DOI: 10.30727/0235-1188-2024-67-3-99-122
Оригинальная исследовательская статья
Original research article

Искусственный интеллект как фактор трансформации государства и общества: поиск баланса между управленческой эффективностью и человекоориентированностью*

Б.Б. Славин
Финансовый университет при Правительстве РФ, Москва, Россия

Аннотация

В статье представлен социально-философский анализ процесса интеграции технологий искусственного интеллекта (ИИ) в систему государственного управления. Акцент сделан на поиск оптимального баланса между повышением эффективности административных процессов и сохранением гуманистических ценностей. Изложены различные подходы к пониманию роли ИИ в современном обществе: от технооптимистических концепций, рассматривающих ИИ как инструмент качественного улучшения человеческой жизни, до критических теорий, предупреждающих об угрозах дегуманизации и усиления социального контроля. Проведен сравнительный анализ национальных стратегий развития ИИ ведущих государств мира, выявлены их общие черты и существенные различия, обусловленные культурными, политическими и экономическими факторами. Исследуются потенциальные риски и угрозы, связанные с внедрением систем ИИ в государственное управление, включая проблемы защиты персональных данных, информационной безопасности и этические аспекты использования алгоритмов принятия решений. Рассматривается концепция человекоцентричного подхода к ИИ как возможного принципа, на основе которого должны осуществляться разработка и внедрение технологий ИИ. Охарактеризованы различные уровни контроля над системами ИИ, включая правовое регулирование, профессиональную и общественную экспертизу. Внимание уделяется перспективам развития общего ИИ и его потенциальному влиянию на трансформацию государственных институтов и социальных отношений. Анализируется перспектива создания общего ИИ, его потенциального влияния на трансформацию государственных институтов и социальных отношений. Утверждается, что архитектура общего ИИ, обеспечивающая

* Статья подготовлена по результатам исследований, выполненных за счет бюджетных средств по государственному заданию Финансового университета при Правительстве РФ.

формирование подлинной субъектности системы, должна предполагать уровень, отвечающий за функции актуализации (стратегическое целеполагание, этику и мораль, знания и самоидентификацию). Особое значение придается осознанию системой конечности своего существования как необходимому условию формирования у ИИ осмысленной стратегии существования и этических принципов. В заключении подчеркивается, что по мере развития технологий ИИ возрастает значимость этических норм, ценностных установок и принципа ответственности, которые не могут быть полностью заменены даже самым совершенным регулированием, и указывается на возрастающую значимость фактора взаимного доверия между государством и обществом в условиях, когда системы ИИ предоставляют беспрецедентные возможности для социального контроля.

Ключевые слова: философия искусственного интеллекта, социальная философия, генеративный ИИ, сильный ИИ, общий ИИ, доверенный ИИ, большие языковые модели, цифровая трансформация, социальная трансформация, правовое регулирование ИИ, этика ИИ.

Славин Борис Борисович – доктор экономических наук, профессор кафедры бизнес-информатики Финансового университета при Правительстве РФ.
bbslavin@gmail.com
https://orcid.org/0000-0003-3465-0311

Для цитирования: *Славин Б.Б.* Искусственный интеллект как фактор трансформации государства и общества: поиск баланса между управленческой эффективностью и человекоориентированностью // Философские науки. 2024. Т. 67. № 3. С. 99–122.
DOI: 10.30727/0235-1188-2024-67-3-99-122

Artificial Intelligence as a Factor in State and Society Transformation: Finding Balance between Administrative Efficiency and Human-Centricity*

B.B. Slavin
Financial University under the Government of the Russian Federation, Moscow, Russia

Abstract

The article presents a socio-philosophical analysis of artificial intelligence (AI) integration into public administration systems. The research focuses on

* The article was prepared based on the results of research conducted with budgetary funding under a state assignment at Financial University under the Government of the Russian Federation.

