable to AI technologies, it should be noted that this is not enough[36]. Taking into account the fact that in recent years in the public space the concepts of ethics, morality (in general) and public morals have been often used as synonyms, and moreover, the scope of their application captures the areas of not only philosophy and culture, politics and law, but also technologies, economics and others, let us recall the difference between these terms. If, in the simplest sense, ethics defines and contrasts the concepts of "good" and "evil" in the philosophy of life and the worldview system, and morality regulates social relations through norms and principles [21]. Public morals are the real behavior and specific actions, as well as how this behavior is reflected in the actual experience of the subject, group and society. In other words, by public morals we mean objective assessments of practical actions. Therefore, the general narrative of ethics, according to our understanding, is at least an abstract discussion. It only introduces new simulacrum that threatens Human autonomy and stimulates the polarization of society and increases the degree of social tension. We believe that modern social trends form a single moral knot - the foundation on which the steps of moral progress should be built [15].

It is for this reason that there a lot of speculations around these concepts, and the price of this in relation to the use of artificial intelligent systems may be too high. The problem is that there is no single conceptual apparatus for basic definitions, there is no proper level of digital and humanitarian literacy, and there is often a bias or conflict of interest among interested participants. The task is to create new definitions for

relevant categories that will be multidimensional in nature, and will also allow us to consider phenomena in new conditions from different positions and within different assessment systems. We believe that it is urgent to move to a systematic solution of this problem. After all, the creation and implementation of AI systems in various spheres of life is becoming increasingly important every year, to the point that today in practice there are systems that function autonomously, which are so significant that they carry not only potential, but also real risk [37].

This proves that the study of the principles of ethical behavior of intelligent systems today has a multifaceted significance, both from a theoretical and practical point of view. Intelligent systems, as we noted, will be used more often and will affect society and the individual. Under these conditions, it is critical to ensure transparency and flexibility in creating a digital twin of society. It is undoubtedly important to ensure that the credibility of such projects is strengthened. In scientific and especially public discourse, one can find both fantastic and justified fears in connection with the risks of irresponsible use of AI technologies, up to the threat of loss of autonomy and freedom by Human [10]. At the same time, there is no generally accepted and clear plan for building trust in this area, however, the historical experience of introducing innovations shows that trust is usually born if technologies are useful and safe, and their regulation is transparent.

Speaking about the practical side of the ethical regulation of the artificial society model, it is necessary to solve the problem of formalizing ethics in a format accessible to AI technologies, as well as to implement ethical decision-making in the autonomous system itself based on AI restrictions in accordance with moral norms or by teaching AI to recognize ethical conflicts and accept their individual decisions [5]. This requires increased accountability in the development and use of these systems, as well as the greatest possible operational control in view of the serious consequences of making the wrong ethical choice.

The existing experience of systematizing the main ethical issues related to the creation and development of complex projects based on AI technologies allows us to highlight the problems of confidentiality and data protection, transparency, and, perhaps, as a common denominator, trust in predictive data.

Conclusion

The study allows us to conclude that today it is necessary to study technological and social development in a complex, and this global view will allow us to capture and find a solution to the problems caused by the development and spread of AI and products such as artificial intelligence systems in time.

The authors' research has analyzed the concept of trust through the prism of technological realities. The authors gave an overview of historical and current interpretations of the concept of trust and demonstrated that this concept is very promising and even essential as a tool to control the risks arising in the creation of digital technology products, which is particularly evident in the case of intelligent systems.

At the same time, humanitarian values today require a certain rethinking and new definitions that will be relevant to the new emerging reality. The authors show that the use of notions of ethics and morality without a qualitative revision of their meaning to meet modern requirements is untenable and is a formal intellectual exercise that will not really lead to the creation of trusted AI and to the formation of trust in human interaction with technology. The authors argue that this issue is a challenge because modern civilization in its current stage of development includes a wide range of stakeholders representing political, economic and social spheres, from the global community to the individual, and it is obvious that to build this multidirectional dialogue, a new language and new definitions of terms that define requirements and allow describing values should be formed. According to the authors, the contribution to the development of this issue lies in the fact that considerable research has been carried out into the interpretations of the relevant concepts, and the conclusions presented can be used for the next specific steps.

For a comprehensive understanding and management of ongoing processes, the authors suggest the following activities:

1. It is necessary to build an interdisciplinary dialogue to integrate theory and practice from numerous fields, taking into account the global level of the issues that we are exploring.
2. To do this, it is necessary to create a common knowledge base and a platform for communication between all stakeholders.
3. Under these conditions, sustainable and constructive interaction will become possible, in which new definitions of the concepts under discussion can be formulated and agreed upon in order to further integrate them in practice, both socially and scientifically.

Thus, the authors believe that in modern conditions trust is a category that needs and can be built in multiple dimensions - between the stakeholders, between human and technology, that is, at all levels and on all scales.

References

1. Abramova A.V. Ethics in the field of artificial intelligence - from discussion to scientific justification and practical application: analytical report / A.V. Abramova, A.G. Ignatiev, M.S. Panova; MGIMO MFA of Russia, Center for Artificial Intelligence; XIII Convention of the Russian Association for International Studies (RAMI) (Moscow, October 14–16, 2021). - M.: MGIMO-University, 2021. - 24 p.

2. AI Designed a Fairer Tax System // Hightech.fm URL: https://hightech.fm/2020/05/07/ai-economist (Accessed: 11/14/2022).

3. Alekseev A.Yu. Cognitive-technological projects of artificial personality // Man: Image and Essence. Humanitarian aspects. - 2014. - No. 1. - P. 156-174.

4. Anton, Eduard & Kus, Kevin & Teuteberg, Frank. (2021). Is Ethics Really Such a Big Deal? The Influence of Perceived Usefulness of AI-based Surveillance Technology on Ethical Decision-Making in Scenarios of Public Surveillance. 10.24251/HICSS.2021.261.

5. Avdoshin S.M., Pesotskaya E.Yu. Trusted artificial intelligence as a way of digital protection // Business Informatics. - 2022. - №2. - S. 62-73.

6. Barrat J. (2013). Our final invention: artificial intelligence and the end of the human era (First). Thomas Dunne Books.

7. Boldachev A.V. Essay: myths and dead ends of artificial intelligence // Philosophical problems of information technology and cyberspace. - No. 2 (18). - P. 27-41.

8. Bookstaber, Richard. (2017). The End of Theory: Financial Crises, the Failure of Economics, and the Sweep of Human Interaction. 10.1515/9781400884964.

9. Bostrom, N. Artificial Intelligence. Stages. Threats. Strategies. - St. Petersburg: Mann, Ivanov and Ferber, 2016. - 490 p.

10. Bulavinova M.P. Risks and threats of new technologies based on artificial intelligence. (review) // Social and humanitarian sciences. Domestic and foreign literature. Ser. 8, Science of Science: Abstract Journal. - 2018. - No. 2. - P. 23-41.

11. Comi, Antonio & Nuzzolo, Agostino. (2014). Individual pre-trip path choice modelling in transit advisor tools: theoretical and empirical evidences.

12. Dubrovsky, David. (2021). The Task of the Creation of Artificial General Intelligence and the Problem of Consciousness. Russian Journal of Philosophical Sciences. 64. 13-44. 10.30727/0235-1188-2021-64-1-13-44.

13. Dubrovsky, David. (2022). An Epistemological Analysis of the Social and Humanitarian Significance of Artificial Intelligence Innovations in Context of Artificial General Intelligence. Russian Journal of Philosophical Sciences. 65. 10-26. 10.30727/0235-1188-2022-65-1-10-26.

14. Dzyaloshinskiy I.M. Artificial Intelligence: Humanitarian Perspective // Bulletin of the Novosibirsk State University. Series: History, Philology. - 2022. - №6. – P. 20-29.

15. Ethics And 'Digital' - A Short Summary. Robot Doctor, Robot Teacher, Robot Policeman: Social Risks And Ethical Challenges For Industry: Policy Brief For Volume 2 Of The Report "Ethics And Digital: Ethical Challenges Of Digital Technology". - Moscow: RANEPA, 2020. - 45 c.

16. Finn V.: "Not all functions of natural intelligence can be formalized and automated" // Kommersant URL: https://www.kommersant.ru/doc/4198609?ysclid=lafntgvfcq626184900 (date of access: 11/14/2022).

17. Goertzel, Ben. (2015). Superintelligence: Fears, Promises and Potentials: Reflections on Bostrom's Superintelligence, Yudkowsky's From AI to Zombies, and Weaver and Veitas's "Open-Ended Intelligence". Journal of Ethics and Emerging Technologies. 25.55-87. 10.55613/jeet.v25i2.48.

18. Gurjar, Arunaben & Patel, Shitalben. (2022). Fundamental Categories of Artificial Neural Networks. 10.4018/978-1-6684-2408-7.ch001.

19. Gurov O. (2020). Ethical Interaction with Intellectual Systems. Artificial societies. vol. 15, no. 3 DOI: 10.18254/S207751800010905-4

20. Gurov O. (2021). Artificial Intelligent Systems and Social Problems Solving: Challenge of Trust. artificial societies. vol. 16, no. 1 DOI: 10.18254/S207751800014095-3

21. Gurov O.N. (2022) Metaverse—Flight From Dusk to Darkness? The Art and Science of Television, 18 (1), 11–46. https://doi.org/10.30628/1994-9529-2022-18.1-11-46

22. Heidegger M. Time and being: articles and speeches. - M.: Respublika, 1993. - 447 p.

23. Karpov V. E., Gotovtsev P. M., Roizenzon G. V. On the issue of ethics and systems of artificial intelligence // Philosophy and Society. - 2018. - No. 2 (87). - P. 84-105.

24. Kurzweil, R. The evolution of the mind. How expanding the capabilities of our mind will solve many of the world's problems. - M.: EKSMO, 2015. - 352 p.

25. McHale, Brian. (2013). Afterword: Reconstructing Postmodernism. Narrative. 21. 357-364. 10.1353/nar.2013.0020.

26. Muggleton, Stephen, and Nicholas Chater (eds), Human-Like Machine Intelligence (Oxford, 2021; online edn, Oxford Academic, 19 Aug. 2021), https://doi.org/10.1093/oso/9780198862536.001.0001, accessed 12 Dec. 2022.

27. Munira, Fayzulloeva. (2022). Ethical Ideas of the Ancient World. Annals of Bioethics & Clinical Applications. 5. 10.23880/abca-16000222.

28. Omrani, Nessrine & Rivieccio, Giorgia & Fiore, Ugo & Schiavone, Francesco & Garcia-Agreda, Sergio. (2022). To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. Technological Forecasting and Social Change. 181. 121763. 10.1016/j.techfore.2022.121763.

29. Palladino, Nicola. (2021). The role of epistemic communities in the «constitutionalization» of internet governance: The example of the European Commission High-Level Expert Group on Artificial Intelligence. Telecommunications Policy. 45. 102149. 10.1016/j.telpol.2021.102149.

30. Ralph, Jason. (2013). The liberal state in international society: Interpreting recent British foreign policy. International Relations. 28. 3-24. 10.1177/0047117813486822.

31. Rawat, Romil & Yadav, Rishika. (2021). Big Data: Big Data Analysis, Issues and Challenges and Technologies. IOP Conference Series: Materials Science and Engineering. 1022. 012014. 10.1088/1757-899X/1022/1/012014.

32. Schneider, S. (2019). Artificial You: AI and the Future of Your Mind. Princeton University Press. https://doi.org/10.2307/j.ctvfjd00r

33. Schultz, Robert A. "Political Realism and the Society of Societies." In Information Technology and the Ethics of Globalization: Transnational Issues and Implications. Hershey, PA: IGI Global, 2010. https://doi.org/10.4018/978-1-60566-922-9.ch006

34. Thanetsunthorn, Namporn. (2021). Organization development and cultural values of trust in international contexts. Review of International Business and Strategy. ahead-of-print. 10.1108/RIBS-04-2021-0054.

35. The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies // Salesforce URL: https://blog.salesforceairesearch.com/the-ai-economist/ (Accessed 11/14/2022).

36. Kopalle, Praveen & Gangwar, Manish & Kaplan, Andreas & Ramachandran, Divya & Reinartz, Werner & Rindfleisch, Aric. (2021). Examining Artificial Intelligence (AI) Technologies in Marketing Via a Global Lens: Current Trends and Future Research Opportunities. International Journal of Research in Marketing. 39. 10.1016/j.ijresmar.2021.11.002.

37. Vakkuri, Ville & Kemell, Kai-Kristian & Jantunen, Marianna & Abrahamsson, Pekka. (2020). "This is Just a Prototype": How Ethics Are Ignored in Software Startup-Like Environments. 10.1007/978-3-030-49392-9_13.

Исследования в цифровой экономике. 2023. 1. P. 137–176
DOI: 10.24833/14511791-2023-1-137-176

Получено: 8 октября 2022 г.,
Пересмотрено 10 ноября,
Принято: 3 декабря 2022 г.

ВЛИЯНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА НА РАБОЧЕЕ МЕСТО: ТЕКУЩЕЕ СОСТОЯНИЕ И БУДУЩИЕ ПЕРСПЕКТИВЫ

И.А. Ляпин

Аннотация. Стремительные технологические разработки в области технологий искусственного интеллекта и его многочисленных применений на рабочем месте вызвали в обществе множество опасений и предположений о возможной потере работы. Текущие исследования не сообщают о снижении спроса на высококвалифицированных работников, вызванных технологиями ИИ. Кроме того, нет никаких признаков того, что они заменят рабочие места в ближайшем будущем. Все больше компаний продолжают использовать и масштабировать ИИ практически во всех отраслях. Большинство предприятий получают отдачу от ИИ. Большая часть прогресса, достигнутого ИИ, связана с высокообразованными и квалифицированными профессиями. Для дальнейшего развития компаниям необходимо определить новые виды задач, которые необходимо выполнять, и распределить их между людьми и машинами. ИИ не только приведет к автоматизации, но и дополнит труд. Широкое внедрение технологий ИИ во всех отраслях по-прежнему сталкивается с проблемами. Часть ограничений носит технический характер, также существуют потенциальные проблемы с конфиденциальностью данных, злонамеренным использованием и безопасностью. Предприятия и органы власти должны использовать развитие ИИ и получать выгоду от повышения производительности и качества труда. Акцент должен быть сделан на обеспечение того, чтобы переход на новые технологии был плавным и проходил без осложнений.

Ключевые слова: Искусственный Интеллект, Рабочее Место, Рынок Труда, Профессиональное Влияние

Для цитирования: Ляпин И.А. (2023). Влияние искусственного интеллекта на рабочее место: текущее состояние и будущие перспективы. – *Исследования в цифровой экономике.* №1. С. 137–176. DOI: 10.24833/14511791-2023-1-137-176

Информация об авторе:

Ляпин Иван Алексеевич – МБА. Руководитель направления, ФАМ-Холдинг.
117405, г. Москва, ул. Дорожная, д.60 Б, 5-й этаж, оф. 516.
lyapin.ivan@gmail.com

Введение

Недавние технологические достижения в области искусственного интеллекта (ИИ) вызвали многочисленные дискуссии о крупномасштабной потере рабочих мест из-за его способности автоматизировать широкий и постоянно расширяющийся набор задач и его способности влиять на все сектора экономики.

Кроме того, варианты использования последнего десятилетия вызвали обеспокоенность по поводу благополучия сотрудников и о рабочей среде в более широком смысле, основанную на идее, что ИИ может вскоре стать широко используемым на рабочем месте и угрожать месту человека в нем. Однако у ИИ также есть возможность дополнить и расширить возможности человека, что приведет к повышению производительности, увеличению спроса на человеческий труд и повышению качества работы.

С одной стороны, динамичное развитие сопровождает технологические изменения и создает неопределенность в отношении будущего сферы труда. С другой стороны, многие авторы согласны с тем, что профессии лучше всего понимать как абстрактные наборы навыков и что технологии напрямую влияют на спрос на конкретные навыки, а не воздействуют сразу на все профессии.

В данной статье исследуются взгляды на текущий уровень внедрения технологий искусственного интеллекта, последние изменения в составе рабочей силы в связи с развитием технологий, анализируются отрасли, наиболее подверженные автоматизации рабочей среды с использованием новых технологий. Затем в нем представлен обзор возможных изменений требований к рабочему месту и рабочей силе, обобщены проблемы быстрого внедрения и даны рекомендации по решению проблем.

Настоящие изменения в составе рабочей силы

В последнее время ИИ добился наиболее значительного прогресса в профессиях, связанных с нерутинными и познавательными задачами, которые обычно выполняются работниками средней и высокой квалификации. Рабочие этого типа полагаются на способности, которые ИИ в настоящее время не может предоставить, такие как индуктивное мышление или социальный интеллект. Кроме того, образованным работникам легче внедрять и изучать новые технологии и, как правило, чаще участвовать в обучении для повышения своего профессионального уровня, что позволяет им лучше использовать возможности, предоставляемые ИИ на рынке труда.

Проведенные в 2022 году исследования изменений состава рабочей силы и рабочего процесса, сопровождающие инвестиции в ИИ, дают представление о том, как ИИ может преобразовать рабочее место, производственные процессы и организацию [6]. Как технология, применимая в основном для работы с когнитивными способностями, ИИ улучшает способность человека принимать решения, что, в свою очередь, повышает автономность рабочей силы в прогнозировании и снижает потребность в управленческом персонале.

По сравнению с предыдущими технологиями автоматизации, которые способствовали повышению производительности, развитие технологий ИИ не связано со снижением потребности в высококвалифицированных работниках на организационном уровне.

Исследования также помогают сделать выводы о влиянии на рынок труда. ИИ ассоциируется с высококвалифицированным трудом, и нет никаких доказательств того, что он вытеснит задачи, которые, как обычно прогнозируется, будут им вытеснены, такие как рутинные задачи по управлению персоналом, юридические, административные. Далее мы прольем свет на то, как ИИ влияет на рабочие места в компаниях, не использующих ИИ, и на влияние на рынок труда в целом.

Другие изученные нами исследования показывают существенное увеличение спроса на навыки, связанные с ИИ [5]. Количество вакансий, требующих навыков ИИ, увеличилось в десять раз с 2010 по 2019 год и в четыре раза по отношению ко всем вакансиям.

Большая часть объявлений о вакансиях сосредоточена в секторах информационных технологий, профессиональных услуг, финансов и производства, но рост спроса наблюдается и в других отраслях. Точно так же потребность в навыках ИИ вытекает из профессий, связанных с компьютерами, инженерией и наукой, а также из таких профессий, как сельское хозяйство, лесное хозяйство и рыболовство, которые в начале наблюдения имели почти нулевой спрос на навыки ИИ.

В изученных объявлениях о вакансиях с сопоставимыми другими навыками добавление требований к навыкам ИИ увеличивает предлагаемую заработную плату на 16%. С учетом постоянных премий в некоторых отдельных фирмах, надбавка к заработной плате ИИ может составлять почти 20%. Эта надбавка значительно выше, чем надбавка за другие навыки анализа.

Также были обнаружены значительные надбавки к заработной плате для профессий, не связанных с ИИ, в фирмах с более высокой интенсивностью инвестиций в ИИ. Результаты показывают, что рабочие места, требующие навыков работы с программным обеспечением, когнитивных, социальных навыков, управления проектами и управления людьми, дополняют рабочие места ИИ, в то время как рабочие места, связанные с клиентским сервисом, могут быть заменены ИИ.

Недавние исследования [1] более формально сформулировали традиционное объяснение того, почему совершенствование технологий и методов экономии труда не приводит к сокращению рабочих мест: производительность увеличивает спрос на топливо в экономике в целом, что, в свою очередь, создает больше рабочих мест, хотя, как правило, в других областях, а не в тех, где происходят первоначальные улучшения производительности. Существует множество путей, по которым может осуществляться связь между повышением производительности и спросом.

Текущие примеры использования ИИ на рабочем месте

Внедрение ИИ растет практически во всех отраслях, но его возможности отличаются.

Исследование и опрос, проведенные McKinsey & Company, показывают, что организации внедрили по крайней мере одну из возможностей ИИ в продуктовый процесс по крайней мере в одной функции или бизнес-подразделении. Ответы также показывают увеличение доли фирм, использующих ИИ в продуктах или процессах во многих функциях [22]. Это свидетельствует о том, что все больше компаний используют и масштабируют ИИ. Компании, интенсивно внедряющие ИИ, продвинулись в этих усилиях гораздо дальше.

По секторам результаты показывают рост внедрения ИИ почти во всех отраслях в последние годы. Среди лидеров роста розничная торговля является наиболее эффективным сектором, где компании внедрили хотя бы одну возможность искусственного интеллекта в одной или нескольких функциях или бизнес-подразделениях.

По функциям мы находим большинство вариантов использования в операциях (RPA, профилактическое обслуживание, производственная аналитика, инвентаризация и цепочка поставок, робототехника, выставление счетов), общих решениях (аналитика, автоматизированное машинное обучение) и специализированных решениях (геоаналитика, разговорная аналитика, распознавание изображений, аналитика электронной коммерции), маркетинг (персонализированный и контекстный маркетинг, аналитика), продажи (прогнозирование, лидогенерация, автоматизация ввода данных, оценка лидов, обучение агентов, аналитика продаж), клиентский сервис (социальное прослушивание, продажа билетов, маршрутизация звонков, голосовая аутентификация, ответы, чат-боты, аналитика звонков, опросов и обзоров), HR (найм, управление производительностью, управление удержанием, мониторинг сотрудников, управление зданием, HR-аналитика). Есть также множество вариантов использования в области технологий, обработки данных и других инновационных функций.

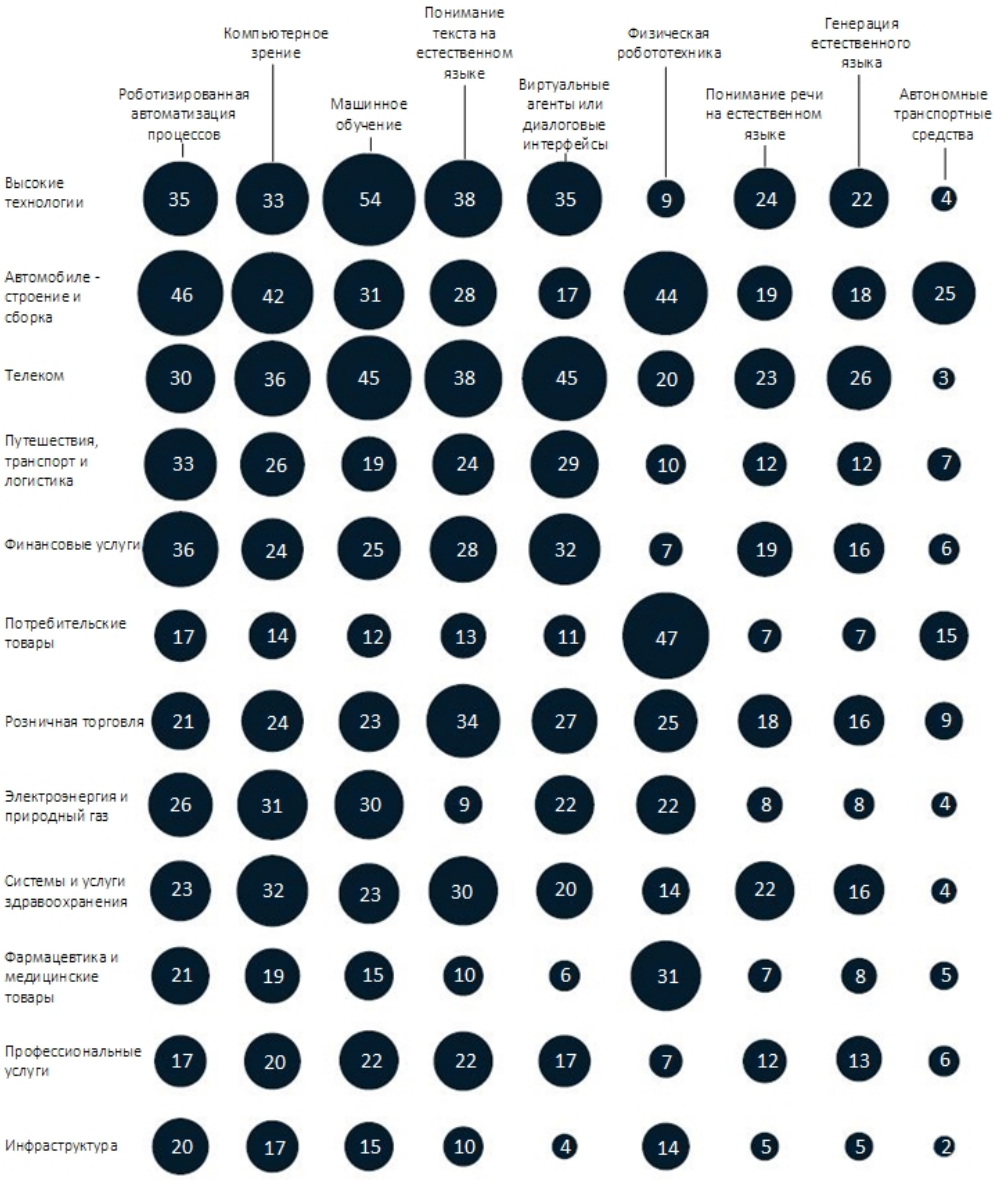

Рис. 1. Отрасли, использующие возможности ИИ.

Мы видим, что фирмы применяют возможности ИИ в функциях, которые помогают им получать прибыль в своих отраслях. Например, респонденты из компаний, производящих потребительские товары, чаще сообщают об использовании физической робототехники, которая может помочь в сборке, чем о большинстве других типов возможностей. Респонденты из сферы телекоммуни-

каций сообщают, что их компании используют виртуальные агенты, которые могут использоваться в приложениях для обслуживания клиентов чаще, чем другие возможности (рис. 1). Другие высокоэффективные компании сообщают, что внедрили ИИ в продажи и маркетинг.