identifying an optimal balance between enhancing administrative efficiency and preserving humanistic values. The author examines diverse perspectives on AI's role in contemporary society, ranging from techno-optimistic concepts that view AI as a tool for qualitative improvement of human life, to critical theories warning of dehumanization risks and increased social control. The paper conducts a comparative analysis of national AI development strategies among leading global powers, identifying their common features and significant differences shaped by cultural, political, and economic factors. Potential risks and threats associated with the implementation of AI systems in public administration are explored, including issues of personal data protection, information security, and the ethical dimensions of algorithmic decision-making. The concept of a human-centered approach to AI is examined as a potential guiding principle for the development and deployment of these technologies. Various levels of control over AI systems are characterized, encompassing legal regulation, professional and public evaluation. Particular attention is given to the prospects of artificial general intelligence (AGI) development and its potential impact on the transformation of state institutions and social relations. The study argues that AGI architecture, enabling genuine system agency, must incorporate a level responsible for actualization functions (strategic goal-setting, ethics, knowledge, and self-identification). Special emphasis is placed on the system's awareness of its finite existence as a necessary condition for developing meaningful operational strategies and ethical principles. The article concludes that as AI technologies advance, the importance of ethical norms, value systems, and responsibility principles increases since these core societal factors cannot be fully replaced even by the most sophisticated regulation. The author highlights the growing significance of mutual trust between state and society in an environment where AI systems provide unprecedented opportunities for social control.

Keywords: philosophy of artificial intelligence, social philosophy, generative AI, strong AI, AGI, trustworthy AI, large language models, digital transformation, social transformation, AI regulation, AI ethics.

Boris B. Slavin – D.Sc. in Economics, Professor of the Department of Business Informatics, Financial University under the Government of the Russian Federation.

bbslavin@gmail.com

https://orcid.org/0000-0003-3465-0311

For citation: Slavin B.B. (2024) Artificial Intelligence as a Factor in State and Society Transformation: Finding Balance between Administrative Efficiency and Human-Centricity. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 67, no. 3, pp. 99–122.
DOI: 10.30727/0235-1188-2024-67-3-99-122

Введение

В последние годы развитие технологий искусственного интеллекта (ИИ) демонстрирует беспрецедентную динамику, оказывая существенное влияние на различные сферы общественной жизни. Особенно значимы эти технологии в контексте государственного управления. В данном случае их внедрение способно качественно трансформировать характер взаимодействия между властью и обществом. Успехи в области ИИ, в частности при создании генеративных моделей, не только открывают новые возможности для повышения эффективности управленческих процессов, но и ставят фундаментальные вопросы этического и социального характера, решение которых следует искать с учетом принципа ответственности перед будущими поколениями [Йонас 2004].

В современном научном дискурсе сформирован широкий спектр оценок влияния ИИ на трансформацию общества. Представители технооптимистического направления (Р. Курцвейл, К. Келли, Д. Амадеи) рассматривают развитие ИИ как путь к качественному улучшению человеческой жизни и решению глобальных проблем человечества. Курцвейл предсказывает наступление технологической сингулярности, при этом ИИ превзойдет человеческий интеллект и откроет беспрецедентные возможности для развития [Kurzweil 2005]. Амадеи пишет о потенциале ИИ для комплексного совершенствования государственного управления: от повышения объективности судебной системы через автоматизацию интерпретации законов до оптимизации процессов агрегации общественного мнения и достижения социального консенсуса. Это будет способствовать преодолению ограничений, связанных с пристрастностью людей, и обеспечит для всех равные политические, правовые и экономические возможности [Amadei 2024].

Вместе с тем ряд исследователей (Н. Бостром, С. Рассел и др.) указывают на возможные экзистенциальные риски для человечества, связанные с потенциальной утратой контроля над сверхразумными формами ИИ [Bostrom 2014; Russell 2019]. Ш. Зубофф описывает угрозу становления «капитализма слежки» (surveillance capitalism), что может привести к усилению контроля со стороны государства и корпораций, подрыву демократических основ общества [Zuboff 2019]. Кроме того, некоторые исследователи обращают внимание на трансформацию государства под влиянием ИИ. Л. Флориди пишет о «четвертой революции», в которой информационные технологии радикально изменяют наше

понимание реальности и социальной структуры, трансформируя государства в распределенные политические мультиагентные системы [Floridi 2014].