На региональном уровне результаты опроса показывают значительное увеличение уровня внедрения в развитых странах Азиатско-Тихоокеанского региона, Европы, Латинской Америки и Северной Америки. Высокий уровень внедрения ИИ ставит эти регионы вместе с Китаем на один уровень внедрения, предполагая, что ИИ является глобальным явлением на рабочем месте, в то время как интенсивность на уровне компании может значительно различаться.

Самые эффективные компании, внедряющие ИИ, используют практики извлечения ценности.

Согласно нескольким исследованиям и работам, большая часть деятельности фирм необходима для получения ценности и масштабировании её. Во многих случаях это аналитика и ее согласование с бизнесом, и ИТ-лидеры фокусируются на возможности извлечения ценности из ИИ в каждой области бизнеса. Мероприятия включают, помимо прочего, инвестиции в таланты и обеспечение того, чтобы бизнес-персонал и технический персонал обладали навыками для успешного масштабирования. Изученный опрос предполагает, что эти методы необходимы для масштабирования ИИ, но необходимо иметь в виду, что респонденты опроса чаще говорят, что их компании уже используют такие методы.

Тем не менее, даже высокоэффективные компании, применяющие ИИ говорят, что их передовые сотрудники используют идеи ИИ для повседневных задач при принятии решений и менее половины систематически отслеживают четко определенный набор ключевых показателей эффективности для ИИ, которые критичны для достижения их принятия конечным пользователем и привнесения ценности. Точно так же только 35% респондентов из числа высокоэффективных компаний с ИИ сообщают, что у них есть активная программа непрерывного обучения сотрудников технологиям ИИ [22].

Степень воздействия ИИ на профессии

Многие компании по всему миру в настоящее время внедряют ИИ на рабочем месте, а также вкладывают значительные средства в его внедрение и адаптацию, что делает ИИ частью почти каждого этапа работы сотрудников. Он широко используется в рекрутинге и адаптации сотрудников, в развитии талантов и управлении процессами.

ИИ становится основным инструментом либо для полной автоматизации рабочего процесса, либо для улучшения и помощи работникам в выполнении их задач с использованием робототехники, автоматизации процессов и других технологий.

Деятельность ИИ добилась наиболее значительного прогресса в отношении профессий высокообразованных белых воротничков. Таким образом, высококвалифицированные профессии относятся к числу профессий, в которых технологии ИИ наиболее широко используются. Это специалисты в области науки и техники, бизнеса и управления, менеджеры, руководители, люди, работающие в юридической, социальной и культурной сферах.

С другой стороны, профессии, где воздействие является самым низким, представляют профессии, в которых люди вовлечены в физическую деятельность. Например: уборщики, сельскохозяйственные рабочие, работники лесного и рыбного хозяйства, работники по приготовлению пищи. Эти профессии полагаются на способности, в которых ИИ пока не реализован.

Исследование показывает, как воздействие ИИ распределяется по разным странам. В большинстве случаев воздействие ИИ на профессию невелико, фундаментальные различия между профессиями обычно выше [21]. Мы можем заметить небольшую разницу в профессиональном воздействии ИИ между проанализированными странами Балтии и Северной Европы. Кроме того, изучая экспозицию между профессиями, мы можем заметить, что распределение по профессиям в наиболее подверженных и наименее подверженных воздействию странах очень похоже. Экспозиция будет по-прежнему различаться по регионам и отраслям из-за различий во многих социально-экономических факторах, разных рынках труда и их динамике: в развитых экономиках с достаточно высоким уровнем заработной платы экспозиция может быть более чем в два раза выше, чем в таких странах, как Индия.

Технологии искусственного интеллекта быстро распространяются по всему миру и, как ожидается, окажут большое влияние на мировую экономику. Давая эффект, который мы можем предполагать; стало критически важно иметь данные, которые позволяют нам изучать возможные последствия. Большая часть работы предстоит сделать, чтобы выявить возможное влияние ИИ на труд. Поэтому становится все более важным иметь возможность работать с проверенными и надежными глобальными данными.

Драйверы для внедрения ИИ на рабочем месте

Большинство предприятий могут увидеть отдачу от ИИ.

Со ссылкой на исследование, проведенное McKinsey & Company, в котором было изучено около 33 случаев внедрения ИИ в восемь бизнес-функций. Они также изучили влияние внедрения ИИ в каждом из этих видов деятельности на доходы и расходы в отделах, использующих ИИ [22]. Настоящим мы приводим результаты, которые пришли к выводу, что ИИ представляет значительную ценность для компаний.

Значительная доля респондентов сообщает, что внедрение технологий ИИ в компаниях увеличило выручку более чем на 10 процентов. Наибольшая прибыль получена от ИИ, используемого в продажах и маркетинге (ценовая политика, вероятность совершения покупки), управлении цепочками поставок (прогнозирование продаж и спроса), разработке продуктов и услуг (разработка новых для рынка продуктов на основе ИИ). В своем отчете Accenture оценивает, что фирмы, интенсивно инвестирующие в технологии на основе ИИ, могут увеличить выручку от продаж на 38 процентов [37]. Также есть возможность повысить уровень занятости на 10 процентов.

Кроме того, около половины участников сообщают о снижении затрат в результате внедрения ИИ. Большинство случаев экономии затрат можно найти в производстве (оптимизация производства, энергии и общей производительности) и управлении цепочками поставок (оптимизация цепочек поставок).

ИИ предоставляет аналитику, которая помогает людям принимать правильные решения.

Преимущества нового сотрудничества человека и ИИ становятся очевидными. Хорошим примером является недавний алгоритм на основе искусственного интеллекта, разработанный группой ученых из Гарварда и предназначенный для более точного выявления клеток рака молочной железы. Когда мы сравниваем работу человека с работой ИИ, патологоанатомы идентифицируют болезнь с точностью 96 процентов, в то время как только алгоритм делает это в 92 процентах случаев. Но когда работу выполняет альянс человека и ИИ, точность диагностики возрастает до 99,5 случаев выявления рака в биоптатах.

Такой успешный вариант использования может быть создан благодаря постоянному сотрудничеству и обучению человека и ИИ, чтобы помочь повысить его эффективность.

Экономика и общество в целом получают больше возможностей от интенсификации ИИ.

Помимо ценности для бизнеса, как уже было сказано выше, технологии ИИ оказывают влияние на экономику, поддерживая ее рост, и могут обеспечить значительный прогресс в решении большинства социальных проблем. Это может влиять на продолжительность жизни и способствовать повышению рождаемости, производительности труда, что является одним из основных факторов экономического подъема. ИИ также находит применение во многих других областях: медицинские исследования, изучение климата, материаловедение и т. д.

Тренды и будущие перспективы применения ИИ на рабочем месте

Можно с уверенностью сказать, что будущее неопределенно не только в отношении рабочего места, но практически в любой сфере жизни. И также трудно быть уверенным, что делать с будущим. Но нам важно попытаться его предугадать. Что касается экономических перспектив рабочей силы, то имеющийся у нас

опыт не позволяет предсказать, какие профессии будут востребованы в отдаленном будущем. Даже если бы у нас было разумное понимание того, какие рабочие места будут упразднены в будущем, мы должны быть достаточно уверены, какие рабочие места будут востребованы тогда же.

Чтобы сделать прогнозирование более эффективным и точным, возможно, имеет смысл планировать будущее на уровне отдельной компании, а не экономики в целом. Из предыдущих абзацев мы знаем, что фирмы со стратегиями и инвестициями, ориентированными на ИИ, лучше приспособлены и подготовлены к адаптации новых технологий и лучше используют новые подходы. Кроме того, если мы делаем предположение, что технологии на основе ИИ могут снизить спрос на рабочие места в будущем, гораздо эффективнее будет сделать программу развития на уровне отдельного работодателя, поскольку региональные или глобальные прогнозы могут быть использованы только небольшой группой сотрудников в любой момент. Кроме того, в случае если человек хочет перейти от одной роли к другой, это намного проще сделать в рамках одной организации, где его специфические для компании навыки и знания остаются актуальными и востребованными.

Чтобы решить проблему неопределенности, прежде всего, мы должны признать, что даже предположительно хорошие модели прогнозирования дают нам наиболее вероятные результаты или оценки наиболее вероятного решения нашей проблемы. Практический опыт показал, что во многих случаях даже самый очевидный исход маловероятен, поэтому важно прогнозировать второе и даже третье наиболее вероятное решение. Одним из возможных способов решения этих проблем является сценарное планирование. Другим является моделирование, где мы можем создать модель прогнозирования и увидеть возможные результаты, изменив значения исходных переменных.

Без сомнения, компании должны быть готовы к серьезным изменениям в работе. Можно было бы автоматизировать около пятидесяти процентов действий (не самих рабочих мест), выполняемых рабочими и помощниками. Исследование 2 тыс. видов трудовой деятельности по 800 профессиям дает данные об отдельных видах деятельности, которые легче автоматизировать [3]. Это физически выполняемые действия в фиксированной и стабильной среде, сбор и обработка данных. Их можно приблизительно оценить как половину всей деятельности, которую люди выполняют во всех отраслях. Наименее подверженные опасности действия включают принятие решений, предоставление экспертных знаний, общение с заинтересованными сторонами. Учитывая текущий уровень технологий, полностью автоматизировать можно только 5 процентов существующих на данный момент профессий, но почти все они будут затронуты в будущем.

Может ли ИИ помочь решить проблемы рынка труда? Большинство компаний, использующих технологии ИИ на рабочем месте, склонны так думать. Они либо увеличивают интенсивность использования ИИ, чтобы уменьшить общие потребности в найме, либо разрабатывают план его внедрения. Как уже упоми-

налось, у ИИ есть своя проблема для рынка труда: необходимость нанимать квалифицированных и часто дорогих специалистов по ИИ. Есть ряд случаев, когда компании замедляли процесс найма из-за ограниченного количества талантов из ИИ. Несмотря на то, что спрос на специалистов по ИИ и науке о данных постоянно растет, есть способы смягчить замедление набора персонала.

Ниже приведены наши рекомендации о том, как компании могут подготовить свой бизнес к будущему.

Предприятиям следует пересмотреть задачи, которые необходимо выполнять.

Рекомендуется переоценить набор человеческих команд и технологий, доступных для распределения задач и действий между людьми или роботами. Это не разовая задача, процесс такого распределения должен быть непрерывным для повышения его эффективности. Большинство существующих ИИ не способны работать автономно и требуют помощи человека для вмешательства, ввода данных, внесения необходимых корректировок в их работу. Во многих случаях, когда ИИ должен принимать решения, его должны обучать люди, чтобы результаты их работы были приемлемыми и полезными для дальнейшего использования.

ИИ не только приведет к автоматизации, но и дополнит труд. В качестве другого примера сотрудничества человека и ИИ мы можем представить случай, когда аналитику необходимо написать отчет на основе набора данных, предоставленного и собранного ИИ (высокая пригодность для машинного обучения), но затем предоставить в нем рекомендации и указания, которые потребуют абстрактных рассуждений, который должен восприниматься как исходящий от человека. Ожидается, что ИИ повысит производительность не только за счет того, что позволит компаниям заменить рабочую силу более дешевым капиталом, но и дополнит работников, позволяя им повысить свою производительность за счет использования своих навыков социального взаимодействия и способности рассуждать о новых ситуациях.

Компаниям необходимо внедрить управление ИИ.

Чтобы быть в курсе постоянно развивающихся моделей ИИ, компаниям необходимо развернуть модели управления, специально разработанные для приложений ИИ, которые могут повлиять на все соответствующие заинтересованные стороны. Управление должно включать управление рисками, ИИ и бизнес-лидеров с новыми политиками и процедурами, ролями и обязанностями. Примером возможной модели управления является модель трех линий, которая была разработана в 2008–2010 годах Федерацией европейских ассоциаций по управлению рисками и принята Институтом внутренних аудиторов в 2013 году. Модель описывает три линии защиты, которые необходимо установить в рамках управления ИИ:

- Создатели, исполнители и операции.
- Менеджеры, супервайзеры и обеспечение качества.
- Аудиторы и специалисты по этике.

В дополнение к трехлинейной модели также рекомендуется использовать сквозную модель управления. Он включает в себя все этапы жизненного цикла организации.

Кроме того, несмотря на возможность использовать и улучшать большую часть существующих средств управления и контроля ИТ, очень важно, чтобы бизнес-лидеры изучали и понимали некоторые основы ИИ и науки о данных.

Компании должны понимать свои уникальные уязвимости.

Компании, работающие в разных секторах экономики, таких как финансы, розничная торговля, коммунальные услуги и т. д., сталкиваются с различными рисками от потенциального внедрения ИИ: когда он может вмешаться в наборы данных и нанести серьезный ущерб. Для компаний важно определить свои уникальные уязвимости и определить риск потенциального использования их конкретных алгоритмов искусственного интеллекта. Для принятия решения о внедрении ИИ и расстановки приоритетов усилий эффективно оценить величину возможных возникающих финансовых, операционных и репутационных рисков и ценность их снижения.

Компании должны контролировать свои данные.

Скорее всего, традиционные способы контроля над обработкой данных недостаточно надежны для обнаружения конкретных проблем, которые могут вызвать риск, связанный с использованием ИИ. Важно уделять особое внимание проблемам с историческими данными и данными, полученными от третьих лиц. При реализации алгоритмы ИИ могут расползаться по наборам данных и позволять принимать предвзятые решения, поэтому они должны быть хорошо спроектированы и управляться с самого начала.

Компании должны фокусироваться на сотрудниках.

Кратчайший способ внедрить ИИ на рабочем месте — решить проблему нехватки ИИ-специалистов. Важно брать и нанимать новых специалистов, которые уже обладают нужными вам навыками или могут легко их приобрести. Компании должны рассмотреть возможность преподавания науки о данных, машинного обучения, разработки программного обеспечения, статистики и бизнес-лидеров из этих областей. Такой подход также может помочь заполнить пробелы в общении и улучшить сотрудничество между группами, работающими с ИИ.

Чтобы снизить возможные риски предвзятого ИИ на рабочем месте, компаниям необходимо разнообразить свои команды людьми разного происхождения, пола, расы. Разнообразная команда, объединяющая бизнес-лидеров, специалистов по данным, инженеров и других специалистов с разным опытом, позволит по-разному взглянуть на возможные угрозы и способы их смягчения.

Внедряя ИИ, компании смогут уменьшить потребность в рутинной работе и сделать жизнь своих сотрудников проще и интереснее. Объявляя о планах и запуская процесс внедрения ИИ, компании могут увеличить ценность, которую могут дать им сотрудники. ИИ может даже помочь оказать эмоциональную поддержку на рабочем месте, сделав сотрудников счастливее.

Компании могут оценивать и прогнозировать рентабельность инвестиций в ИИ.

Компаниям не нужно внедрять модели ИИ, которые не фокусируются на важных бизнес-задачах. Чем сложнее задача, тем важнее, чтобы бизнес стремился к правильному решению проблемы. Трудно количественно измерить рентабельность инвестиций ИИ, которая часто инициируется стратегическими решениями или новейшими технологиями, которых никогда не было на рынке. Однако в настоящее время предприятия с большей вероятностью будут измерять значение воздействия, используя новые методы оценки. Они могут охватывать не только «твердые» доходы, такие как повышение производительности, но и «твердые» затраты, такие как расходы на новое оборудование. Они также могут отражать «мягкие» доходы, такие как улучшение условий труда сотрудников, и «мягкие» затраты, такие как повышенная потребность во времени специалистов в данной области. Целостный подход к ИИ, множество существующих вариантов использования и практик также помогают оценить рентабельность инвестиций в новые инициативы. Также можно использовать сам ИИ для моделирования воздействия и неопределенностей инициатив ИИ и помощи в более эффективном распределении ресурсов.

При моделировании важно убедиться, что все различные подходы и возможные изменения проверены, а их влияние измерено.

Также может быть полезно рассмотреть портфельный подход, чтобы избежать неожиданностей окупаемости инвестиций, поскольку компании могут также работать с инновациями продуктов: создавать и оценивать набор инициатив, которые повысят вероятность получения общих результатов, необходимых компаниям.

И закончим с того, с чего начали: для эффективного и точного прогнозирования рентабельность инвестиций в ИИ, компаниям необходимо управлять и регулировать не отдельные проекты, а полный жизненный цикл. Это может помочь им постоянно развивать стратегию, точно настраивать исполнение и находить новые варианты использования данных — как избегая неожиданных побочных эффектов, так и находя новую ценность.

Влияние на занятость

Несмотря на то, что ИИ считается новой технологией, он уже вызывает высокий спрос на соответствующие таланты во многих отраслях [36]. Он уже здесь, чтобы остаться во многих секторах экономики. Этот спрос предсказуемо высок в секторах научно-технических услуг, он также показывает высокие уровни в производстве, финансах и страховании, администрировании. Информационный сектор — это сектор, который люди намереваются использовать в первую очередь, когда думают об ИИ. Однако есть немного эмпирических данных о том, как ИИ влияет на рынок труда на сегодняшний день и изменит его в будущем.

Потерянные рабочие места: к 2030 году ожидается значительный спад в некоторых профессиях.

Быстро развивающиеся технологии искусственного интеллекта вытеснят некоторых людей. Исследования показывают, что около 15 процентов всех работающих, или около 400 миллионов человек, могут потерять свои рабочие места из-за ИИ в период 2016–2030 годов [3].

Техническая возможность автоматизации рабочих мест является основной причиной спада. Есть и другие причины, такие как стоимость разработки, изменения на рынке труда, вызванные количеством и качеством предложения рабочей силы и соответствующей заработной платой, выгоды от перемещения людей, которые упоминаются в вариантах использования внедрения, и, что не менее важно, принятие общество и его нормы.

Полученные рабочие места: за тот же период будет создано больше рабочих мест.

Несмотря на то, что многие рабочие места, как ожидается, будут смещены, прогнозируется растущий спрос на работу, что, в свою очередь, создаст больше рабочих мест. Этому утверждению способствуют такие факторы, как более высокие расходы на социальные нужды, растущие доходы, ожидаемые новые инвестиции в значимые проекты в энергетике, инфраструктуре, развитие новых технологий и их практическое применение. Основными бенефициарами этих изменений станут страны с развивающейся экономикой, такие как многочисленные азиатские страны, где население трудоспособного возраста резко увеличивается. Такой рост экономики и населения повлияет на динамику бизнеса и повлияет на рост производительности, а также будет способствовать увеличению числа новых рабочих мест на рынке.

Измененные рабочие места: ожидается, что рабочие места изменятся, поскольку технологии будут взаимодействовать с людьми на рабочем месте.

Будет очевидна выгода от дополнения человеческого труда машинами, и этот фактор приведет к преобладанию частичной автоматизации. Работа с повторяющимися задачами перенесет текущую работу на устранение неполадок и обслуживание автоматизированных систем.

Проблемы внедрения ИИ и их смягчение

Широкое внедрение технологий ИИ во всех отраслях по-прежнему сталкивается с проблемами. Часть ограничений связана с техническими аспектами, такими как необходимость значительного обучения данных и трудности с применением одних и тех же алгоритмов в разных приложениях. Трудно определить все проблемы, связанные с последними инновациями. Потенциальные проблемы с обучающими данными, оптимизация алгоритмов, защита личных данных, мошенническое использование и безопасность — вот основные проблемы, на которые следует обратить внимание.

Тем не менее, мы можем выделить основные проблемы, с которыми компании сталкиваются сейчас. Ожидается, что изменится набор профессий, а также требования к рабочим навыкам и образованию. Рабочую деятельность необходимо будет пересмотреть, чтобы обеспечить наиболее эффективное сотрудничество людей и ИИ.

Ожидается, что больше внимания будет уделяться обучению новым навыкам.

Результаты исследования показывают, что большинство фирм готовятся к кадровым изменениям, связанным с ИИ [22]. Около 60% респондентов заявили, что в их компаниях, внедривших ИИ, работники прошли переподготовку за последний год. Кроме того, 83% участников исследования планируют переобучить некоторых своих сотрудников в ближайшие три года из-за внедрения ИИ, а 38% планируют переобучить более четверти персонала.

Технологии искусственного интеллекта на рабочем месте увеличат скорость смены требуемых навыков. Требования к техническим навыкам будут быстро расти. Также будет расти спрос на когнитивные навыки, такие как анализ информации, критическое мышление, креативность. Это окажет большее давление на существующую рабочую силу и фирмы, чтобы они разрабатывали решения, которые масштабируют их навыки вместе с прогрессом.

Некоторым работникам, вероятно, придется пересмотреть род занятий.

Требования к изменению и развитию новых рабочих навыков могут привести к необходимости изменения категорий занятий. Некоторые из этих изменений ожидаются в разных секторах и компаниях, но мы также должны ожидать, что многие из них произойдут внутри секторов и географических регионов. Профессии, связанные с обработкой данных и высокоструктурированной средой, будут сокращаться, в то время как профессии, связанные с непредсказуемой средой, будут удовлетворять растущий спрос на работу. Некоторые другие востребованные профессии будут включать учителей, менеджеров, технических специалистов и других специалистов.

Ожидается, что автоматизация повлияет на заработную плату в странах с развитой экономикой.

Изменение профессиональной деятельности, вероятно, окажет давление на заработную плату. Искусственный интеллект, скорее всего, заменит высокоавтоматизированные виды деятельности, и произойдет переход к более квалифицированным рабочим местам. Количество высокооплачиваемых рабочих мест значительно увеличится, что приведет к тому, что автоматизация усилит поляризацию заработной платы и неравенство доходов.

Многие страны уже сталкиваются с проблемой надлежащего обучения и переподготовки своих работников, чтобы они могли соответствовать существующим требованиям работодателей [2]. В странах ОЭСР расходы на образование и обучение рабочей силы снижались в течение последних двух десятилетий. Расходы на перевод рабочих и помощь в переселении также продолжали снижаться как доля ВВП.

В поисках решения проблем мы не должны тормозить непрерывную разработку и применение новых технологий. Компании и правительства должны извлечь выгоду из разработок в области технологий искусственного интеллекта, которые приносят не только социальные выгоды, но и повышают производительность и производительность труда. Акцент должен быть сделан на обеспечение того, чтобы переход на новые технологии был плавным и беззаботным. Автоматизация и рост производительности вносят основной вклад в создание рабочих мест и экономическое развитие. Вот почему поддержка инвестиций и развитие новых технологий искусственного интеллекта имеет решающее значение.

Обращаясь ко всем вызовам, важно не упустить важное наблюдение: технологии искусственного интеллекта будут влиять не столько на количество рабочих мест, сколько на требования к работе и задачи.

Ниже приведен список с нашими рекомендациями по направлениям, на которых следует сосредоточить основные усилия.

Стимулирование динамизма бизнеса.

Возможность дальнейшего развития предпринимательства приведет к формированию нового бизнеса, который, в свою очередь, создаст новые рабочие места и повысит производительность труда. Динамика бизнеса зависит от открытой и понятной среды для малого бизнеса и здоровой конкурентной среды для крупного бизнеса. Для ускорения формирования нового бизнеса и повышения конкурентоспособности крупных и малых предприятий потребуются более простые и совершенные правила, налоговые и другие стимулы.

Постоянно совершенствующаяся система образования и доступные тренинги для смены места работы.

Бизнес вместе с властями мог бы внести больший вклад в развитие и развивать базовые и продвинутые навыки через школьные образовательные системы и корпоративные тренинги. Требуется новый акцент на творчестве, критическом и системном мышлении, а также на адаптивном обучении на протяжении всей жизни.

Инвестиции в человеческий капитал.

Как было сказано выше, большое внимание следует уделять обучению сотрудников. Правительства и органы власти могут стимулировать бизнес к инвестированию в человеческий капитал путем введения новых налоговых льгот и льгот. Другими способами поддержки являются создание рабочих мест, наращивание потенциала путем обучения, повышение заработной платы, инвестиции в другие виды капитала, не забывая о НИОКР и технологиях.

Редизайн рабочего места.

Рабочий процесс и само рабочее место должны быть переработаны, чтобы соответствовать новым способам работы, когда люди взаимодействуют с машинами и алгоритмами. Это довольно сложная задача — выполнить требования безопасности при проектировании творческой среды. Компании также меняются, так как рабочий процесс включает в себя больше, чем раньше, взаимодействие человека и машины, а организации переходят на гибкие способы работы с меньшей иерархией.

Надлежащим образом спланированный и плавный переход рабочей силы.

Поскольку ожидается, что технический прогресс приведет к смене отрасли, места работы и, в некоторых случаях, самой профессии, многим людям потребуется поддержка, чтобы такой переход был плавным и менее напряженным. Для обеспечения этого должны быть разработаны, испытаны, приняты и реализованы специальные программы перехода. Соответственно, новые программы потребуют разработки новых систем и подходов.

Поощрение спроса на труд.

Лица, принимающие решения, должны рассмотреть пути увеличения спроса на работу. Одним из подходов являются новые инвестиции в инфраструктуру, меры по борьбе с изменением климата и другие инвестиционно-емкие проекты. Тип рабочих мест, создаваемых в таких случаях, будет в основном среднеоплачиваемым, что, как ожидается, больше всего повлияет на другие отрасли, такие как строительство, монтаж инженерных систем и т. д.

Безопасное внедрение ИИ.

Несмотря на то, что мы пользуемся преимуществами производительности этих быстро развивающихся технологий, нам необходимо активно защищаться от рисков и смягчать любые опасности. Работа с данными всегда связана с рисками, которые необходимо учитывать. Это безопасность данных, защита личных данных, мошенническое использование, потенциальные проблемы с предвзятостью. Правительства, предприятия и отдельные лица должны эффективно и ответственно решать все эти вопросы.

Организуйтесь для Agile.

Поскольку работникам предлагается выполнять больше творческой работы и проектов и меньше заниматься повторяющейся работой, рекомендуется, чтобы они делали это с большей автономией и возможностями для принятия решений.

Корпоративная культура является существенным и важным моментом в данном случае. Для поощрения творчества и экспериментов необходимо культивировать открытую культуру, которая должна распространяться на привлечение людей к принятию решений, которые изменят их рабочую среду и задачи, которые они выполняют. Компании также должны изменить свои рабочие процессы и структуры, которые позволят проектным группам беспрепятственно собирать и разбирать проект, предоставляя работникам менее традиционные функциональные ограничения.