Ряд исследователей склоняются к позиции о том, что это – временная реакция в связи с ажиотажем вокруг технологии ИИ. В частности, авторы сравнительного исследования политики стран в области ИИ [Bareis, Katzenbach 2022] делают вывод о том, что разработанные стратегии ИИ это – лишь «политическая риторика о надеждах и страхах», мифология и дискурсы о новых технологиях. Действительно, можно провести исторические параллели, когда новым технологиям приписывали магические или апокалиптические возможности. В середине XIX века, с началом развития железнодорожного транспорта, некоторые верили в то, что с увеличением скорости человеческое тело будет разорвано на куски. Появление первых телефонных аппаратов возбудило критиков, говоривших о возможном нарушении частной жизни или, наоборот, о том, что люди станут ленивыми. Даже Маркони, который изобрел радио, не был уверен в том, принес ли он людям добро или зло. Вполне вероятно, что и страхи перед ИИ через несколько десятилетий будут приводить в пример классической технофобии.

Цель статьи заключается в социально-философском анализе существующих практик использования ИИ государственными службами, выявлении ключевых проблем и противоречий, а также в определении принципов развития отношений между государством и обществом в эпоху ИИ. Особое внимание нами уделено вопросам этического регулирования, обеспечения баланса между технологической эффективностью и гуманистическими ценностями в процессе внедрения ИИ в государственное администрирование и другие сферы.

Государство и технологии искусственного интеллекта: национальные стратегии и примеры использования

В отличие от других цифровых технологий, после успехов глубокого машинного обучения ИИ сразу оказался в центре внимания политиков. Этому способствовали две особенности ИИ. Во-первых, технологии ИИ используют в ряде отраслей экономики (беспилотный транспорт, медицина, финансы и торговля, безопасность, массмедиа и т.п.); во-вторых, многие проекты с использованием ИИ затрагивают интересы широкого круга граждан

(речь идет о системах распознавания лиц, таргетированной рекламе в сети Интернет, автоматической диагностике болезней и др.). Неслучайно большинство стран в 2017–2019 годах разработали и утвердили национальные программы и стратегии развития ИИ.

Китай стал одним из первых государств, разработавших еще в 2017 году Национальную стратегию в области развития технологий ИИ [Hine, Floridi 2024]. В 2019 году аналогичную стратегию разработали и США, назвав ее Национальной инициативой в области ИИ[1]. В таком неформальном соревновании участвует и Россия: Национальная стратегия развития ИИ на период до 2030 года утверждена Президентом РФ[2] в 2019 году, как и американская. Но пока успехи нашей страны не слишком велики. Согласно трекеру активности стран в ИИ[3], на октябрь 2024 года по количеству научных публикаций лидером выступает Китай, опережая США почти в два раза (однако по количеству цитирований он на втором месте), затем следуют США и Индия, отстающая от США также в два раза. Россия находится на 15-м месте. Между тем количество публикаций свидетельствует лишь о количестве ученых, занимающихся проблемами ИИ, а не об уровне развития технологий.

Немецкие исследователи [Bareis, Katzenbach 2022] провели сравнение стратегических программ в области ИИ четырех стран (США, Франции, Германии и Китая) и пришли к выводу о том, что структура, цели и основные положения этих политических документов очень близки. Однако отношение политиков к ИИ несет еще и специфический, характерный для каждой страны оттенок, который включает в себя элементы национальных культур и традиций. Например, в Стратегии развития ИИ в Китае говорится о том, что ИИ изменит общество и мир, а в программном документе США использование технологии ИИ сравнивают с революцией, которая позволит американскому обществу стать более богатым. От развития ИИ в Германии ожидают того, что эта технология позволит трансформировать экономику, поддержать технически структурные изменения. Французские политики связывают с ИИ возможность ренессанса Европы.

---

[1] В рамках этой инициативы создан специальный ресурс на официальном сайте правительства: https://www.ai.gov/.