Обсуждение

Каково будущее рабочего места — тема, которая вызывает широкий резонанс и обсуждается различными заинтересованными сторонами: властями, предприятиями и отдельными работниками. В течение довольно долгого времени исследователи, власти и консультанты пытались определить, как может вы-

глядеть рабочее место в будущем и как использовать его на благо общества и экономики, и как его изменения повлияют на людей и самые молодые поколения, которые будут начать работать в будущем.

Несмотря на высказанные разные мнения, общее мнение состоит в том, что технологическое развитие быстро стирает различия между работой, выполняемой людьми и машинами, и алгоритмами, что приведет к трансформации глобального рынка труда. При правильном управлении эта трансформация может привести к новой эре работы, новым рабочим местам и образу жизни людей. Но в случае плохого управления это может создать новые проблемы, такие как усиление неравенства, увеличение разрыва в навыках и поляризация общества. Несомненно, сейчас самое время влиять на будущее.

Хотя подробное обсуждение возможного влияния автоматизации рабочих мест в разных странах выходит за рамки этой статьи, важно отметить влияние надлежащих инвестиций в технологии ИИ на затраты на рабочую силу и навыки, которые мы можем предвидеть в странах с развитой экономикой, и сделать некоторые выводы уже сейчас. Например, по заключению одного из исследований, в 1997 г. добавленная стоимость [35] на доллар затрат на рабочую силу в промышленном производстве Мексики была в два раза выше, чем в США. К 2013 г. эта разница значительно сократилась и составила менее 15%. Принимая во внимание объем инвестиций в технологии ИИ в развитых странах, можно прогнозировать сдвиг экономического преимущества в затратах на рабочую силу, что повлияет на структуру промышленности развивающихся стран, таких как страны Юго-Восточной Азии или другие страны с развивающейся экономикой, путем изменения рабочих задач в таких отраслях, как производство одежды, обуви, текстиля или сборки электроники.

Наши результаты предсказывают резкое увеличение спроса на навыки ИИ. Количество должностей, требующих таких навыков, увеличилось в десять раз с 2010 по 2019 год и в четыре раза по доле вакансий [6]. Наибольшая доля вакансий, связанных с ИИ, приходится на секторы информационных технологий, профессиональных услуг, финансов и промышленного производства. Растущее изменение спроса заметно почти во всех отраслях, включая, например, сельское хозяйство, лесное хозяйство и рыболовство, в которых практически не было открытых вакансий. много лет назад.

Также было обнаружено, что существуют существенные различия между предприятиями, которые делают упор на навыки ИИ при поиске талантов, и теми, кто не обращает на это внимания. В перекрестном анализе фирмы с более высокой рыночной стоимостью, более высокой интенсивностью НИОКР и более высокими запасами денежных средств имеют более высокую долю вакансий ИИ.

В целом несколько исследований показывают значительное увеличение найма квалифицированного персонала по ИИ. Навыки ИИ также обеспечивают значительную надбавку к заработной плате в различных отраслях.

Приведенные в статье данные о составе рабочей силы, которые сопровождаются инвестициями в технологии ИИ, также могут способствовать нашему пониманию того, как ИИ влияет на рабочие и производственные процессы. Как предсказательная технология ИИ улучшает возможности людей делать прогнозы и принимать решения, что, в свою очередь, снижает спрос на управленческие должности. Несмотря на это, технологии ИИ не приводят к снижению спроса на высококвалифицированных работников, напротив, увеличивают долю высококвалифицированных работников на уровне компании. Дальнейшее понимание взаимосвязей между технологиями ИИ, операционными процессами и их влиянием на организацию может стать интересной темой для изучения в будущем.

Однако уже сейчас понятно, насколько важно разрабатывать и реализовывать программы обучения на протяжении всей жизни, планы переквалификации и повышения квалификации на национальном уровне. Принимая во внимание рекомендации, данные в этой статье, это также может быть плодотворным для будущих исследований. Это исследование может привести к дальнейшему сотрудничеству между правительствами и корпоративным переподготовкой в отдельных странах и регионах.

Кроме того, быстрые изменения в новых технологиях, приводящие к смене рабочих ролей и навыков, требуют гораздо большей скорости, чем когда-либо в прошлом. Вот почему решения и возможное сотрудничество между предприятиями, политиками и отдельными лицами в области обучения на протяжении всей жизни приобретают все большее значение и открывают широкие возможности для будущего партнерства.

Выводы

Технологии автоматизации и искусственного интеллекта меняют экономику и влияют на бизнес, ожидается, что они будут способствовать экономическому росту, повышая производительность труда и решая социальные проблемы. Наряду с этим, эти технологии будут способствовать трансформации рабочего места и восприятия самой работы.

Ожидается, что ИИ изменит рынок труда, изменив набор требований к навыкам и сочетание профессий. Тем не менее, ожидается, что это не приведет к сокращению занятости.

По некоторым профессиям прогнозируется серьезный спад, но наряду с этим появятся рабочие места других типов. Автоматизация рабочих мест не обязательно приводит к исчезновению профессий. Это скорее ведет к трансформации профессий, поскольку машины, алгоритмы и человеческий труд будут дополнять друг друга на рабочем месте.

Власти и бизнес несут ответственность за то, чтобы этот переход прошел гладко, создавая условия и среду, способствующие инновациям и развитию новых технологий.

Работникам может потребоваться переквалификация или повышение квалификации, чтобы адаптироваться к реорганизации задач и появлению новых задач, а также пережить возможную потерю работы и сориентироваться при переходе на новую работу. Это будет означать не только приобретение навыков, связанных с ИИ, но и приобретение навыков в областях, в которых ИИ не может работать так хорошо.

Продуктивное сотрудничество между людьми и роботами положит начало периоду новой работы и обеспечит конкурентное преимущество. Потенциал эффективного применения ИИ зависит от сотрудничества людей и машин, которое поможет создать новый опыт для конечных пользователей и разработать продукты, услуги и рынки, которых раньше не существовало. Это обеспечивает реальную возможность ИИ для человеческого общества.

Поскольку в настоящее время многие компании участвуют в цифровой трансформации бизнеса и кадровых стратегиях, у них есть множество возможностей использовать новые технологии, такие как автоматизация, для улучшения создания экономической ценности за счет внедрения новых видов деятельности, улучшения качества рабочего места, изменить традиционные или создать новые профессии, которые никогда не выполнялись ранее людьми, чтобы использовать весь потенциал своей рабочей силы.

Ссылки

1. Acemoglu D., Restrepo P. The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. // American Economic Review. – 2018. URL: https://clck.ru/32izef

2. Acemoglu, Daron, and David Autor. “Skills, tasks and technologies: Implications for employment and earnings.” In Handbook of labor economics. Vol. 4. // Elsevier. – 2011.

3. Ai, automation, and the future of work: ten things to solve for. // McKinsey Global Institute. – 2018. URL: https://clck.ru/32izco

4. Alderucci, Dean, Lee Branstetter, Eduard Hovy, Andrew Runge, and Nikolas Zolas. “Quantifying the Impact of AI on Productivity and Labor Demand: Evidence from U.S. Census Microdata.” – 2020.

5. Alekseeva L., Azar J., Giné M., Samila S., Taska B.: The Demand for AI Skills in the Labor Market. – 2021. URL: https://clck.ru/32izah

6. Babina T., Fedyk A., He A., Hodson J.: Firm Investments in Artificial Intelligence Technologies and Changes in Workforce Composition. – 2022.

7. Bakhshi, Hasan, et al., The Future of Skills: Employment in 2030. // Pearson, Nesta and The Oxford Martin School. – 2017.

8. Baldwin, Richard, The Great Convergence: Information Technology and the New Globalization. // Harvard University Press. – 2016.

9. Behrendt, Christina and Quynh Anh Nguyen, Innovative Approaches for Ensuring Universal Social Protection for the Future of Work, ILO Future of Work Research Paper Series No. 1. // International Labour Organization. – 2018.

10. Berg, Andrew, Edward Buffie and Luis-Felipe Zanna, Should We Fear the Robot Revolution? (The Correct Answer is Yes), IMF Working Paper No. 18/116. // International Monetary Fund. – 2018.

11. Bessen J. Automation and Jobs: When Technology Boosts Employment. // Boston University School of Law Law & Economics Paper. – 2018. URL: https://clck.ru/32izeH

12. Brynjolfsson, Erik, Tom Mitchell, and Daniel Rock. "What can machines learn, and what does it mean for occupations and the economy?". – 2018.

13. Brynjolfsson, Erik and Tom Mitchell. "What Can Machine Learning Do? Workforce Implications." // Science. – 2017.

14. Cappelli P., The consequences of AI-based technologies for jobs. // Wharton School of the University of Pennsylvania and Research Associate at the NBER. – 2020. URL: https://clck.ru/32izaJ

15. Cline, Bill, Maureen Brady, David Montes, Chris Foster and Davim, Catia The Augmented Workforce: 4 areas for financial insitutions to consider when pursuing intelligent automation for greater value and productivity. // KPMG Insights. – 2018. https://home.kpmg.com/xx/en/home/insights/2018/06/augmented-workforce-fs.html.

16. Dellot, Benedict, "Why automation is more than just a job killer", RSA Blog, 20 July 2018, https://www.thersa.org/discover/publications-and-articles/rsa-blogs/2018/07/the-four-types-of-automation-substitution-augmentation-generation-and-transference.

17. Deloitte, Reconstructing Jobs: Creating good jobs in the age of artificial intelligence. – 2018. https://www2.deloitte.com/content/dam/insights/us/articles/AU308_Reconstructing-jobs/DI_Reconstructing-jobs.pdf.

18. Deming, David, and Lisa B Kahn. "Skill requirements across firms and labor markets: Evidence from job postings for professionals." Journal of Labor Economics. – 2018.

19. Felten, Edward W., Manav Raj, and Robert Seamans. "The Occupational Impact of Artificial Intelligence: Labor, Skills, and Polarization." – 2019.

20. Franka M.R., Autorb R, Bessen J.E. Toward understanding the impact of artificial intelligence on labor. // Columbia University. – 2019. URL: https://clck.ru/32izby

21. Georgieff A., Hyee R. Artificial Intelligence and Employment: New Cross-Country Evidence. // AI for Human Learning and Behavior Change. – 2022. URL: https://clck.ru/32izdV

22. Global AI Survey: AI proves its worth, but few scale impact. // McKinsey Analytics. – 2019. URL: https://clck.ru/XD4Kx

23. Graetz, Georg and Guy Michaels. "Robots at Work." Review of Economics and Statistics. – 2018.

24. Hershbein, Brad, and Lisa B Kahn. "Do recessions accelerate routine-biased technological change? Evidence from vacancy postings." American Economic Review. – 2018.

25. Hirsch-Kreinsen, Hartmut, "Digitization of industrial work: development paths and prospects", Journal of Labour Market Research, vol. 49, no. 1 - 2016.

26. International Federation of Robotics, The Impact of Robots on Productivity, Employment and Jobs: A positioning paper by the International Federation of Robotics. – 2017.

27. Jesuthasan, Ravin and John Boudreau, Thinking Through How Automation Will Affect Your Workforce, Harvard Business Review. - April 2017.

28. Jovanovic, Boyan and Peter L Rousseau. "General Purpose Technologies." In Handbook of Economic Growth. Vol. 1, Part B, ed. Philippe Aghion and Steven N. Durlauf. // Elsevier. – 2005.

29. Kleiner, Morris and Ming Xu. “Occupational Licensing and Labor Market Fluidity.” Manuscript. // University of Minnesota. – 2017.

30. Lane M., Saint-Martin A. The impact of Artificial Intelligence on the labour market: What do we know so far? // OECD Social, Employment and Migration Working Papers. – 2021. - No. 256. URL: https://clck.ru/32izfW

31. McKinsey & Company, Skill Shift: Automation and the Future of the Workforce, Discussion Paper. // McKinsey Global Institute (MGI). – 2018.

32. Michaels, Guy, Ashwini Natraj, and John Van Reenen. “Has ICT Polarized Skill Demand? Evidence from Eleven Countries over Twenty-Five Years.” The Review of Economics and Statistics. – 2013.

33. Nedelkoska, Ljubica and Glenda Quintini, Automation, skills use and training, OECD Social, Employment and Migration Working Papers, No. 202, OECD, http://dx.doi.org/10.1787/2e2f4eea-en, 2018

34. Purdy M., Daugherty P. How ai industry profits and innovation boos. – 2017. URL: https://clck.ru/32izf9

35. Till Alexander Leopold, Vesselina Stefanova Ratcheva, Saadia Zahidi. The Future of Jobs Report. // Centre for the New Economy and Society. – 2018.

36. Toney A., Flagg M. U.S. Demand for AI-Related Talent. // CSET Data Brief. – 2020. URL: https://clck.ru/32izdu

37. Shook E., Knickrehm M. Reworking the revolution. – 2018. URL: https://clck.ru/32izep

38. van der Zande, Jochem, et al., The Substitution of Labor: From technological feasibility to other factors influencing job automation, Innovative Internet: Report 5. // Stockholm School of Economics Institute for Research. – 2018.

39. Vats, Anshu, Abdulkarim Alyousef and Stephen Clements, How Can Nations Prepare For the Industries of Tomorrow? “Make” It Happen – Harnessing the Maker Movement to Transform GCC Economies. // Oliver Wyman. – 2017.

40. Webb, Michael and Daniel Chandler. “How Does Automation Destroy Jobs? The ’Mother Machine’ in British Manufacturing.” // Stanford Mimeo. – 2018.

41. Webb M. The Impact of Artificial Intelligence on the Labor Market. // Stanford University. – 2020. URL: https://web.stanford.edu/~mww/webb_jmp.pdf

Journal of Digital Economy Research. 2023. 1. P. 137–176
DOI: 10.24833/14511791-2023-1-137-176

Received: 10 September 2022,
Revised 5 October 2022,
Accepted: 15 December 2022

THE IMPACT OF ARTIFICIAL INTELLIGENCE ON THE WORKPLACE: CURRENT STATE AND FUTURE PERSPECTIVES

by Ivan A. Lyapin

Abstract. Rapid technological developments in Artificial Intelligence technologies and its multiple applications on the workplace have stoked many fears and speculations in the society about possible job loss. Current studies report no decrease in demand of high-skilled workers associated with AI technologies. Also, there are no evidences that they will displace jobs in the near future. More companies keep on using and scaling AI in nearly every industry. Most of the businesses are getting benefits after implementation of AI. Most of the progress AI made relate to highly-educated and skilled occupations. To develop further, companies should decide which tasks and routines to be performed and delegate them to humans or robots. AI will not only lead to automation, but will also complement labor. Implementation of AI technologies widely across all industries still face challenges. Part of the limitations is technical related, there are also potential issues with data protection, its fraudulent use and cybersecurity are the main challenges to focus. Businesses and authorities should harness AI development and benefit from the increased productivity and working performance. Focus should be on ensuring the transition to new technologies is smooth and stressless.

Keywords: Artificial Intelligence, Workplace, Labor Market, Occupational Exposure

For citation: Lyapin I. (2023) The Impact of Artificial Intelligence on the Workplace: Current State and Future Perspectives. *Journal of Digital Economy Research,* vol. 1, no 1, pp. 137–176. (in English). DOI: 10.24833/14511791-2023-1-137-176

Information about the author:

Ivan A. Lyapin – MBA. Business Line Director, FAM-Holding.
117405, Moscow, st. Dorozhnaya, 60 B, 5th floor, office 516
lyapin.ivan@gmail.com
ORCID: 0000-0002-9754-1094

Introduction

Recent technological advancements in Artificial Intelligence (AI) initiated numerous discussions about large-scale job loss, because of its ability to automate a wide and continuously expanding set of tasks, and its potential to affect every sector of the economy.

Furthermore, use cases of the recent decade raised the concerns about employee well-being and the broader work environment, keeping in mind the opinion that AI may soon become commonly used in the workplace and threaten and humans' place in it. However, there is a possibility for AI also to complement and augment human capabilities, leading to higher productivity, greater demand for human labor and improved job quality.

From one perspective, dynamic development accompany technological change and create uncertainty about the future of work. On the other one, many authors agree that occupations are best understood as abstract sets of skills and that technology directly impacts demand for specific skills instead of acting on whole occupations all at once.

This article studies the views on the current level of AI technologies implementation, recent changes in workforce composition due to technology advancement, analyzes the sectors which are most exposed to working environment automation using new technologies. Then it provides the outlook on the possible changes on the workplace and working force requirements, summarizes the challenges of rapid implementation and provides recommendations to address the challenges.

Current changes in the workforce composition

In the recent period AI has made the most significant progress in occupations that are associated with non-routine and cognitive tasks that are usually performed by medium and highly skilled employees. This type of workers rely on abilities AI currently cannot provide, such as inductive reasoning or social intelligence. On top of that, it is easier for educated workers to adopt and learn new technologies and tend to participate more on new trainings which puts them in a better position to use the opportunities provided by AI on the labor market.

Studies, performed in 2022 in changes of workforce composition and workflow accompanying AI investments provide the view of how AI can transform the workplace, production processes and organization [6]. As a technology applicable mostly to work with cognitive abilities, AI improves the ability of a person to make decisions, which, in its turn, increases the autonomy of the workforce in predictions and reduces the demand for managerial staff.

Comparing to the previous automation technologies that contributed to productivity improvement, development of AI technologies is not associated with decrease in demand of high-skilled workers at the organizational level.

Studies also help to make conclusions about influence on the labor market. AI is associated with the high-skilled labor and there are no evidences that it will displace tasks that are commonly predicted to be displaced by AI, such as routine tasks for human resource, legal, administration. We will further shed the light on how AI is influences the jobs at non-AI investing firms and on the overall effect on the labor market as a whole.

Other studies show a substantial growth in need for skills associated with AI [5]. The number of positions requiring AI skills has increased ten times from 2010 to 2019 and four times as a percentage of all vacancies posted.

A large share of job postings is focused in Information Technology, Professional Services, Finance and Manufacturing sectors, but the increase in demand is visible in the other industries. Likewise, the demand in AI skills is following from Computer, Engineering and Science occupations and also in such occupations as Farming, Forestry and Fishing which had almost zero demand for AI skills when the observation started.

In job postings studied with comparable other skills, the addition of AI skill requirement increases the offered wage by 16%. Controlling for firm fixed effects, the AI wage premium is almost 20%. This premium is significantly higher than premium for other skills in the analysis.

There were also findings in substantial wage premium for non-AI associated occupations in the firms with higher AI investment intensity. Findings suggest that jobs that require software, cognitive, social, project management, and people management skills are complementary to AI jobs, while jobs requiring customer service skills might be substituted by AI.

Recent studies [1] formulated more formally the explanation which is traditionally used as an argument why technological progress and labor-saving practices are not followed by job losses: productivity increase is resulting in higher fuel demand in the economy in general and generates more job openings, although, as a rule, in areas other than where the initial performance improvements occur. There are numerous ways in which productivity improvements and demand can be linked.

AI at the workplace current use cases

AI adoption is on the rise in almost every industry, but opportunities are different.

The research and survey made by McKinsey & Company shows the results that the organizations implemented at least one of the AI use possibilities into their day-to-day operations of product process at least for one work function or department. Responses also show the share growth of firms applying AI in their processes or

products within different functions [22]. This is an indication that more companies are utilizing and scaling AI. Highly intensive in AI implementation companies are much further in these efforts.

Analyzing different industrial sectors, we can notice increases in AI implementation almost in every industry in the recent years. Among the grow leaders retail is the top performing sector where companies introduced one or more AI applications in at least one function or department.

Exhibit 1. Industries generally using AI capabilities.

By function, we find the most use cases in operations (RPA, predictive maintenance, production analytics, inventory and supply chain, robotics, invoicing), general solutions (analytics, automated machine learning) and specialized solutions (geo-analytics, conversational analytics, image recognition, e-commerce analytics), marketing (personalized and context marketing, analytics), sales (forecasting, lead generation, data input automation, lead scoring, agent coaching, sales analytics), customer service (social listening, ticketing, call routing, voice authentication, responses, chatbots, call, survey and review analytics), HR (hiring, performance management, retention management, employee monitoring, building management, HR analytics). There are also numerous use cases in Tech, Data processing and other innovative functions.

We see that firms adopting AI technologies in the areas that support them in their industries to capture value and benefits. For instance, respondents representing FMCG companies claim higher use of physical robots, which are usually more often used in production and assembly, than any other application opportunities. Respondents representing telecom industry inform that using virtual agents is common for their companies, which can improve customer-service operations, more frequent than most of the other capabilities (Exhibit 1). Other high-performing companies report that they implemented AI in sales and marketing.

Analyzing the report by geography, we see the results that show that the share of AI adoption increased substantially in technologically advanced countries of Europe, Latin and North America, Asia-Pacific. High level of AI implementation puts these regions together with China at the similar level of AI introduction and implementation at the workplace stating that AI is a global phenomenon at the workplace while the intensity on the company level may vary significantly.

AI implementation champions tend to apply value-capturing techniques

From several researches and studies we see that most of the firms' activities are essential to fix the value at scale. In many cases these are analytics, business alignment and IT leaders focus on potential to capture value from AI applications in every business sector. Activities include, but not limited in investing in talent and ensuring that business and technical staff have the skills for successful scaling. The survey studied suggests that these practices are essential for scaling AI, but it is necessary to keep in mind that respondents of the survey are more likely to say their companies utilize such practices.

However, even AI high performers say that their most skilled staff use insights they got from AI for daily tasks in decision making and less than a half systematically track a well-defined set of KPI's for AI, two practices, according to McKinsey, that are critical to achieving acceptance and value by end users. In a similar way, only 35 percent of high-performing AI respondents declare that they have an active continuous training program for their workers on AI technologies [22].

Occupational Exposure to AI

Many companies across the globe are currently implementing AI in the workplace alongside with significant investments in its adoption and implementation making AI a part almost of every stage of employees' work. It is widely used in recruiting and onboarding, in talent development and process management.

AI becomes the main tool either to fully automate the working process or enhance and assist the workers in performing their tasks by using robotics, process automation and other technologies.

The activities AI has made the most significant progress in relate to highly-educated white-collar occupations. Thus highly-skilled occupations are among those with the most significant exposure to AI technologies. These are Science and Engineering, Business and Administration professionals, managers, Executives, people working in Legal, Social and culture.

On the other hand, occupations where exposure is lowest representing professions where people are involved in physical activities. Such as: Cleaners, Agricultural workers, Forestry and Fishery Laborers, Food cooking workers. These professions rely on the abilities where AI has not been implemented much so far.

The research shows how AI exposure is distributed across different countries. In most of the cases, an occupation's exposure to AI is not high, fundamental differences between occupations are usually higher [21]. We can notice slight difference in AI occupational exposure between Baltics and Northern Europe countries analyzed. Also, studying the exposure between occupations, we can notice that distributions across occupations in most exposed and least exposed countries are very similar. Exposure will keep on being different across regions and industries due to differences in many social and economic factors, different labor markets and their dynamics: in developed economies with quite high salary levels, the exposure could be more than two times higher than in such countries as India.

AI technologies are rapidly spreading around the glove and are expected to have big effects on world economy. Giving the effect that we can suspect; it has become critical to have data that allow us to study possible effects. Much of the work to be done to identify the possible influence of AI to the labor. Therefore, it is becoming increasingly important to be able to operate with trusted and reliable global data.

Drivers for AI implementation at the workplace

Most businesses can see the payoff from AI.

With a reference to research, conducted by McKinsey & Company where they studied about 33 cases of AI implementation in eight business functions. They also studied the effect of implementation of AI in each of these activities on the revenue and cost in the departments using AI [22]. Hereby, we provide the results which concluded that AI is delivering a significant value for companies.

A significant share of respondents reports that adoption of AI technologies in the companies increased the revenue on more than 10 percent. The most gains are reported from AI used in sales and marketing (price policies, probability of making a purchase), supply chain management (sales and demand forecasting), product and service development (designing the AI based products new for the markets). In its report Accenture estimates that the firms intensively investing in AI-based technologies can boost sales revenues by 38 percent [37]. There is also an opportunity to increase the employment level by 10 percent.

Additionally, about a half of participants report cost reduction resulted from AI implementation. The most cases of cost saving can be found in manufacturing (optimization of the output, energy and overall capacity) and supply chain management (supply chain optimization).

AI provides analytics that help humans to take sound decisions.

The benefits of new human-AI collaboration are getting clear. A good example is the recent AI-based algorithm developed by a team of scientists from Harvard aimed to identify breast cancer cells with higher precision. When we compare human performance versus AI's, pathologists identify the decease with 96 percent accuracy while algorithm alone do that in 92 percent of the cases. But when human-AI alliance is performing the job, accuracy of diagnostics increases up to 99.5 cases for cancer biopsies identification.

Such a successful use case could be created due to human-AI ongoing collaboration and training to help improve the effectiveness of it.

Economy and society as whole are getting more opportunities from AI intensification.

Besides value for businesses, as it was already mentioned above, AI technologies make impact on economy, supporting its growth and may provide a significant progress on most societal challenges. It can influence aging and contribute to birth rates increase, labor productivity, which is one of the major factors of economic expansion. AI is also finding applications in many other areas: medical researches, climate exploration, material science etc.

Future and trends of AI at the workplace

We can definitely say that the future is uncertain not just with respect to workplace, but almost to any sphere of life. And it is also hard to be certain what to do about the future. But it is important for us to take the guess about the future. With regard to workforce economic outlook, the experience we possess does not allow to predict the type of jobs which will be in high demand in the distant future. Even if could have reasonable understanding which jobs will be eliminated in the future, we need to be reasonably sure which jobs will be demanded then.

To make the forecasting more effective and accurate, it might make sense to plan the future on the individual company level rather than the economy as a whole. From the previous paragraphs we know that the firms with AI-focused strategies and

investments are better suited and prepared for the new technologies adaptation and better suited to new approaches. Furthermore, if we make an assumption that AI-based technologies may decrease future jobs demand, it looks much more effective to make the development program on individual employer's level because region or global-wide forecasts may only be used by the small group of employees at any time. Also, in case an individual wishes to have a transition from one role to another, it is much easier to do within the same organization where their company specific skills and knowledge remain relevant and demanded.

To address the issue of uncertainty, first of all, we must admit that even presumably good forecasting models provide us the most probable outcomes or estimations of the most likely solution of our problem. Practical experience has shown that in many cases even the most probable outcome is not that likely, this is why it is important to forecast second and even the third most probable solution. To address these outcomes, scenario planning is one of the possible techniques. Another one is simulation, where we can create the forecasting model and see the possible outcomes by changing the values of initial variables.

Without doubts, companies have to be ready to face major changes at work. It would be possible to automate about fifty percent of the activities (not jobs themselves) carried out by workers and assistants. The research of 2 thousand work activities within 800 occupations provides the data about certain types of activities that are easier to automate [3]. These are physically performed activities in fixed and stable environments, data gathering and processing. These may be roughly estimated as a half of all activities people perform across all industries. The lease threatened activities include taking decisions, providing expertise, communicating with stakeholders. Taking into account current level of technologies, it is possible to fully automate only 5 percent of currently existing professions, but almost all of them will be affected in the future.

Can AI help to solve labor market issues? Most of the companies using AI technologies at the workplace tend to think so. They are either increasing the intensity of AI use to reduce general hiring needs or on their way to develop a plan to use it. As was already mentioned, AI comes with its own challenge for labor market: the need to recruit skilled and often expensive AI talent. There is a number of cases when companies slowed down their recruitment process because of the limited availability of AI talent. Nonetheless, despite that the demand for AI and data science staff is continuously increasing, there are ways to mitigate the recruitment slowdown.

Below are our recommendations of how companies could prepare their businesses for the future.

Businesses should reconsider the tasks to be carried out.

It is recommended to reassess the set of human teams and technologies available to allocate the tasks and activities to humans or robots. This is not a one-time-task, the process of such allocation should be ongoing to increase the its efficiency. Most of the existing AI are not able to work autonomously and require human assistance

to interfere, input data, make necessary corrections in their operations. In many cases where AI is supposed to take the decisions, it should be train by the people to have the results of their work acceptable and useful for further use.

AI will not only lead to automation, but will also complement labor. As another example of human – AI collaboration, we can provide the case when an analytic needs to write a report based on dataset provided and collected by AI (high suitability for machine learning) but then provide recommendations and guidance in it which will require abstract reasoning and perceived to come from a human. AI is expected to increase productivity not only by enabling companies to replace labor with cheaper capital, but also by complementing workers enabling them to increase their productivity by leveraging their social interaction skills and ability to reason about novel situations.

Companies need to implement AI Governance.

To stay tuned with continuously evolving AI models, companies need to deploy the governance models specifically designed for AI application which may influence all the relevant stakeholders. The governance should include the risk, AI and business leaders with new policies and procedures, roles and responsibilities. An example of possible governance model to be used is Three Lines Model which was developed in 2008-2010 by the Federation of European Risk Management Associations and was adopted by the Institute for Internal Auditors in 2013. The model describes three lines of defense to apply for AI Governance:

- Creators, Executors and Operations
- Managers, Supervisors and Quality Assurance
- Auditors and Ethicists

In addition to three lines model, it is also recommended to use end-to-end governance model. It includes all the stages of organization's lifecycle.

Also, despite the possibility to employ and enhance much of the existing IT governance and controls, it is very important that business leaders learn and understand some AI and data science basics.

Companies need to understand their unique vulnerabilities.

Companies acting in different sectors of economy, such as finance, retail, utilities etc., all face different kinds of risks from potential AI implementation: where it could interfere into data sets and cause serious damage. It is important for the companies to identify their unique vulnerabilities and define the risk from their specific AI algorithms potentially to be used. It is effective to estimate the value of the possible resulting financial, operational and reputational risks and the value of its mitigation to take the decision about AI implementation and prioritization of efforts.

Companies need to control their data.

Most probably, the traditional ways to control the data handling are not sufficiently robust to detect particular problems that can cause the risk associated with AI use. It is important to pay special attention to issues in historical data and data

acquired from third parties. When implemented, AI algorithms may creep the data sets and allow to take biased decisions, therefore, they must be well designed and managed from a very beginning.

Companies need to focus on employees.

The shortest way to implement AI at the workplace is to address AI talent shortage. It is important to take and hire new specialists who already have the skills you will need or can easily acquire them. Companies should consider teaching data science, machine learning, software engineering, statistics and business leaders a little of all from these fields. Such approach can also help to fill the communication gaps and increase cooperation between the groups that work with AI.

To mitigate the possible risks of biased AI at the workplace companies need to diversify their teams with people with different backgrounds, gender, races. A diverse team bringing together business leaders, data scientists, engineers and other professionals with different backgrounds and experiences will allow different views on the possible threats and ways to mitigate them.

By implementing AI, companies will be able to reduce the need for routine work and make the life of their employees easier and more engaging. By announcing the plans and starting the process of AI implementation, companies can increase the value the employees can give to them. AI can even help provide emotional support at the workplace thus making employees happier.

Companies can assess and predict ROI of AI.

There is no need for companies to implement AI models which don't focus on important business issues. The more complex the task the more important that businesses aim at the right issue to solve. It is hard to quantify and measure AI's ROI which are often initiated by strategic decisions or disruptions which never existed on the market. However, businesses now are more likely to measure the value of impact using new assessment methods. These can capture not just "hard" returns, such as increased productivity, and "hard" costs, such as new hardware spending. They can also capture "soft" returns, such as an improved employee experience, and "soft" costs, such as increased demands on subject matter specialists' time. The holistic approach to AI, many existing use cases and practices also help to estimate the ROI of new initiatives. It is also possible to use AI itself to model the impact and uncertainties of AI initiatives and help to better allocate the resources.

When modelling it is important to make sure that all different approaches and possible changes are tested and their impact is measured.

It may also be useful to consider portfolio approach to avoid ROI surprises as companies may also work with product innovation: create and assess a mix of initiatives that will raise the likelihood of delivering the overall results companies need.

And we will finish from what we started: to effectively and precisely predict AI's ROI, companies need to manage and govern not individual projects, but full life-cycle. That can help them continually evolve strategy, fine-tune execution and find new use cases for data — both avoiding downside surprises and identifying new value.

Influence on employment

Despite the fact that AI is considered to be an emerging technology, it is already driving a strong demand for related talents across many industries [36]. It is already here to stay in numerous economic sectors. This demand is predictably high in Scientific and Technical services sectors, it also shows high levels in Manufacturing, Finance and Insurance and administrative ones. Information sector is the sector people intend to go first when they think about AI. However, there is a little empirical evidence how AI influences the labor market to date and it will transform it in the future.

Jobs lost: by 2030 some occupations are expected to have significant declines.

Rapidly developing AI technologies will displace some people. Research shows that about 15 percent of all the workers, or about 400 million people, could loss their workplaces because of AI in the period 2016–30 [3].

The technical possibility to automate the jobs is the main reason that causing the decline. There are also other reasons, such as development cost, labor market changes caused by the quantity and quality of labor supply and its corresponding wages, the gains due to human displacement that are mentioned in the adoption use cases and last but not least, acceptance by society and its norms.

Jobs gained: There will be more jobs created during the same period.

Despite that many jobs are expected to be displaced, the increasing demand for work is predicted which in its turn will provide more jobs. Such factors as higher spending on social needs, growing incomes, expected new investments in significant projects in energy, infrastructure, development of new technologies and their practical applications, contribute to this statement. The main beneficiaries from these changes are going to be emerging economies, such as numerous Asian countries where the working age population is increasing dramatically. Such a growth in economy and population will affect business dynamism and influence the productivity increase and will keep on the growth of new jobs on the market.

Jobs changed: Jobs are expected to change as technologies will collaborate with humans at the workplace.

There will be evident benefits from human labor complementation by machines and this factor will cause partial automation to prevail. Dealing with repetitive tasks will shift current jobs to automated systems troubleshooting and maintenance.

Challenges of AI implementation and its mitigation

Implementation of AI technologies widely across all industries still face challenges. Part of the limitations is technical related, such as need for significant data training and difficulties to apply same algorithms across different applications. It is hard to identify all the issues related to recent innovations. Potential issues with training data, algorithms optimization, private data protection, fraudulent use and security are the main challenges to focus.

However, we can highlight the main challenges companies are facing now. The set of occupations is expected to change as well as requirements for working skills and educational background. Working activities will have to be reconsidered to make sure humans and AI collaborate most effectively.

Higher emphasis on new skills trainings is expected.

Results of the research report that the majority of firms are getting ready for AI-related personnel changes [22]. About 60 percent of respondents claimed that their companies adopting AI have workers retrained during the past year. On top of that, 83 percent of research participants have plans to retrain some of their workers in the next three years because of AI implementation and 38 percent have plans to retrain more than a quarter of the staff.

AI technologies at the workplace will increase the pace of shift in skills demanded. Requirements for technical skills will grow rapidly. The demand for cognitive skills, such as analyze of information, critical thinking, creativity will grow as well. This will put a greater pressure on the existing workforce and firms to develop the solutions that scale its skills along with the progress.

Some workers will probably need to reconsider occupations.

Requirements to change and develop new working skills may lead to the need of change of occupational categories. Some of these changes are expected across sectors and companies, but we should also expect that many will happen within sectors and geographies. Occupations related to data processing and highly structured environments will see the decline, while occupations related to unpredictable environments will meet increasing demand for work. Some other demanded occupations will include teachers, managers, tech and other professionals.

Automation is expected to influence wages in advanced economies.

The change in occupational activities will likely to put pressure on wages. Highly automatable activities are the most likely to be substituted by AI and shift to higher skilled jobs will happen. The number of high wage jobs will increase significantly causing the rick that automation will increase wage polarization and income inequality.

Many economies are already facing the challenge to properly educate and retrain their workers to ensure they are able to meet the existing demands of employers [2]. In OECD countries, workforce education and training spending has been in decline for two last decades. Worker transfer spending and relocation support also kept on declining as a share of GDP.

In seeking solutions to problems, we must not slow down the continuous development and application of new technologies. Businesses and governments should benefit from developments in AI technologies that bring improved productivity and work performance not as well as social benefits. Focus should be on ensuring the transition to new technologies is smooth and stressless. Automation and productivity growth is a main contributor to job and economic development. This is why support of investments and new AI technologies development is crucial.

Addressing all the challenges it is important not to miss a critical observation: AI technologies will not much influence the quantity of jobs, but rather job requirements and tasks.

Below is the list with our recommendations of the areas where main efforts should be concentrated.

Fostering business dynamism.

The opportunity for entrepreneurship to develop further will result in new business formation which, in its turn, will create new jobs and increase working performance. Business dynamism depends on open and clear environment for small businesses and healthy competitive environment for large business. Simpler and better rules, tax and other incentives will be required to accelerate the formation of new business and increase the competitiveness of large and small enterprises.

Constantly improving education system and affordable trainings for changed workplace.

Businesses together with authorities could contribute more to development and evolve basic and advance skills via school educational systems and corporate trainings. A new focus on creativity, critical and systems thinking, and adaptive lifelong learning is required.

Human capital investments.

As mentioned above, much attention should be paid to the training of employees. Governments and authorities are able to encourage businesses to invest in human capital by introduction of new tax incentives and benefits. Other ways to support are jobs creation, building capabilities by providing training, increase wages, make investments in other types of capital, not forgetting about R&D and technologies.

Workplace redesign.

Working process and workplace itself will require to be redesigned to satisfy the new ways of working where humans collaborate with the machines and algorithms. This is quite a challenge to fulfil safety requirements while designing creative environment. Companies are also changing as working process involves human-machine collaboration more than before and organizations are transforming to agile ways of working with less hierarchy.

Properly planned and smooth transition for workforce affected.

As the technological progress is expected to cause the change of industry, place of work and, in some cases, occupation itself, many people will require support to have such a transition smooth and less stressful. Special transition programs should be developed, tested, adopted and implemented to ensure this. Respectively, new programs will require development of new systems and approaches.

Demand for work incentivization.

Decision-makers should consider the ways to increase the demand for work. One of the approaches is new investments in infrastructure, climate change measures and other investment intensive projects. The type of jobs created in such cases will mostly be middle-wage jobs, which are expected to be most affected in other sectors, such as construction, engineering systems installation etc.

Safe implementation of AI.

While we are taking advantage of the performance benefits of these rapidly evolving technologies, we need to proactively guard against risks and mitigate any hazards. Working with data is always associated with risks which have to be taken into account. These are data security, private data protection, fraudulent use, potential issues with bias. Governments, businesses and individuals should address all these issues effectively and responsibly.

Organize for agility.

Because workers are suggested to do more creative work and projects and involved less in repetitive work, it is recommended that they do that with higher autonomy and decision-making opportunities.

Corporate culture is an essential and important point in this case. To encourage creativity and experimentation it is required to cultivate open culture which should extend to involving people in making decisions that will change their work environment and the tasks they perform. Companies should also change their workflow and structures that allow project teams to seamlessly assemble and disassemble, providing workers less traditional functional constraints.

Discussion

What is the future of the workplace is a topic which resonates broadly and has been discussed among different stakeholders: authorities, businesses and individual workers. For quite a long time, researches, authorities and consultants have been trying to define how the workplace might look like in the future and how to utilize it for the good of society and economy and how will its changes affect the individuals and youngest generations who will start working in the future.

Despite different opinions expressed, the common ground is that technological development is rapidly diffusing the differences between work done by humans and machines and algorithms which will result in global labor market transformation. If managed properly, this transformation could lead to the new era of work, new jobs and way people live. But in case of poor management, it may impose new challenges, such as greater inequality, wider skills gaps and society polarization. Undoubtedly, now is the time to influence the future.

While the detailed discussion of the possible effect of workplace automation in different countries is beyond this article, it is important to notice the effect of proper investments in AI technologies on labor costs and skills which we can foresee in advanced economies and make some conclusions already now. For instance, as one of the studies concluded, in 1997, value-added [35] per dollar of labor cost was twice higher in industrial manufacturing in Mexico than those in the United States. This difference had significantly decreased to less than 15% by 2013. Taking into account the amount of investments in AI technologies in developed countries, it is possible

to predict the shift in economical advantage in labor costs which will influence the industrial structure of developing economies, such as South-East Asia countries or other emerging economies through reshaping the work tasks in such industries as apparel, footwear, textiles or electronics assembly.

Our findings predict a dramatic increase in the demand for AI skills. The number of positions requiring such skills increased ten times from 2010 to 2019 and four times as a share of job openings [6]. The biggest share in AI-related job postings is from Information Technology, Professional Services, Finance and Industrial Manufacturing sectors the growing change in demand is visible almost in all industries, including, for example, Farming, Forestry and Fishing which had almost no opening few years ago.

It was also found that there are significant differences between businesses that are focusing on AI skills when looking for talents and those who do not pay attention to it. In a cross-sectional analysis, firms with higher market values, higher R&D intensity, and higher cash stock have a higher share of AI vacancies.

All in all, several studies show the significant increase of the recruitment of AI skilled staff. AI skills also provide a significant wage premium across different industries.

An evidence noted in the article of the workforce composition which go along with investments in AI technologies could also contribute to our understanding of how AI is shaping the work and production processes. As a predictive technology, AI improves the possibility of individuals to make predictions and take decisions which, in its turn, reduces the demand for managerial positions. Despite that, AI technologies do not lead to the decrease in demand of high-skilled workers, opposite, it is increasing the share of high-skilled workers on the company level. Further understanding of dependencies between AI technologies, operational processes and their impact on organization could be an interesting topic to study in the future.

However, it is already clear how important it is to develop and implement lifelong learning programs, reskilling and upskilling plans on the national levels. Taking into account recommendations given in this article, this could also be a fruitful are for future research. This research could lead to further collaboration between governments and corporate reskilling in selected countries and regions.

Also, rapid changes in new technologies resulting in shifts of the job roles and skills demand at a much higher speed than ever in the past. This is why the decisions and possible collaboration between businesses, policy-makers and individuals in the field of lifelong learning are getting higher importance and provide significant opportunities for future partnerships.

Conclusions

Automation and AI technologies are changing the economies and influence businesses, they are expected to contribute to economic growth increasing work performance and solving social issues. Along with this, these technologies are going to fuel the transformation of the workplace and to perception of work itself.

It is expected that AI will transform the labor market changing the set of skill requirements and mix of occupations. Yet it is not expected to lead to declining employment.

Serious declines are predicted for some occupations, but along with this, jobs of other types will appear. Workplace automation doesn't necessarily lead occupations to disappear. It rather leads to occupations transformation as the machines, algorithms and human labor will complement each other at the workplace.

It is in the responsibility of authorities and businesses to ensure this transition will happen smooth by creating the conditions and environments that foster innovations and new technologies development.

Workers may need to re-skill or up-skill in order to adapt to the reorganization of tasks and new tasks appearance, and to weather potential job loss and navigate transitions to new jobs. This will not only mean acquiring AI-related skills, but also acquiring skills in areas that AI cannot perform so well.

Productive collaboration between humans and robots will start the period of new work and provide competitive advantage. The potential of AI effective application depends on humans and machines collaboration which will help to create new end user experiences and to develop products, services and markets which never existed before. This provides the real opportunity of AI for human society.

As many companies nowadays have been involved in business digital transformation and workforce strategies, there is a big number of opportunities for them to leverage the utilization of new technologies, such as automation, to improve the creation of economic value by introduction of new activities, improvement of workplace qualities, change the traditional or create new occupations which have never been performed before by human workers to utilize the full potential of their workforce.

References

1. Acemoglu D., Restrepo P. The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. // American Economic Review. – 2018. URL: https://clck.ru/32izef

2. Acemoglu, Daron, and David Autor. "Skills, tasks and technologies: Implications for employment and earnings." In Handbook of labor economics. Vol. 4. // Elsevier. – 2011.

3. Ai, automation, and the future of work: ten things to solve for. // McKinsey Global Institute. – 2018. URL: https://clck.ru/32izco

4. Alderucci, Dean, Lee Branstetter, Eduard Hovy, Andrew Runge, and Nikolas Zolas. "Quantifying the Impact of AI on Productivity and Labor Demand: Evidence from U.S. Census Microdata." – 2020.

5. Alekseeva L., Azar J., Giné M., Samila S., Taska B.: The Demand for AI Skills in the Labor Market. – 2021. URL: https://clck.ru/32izah

6. Babina T., Fedyk A., He A., Hodson J.: Firm Investments in Artificial Intelligence Technologies and Changes in Workforce Composition. – 2022.

7. Bakhshi, Hasan, et al., The Future of Skills: Employment in 2030. // Pearson, Nesta and The Oxford Martin School. – 2017.

8. Baldwin, Richard, The Great Convergence: Information Technology and the New Globalization. // Harvard University Press. – 2016.

9. Behrendt, Christina and Quynh Anh Nguyen, Innovative Approaches for Ensuring Universal Social Protection for the Future of Work, ILO Future of Work Research Paper Series No. 1. // International Labour Organization. – 2018.

10. Berg, Andrew, Edward Buffie and Luis-Felipe Zanna, Should We Fear the Robot Revolution? (The Correct Answer is Yes), IMF Working Paper No. 18/116. // International Monetary Fund. – 2018.

11. Bessen J. Automation and Jobs: When Technology Boosts Employment. // Boston University School of Law Law & Economics Paper. – 2018. URL: https://clck.ru/32izeH

12. Brynjolfsson, Erik, Tom Mitchell, and Daniel Rock. "What can machines learn, and what does it mean for occupations and the economy?". – 2018.

13. Brynjolfsson, Erik and Tom Mitchell. "What Can Machine Learning Do? Workforce Implications." // Science. – 2017.

14. Cappelli P., The consequences of AI-based technologies for jobs. // Wharton School of the University of Pennsylvania and Research Associate at the NBER. – 2020. URL: https://clck.ru/32izaJ

15. Cline, Bill, Maureen Brady, David Montes, Chris Foster and Davim, Catia The Augmented Workforce: 4 areas for financial insitutions to consider when pursuing intelligent automation for greater value and productivity. // KPMG Insights. – 2018. https://home.kpmg.com/xx/en/home/insights/2018/06/augmented-workforce-fs.html.

16. Dellot, Benedict, "Why automation is more than just a job killer", RSA Blog, 20 July 2018, https://www.thersa.org/discover/publications-and-articles/rsa-blogs/2018/07/the-four-types-of-automation-substitution-augmentation-generation-and-transference.

17. Deloitte, Reconstructing Jobs: Creating good jobs in the age of artificial intelligence. – 2018. https://www2.deloitte.com/content/dam/insights/us/articles/AU308_Reconstructing-jobs/DI_Reconstructing-jobs.pdf.

18. Deming, David, and Lisa B Kahn. "Skill requirements across firms and labor markets: Evidence from job postings for professionals." Journal of Labor Economics. – 2018.

19. Felten, Edward W., Manav Raj, and Robert Seamans. "The Occupational Impact of Artificial Intelligence: Labor, Skills, and Polarization." – 2019.

20. Franka M.R., Autorb R, Bessen J.E. Toward understanding the impact of artificial intelligence on labor. // Columbia University. – 2019. URL: https://clck.ru/32izby

21. Georgieff A., Hyee R. Artificial Intelligence and Employment: New Cross-Country Evidence. // AI for Human Learning and Behavior Change. – 2022. URL: https://clck.ru/32izdV

22. Global AI Survey: AI proves its worth, but few scale impact. // McKinsey Analytics. – 2019. URL: https://clck.ru/XD4Kx

23. Graetz, Georg and Guy Michaels. "Robots at Work." Review of Economics and Statistics. – 2018.

24. Hershbein, Brad, and Lisa B Kahn. "Do recessions accelerate routine-biased technological change? Evidence from vacancy postings." American Economic Review. – 2018.

25. Hirsch-Kreinsen, Hartmut, "Digitization of industrial work: development paths and prospects", Journal of Labour Market Research, vol. 49, no. 1 - 2016.

26. International Federation of Robotics, The Impact of Robots on Productivity, Employment and Jobs: A positioning paper by the International Federation of Robotics. – 2017.

27. Jesuthasan, Ravin and John Boudreau, Thinking Through How Automation Will Affect Your Workforce, Harvard Business Review. - April 2017.

28. Jovanovic, Boyan and Peter L Rousseau. "General Purpose Technologies." In Handbook of Economic Growth. Vol. 1, Part B, ed. Philippe Aghion and Steven N. Durlauf. // Elsevier. – 2005.

29. Kleiner, Morris and Ming Xu. "Occupational Licensing and Labor Market Fluidity." Manuscript. // University of Minnesota. – 2017.

30. Lane M., Saint-Martin A. The impact of Artificial Intelligence on the labour market: What do we know so far? // OECD Social, Employment and Migration Working Papers. – 2021. - No. 256. URL: https://clck.ru/32izfW

31. McKinsey & Company, Skill Shift: Automation and the Future of the Workforce, Discussion Paper. // McKinsey Global Institute (MGI). – 2018.

32. Michaels, Guy, Ashwini Natraj, and John Van Reenen. "Has ICT Polarized Skill Demand? Evidence from Eleven Countries over Twenty-Five Years." The Review of Economics and Statistics. – 2013.

33. Nedelkoska, Ljubica and Glenda Quintini, Automation, skills use and training, OECD Social, Employment and Migration Working Papers, No. 202, OECD, http://dx.doi.org/10.1787/2e2f4eea-en, 2018

34. Purdy M., Daugherty P. How ai industry profits and innovation boos. – 2017. URL: https://clck.ru/32izf9

35. Till Alexander Leopold, Vesselina Stefanova Ratcheva, Saadia Zahidi. The Future of Jobs Report. // Centre for the New Economy and Society. – 2018.

36. Toney A., Flagg M. U.S. Demand for AI-Related Talent. // CSET Data Brief. – 2020. URL: https://clck.ru/32izdu

37. Shook E., Knickrehm M. Reworking the revolution. – 2018. URL: https://clck.ru/32izep

38. van der Zande, Jochem, et al., The Substitution of Labor: From technological feasibility to other factors influencing job automation, Innovative Internet: Report 5. // Stockholm School of Economics Institute for Research. – 2018.

39. Vats, Anshu, Abdulkarim Alyousef and Stephen Clements, How Can Nations Prepare For the Industries of Tomorrow? "Make" It Happen – Harnessing the Maker Movement to Transform GCC Economies. // Oliver Wyman. – 2017.

40. Webb, Michael and Daniel Chandler. "How Does Automation Destroy Jobs? The 'Mother Machine' in British Manufacturing." // Stanford Mimeo. – 2018.

41. Webb M. The Impact of Artificial Intelligence on the Labor Market. // Stanford University. – 2020. URL: https://web.stanford.edu/~mww/webb_jmp.pdf

Учредитель: Федеральное государственное автономное образовательное учреждение высшего образования «Московский государственный институт международных отношений (университет) Министерства иностранных дел Российской Федерации».

Адрес редакции: 119454, Россия, Москва, пр. Вернадского 76
Тел.: 8 (903) 623-95-15

веб-сайт: https://jder.mgimo.ru
электронная почта: jder@inno.mgimo.ru

Дизайн – Волков Д.Е., Церех Ю.К.; редакторы – Рыжкова А.И., Миляева Е.В.;
верстка – Волков Д.Е.

Отпечатано в отделе оперативной полиграфии и множительной техники МГИМО МИД России.
119454, Москва, проспект Вернадского, д. 76.

Тираж 150 экз. Объём 15,5 усл. п.л. Заказ № 453.

The Founder: Moscow State Institute of International Relations (University) of the Ministry of Foreign Affairs of Russian Federation.

The Publisher Address : 119454, Moscow, Prospect Vernadskogo, 76
Phone/fax: 8 (903) 623-95-15

URL: https://jder.mgimo.ru
e-mail: jder@inno.mgimo.ru

Published by MGIMO University Press. Number of printed copies: 150.