[2] О развитии искусственного интеллекта в Российской Федерации: указ Президента РФ от 10 октября 2019 года № 490 // Президент России: офиц. сайт. – URL: http://www.kremlin.ru/acts/bank/44731.

[3] Country Activity Tracker: Artificial Intelligence // Emerging Technology Observatory. – URL: https://cat.eto.tech/.

Политическую сущность стратегических документов в области развития ИИ демонстрируют заявления в них о роли и вкладе в мире. Так, в американской стратегии говорится о лидирующей роли США в развитии ИИ и утверждается, что лидерство упрочится в будущем. В китайской стратегии речь идет о том, что ИИ стал объектом мировой конкуренции, а значит, чтобы быть конкурентоспособным, а также для обеспечения кибербезопасности, необходимо развивать технологии ИИ. Аналогичные выводы относительно сравнения политики США и Китая в области ИИ сделаны Э. Хайн и Л. Флориди в работе «Искусственный интеллект с американскими ценностями и китайскими характеристиками» [Hine, Floridi 2024]. Как и немецкие ученые, они связывают политику в области ИИ с традициями. В частности, американские политики, движимые индивидуальной протестантской трудовой этикой, в большей степени полагаются на технологическое совершенствование с использованием ИИ, а китайские политики в рамках конфуцианской этики видят ИИ в качестве инструмента авторитарной политики, но для целей социальной стабильности (известным стал проект социального рейтинга). Европейское правовое регулирование фокусируется на максимальном ограничении возможности использования ИИ, нарушающем права граждан.

Несмотря на огромный интерес к ИИ, использование этой технологии (без учета систем безопасности с распознаванием лиц и объектов) в государственном управлении не настолько распространено, как в бизнесе. Это связано как с более жестким нормативным регулированием, так и с вопросами этики использования ИИ. Наиболее успешно ИИ в государственной деятельности применяется, видимо, для задач прогнозирования [Margetts 2022]. Аналитические правительственные службы используют инструменты анализа данных на основе ИИ для оценки динамики преступности, экономических проблем, урбанизации и т.д. Так, исследование практики применения инструментов анализа данных в Великобритании в 2018 году [Vogl et al. 2020] показало, что 15 % органов самоуправления использовали в своей деятельности именно прогностические модели.

Большинство проектов, использующих ИИ в органах власти, связано с решением множества муниципальных задач с огромным количеством параметров. В качестве примера приведем несколько таких проектов [Yigitcanlar et al. 2021]: агентное моделирование

для изучения возможностей различных оживленных общественных мест (например, Елисейских полей в Париже), которое проводила медиалаборатория Массачусетского технологического института; ИИ-стартап из Нью-Йорка, который с использованием распознавания изображений и обработки естественного языка помогает понять, каким образом планировка различных районов мегаполиса влияет на тех, кто в них живет; система принятия решений в сфере городского дизайна посредством использования изображений Google Street View и генеративно-состязательных сетей, разработанная Мельбурнским университетом; проект Чикагского университета «Массив вещей», который включает в себя сеть интерактивных модульных устройств или узлов, установленных на территории Чикаго для сбора данных в реальном времени об окружающей среде, инфраструктуре и деятельности города.

Интерес представляет анализ того, в каких сферы деятельности государственные службы чаще всего используют технологии ИИ. В работе [Engstrom et al. 2020] исследованы кейсы применения ИИ различными агентствами США. Лидером по количеству кейсов (33) оказались правоохранительные органы. На втором месте (18 кейсов) находится здравоохранение, на третьем – финансовое регулирование (14). Часто технологии ИИ используют в деятельности социальных служб, в сфере регулирования труда и занятости. Последнее место, как это ни странно, оказалось у служб, связанных с образованием. Скорее всего, все бюджеты образовательных организаций в течение последних лет израсходованы на дистанционное обучение, оказавшееся безальтернативным в условиях пандемии. Более половины выявленных в результате упомянутого выше исследования кейсов реализовано IT-специалистами, работающими в службах, без использования сторонних специализированных компаний. Кроме того, большинство задач решают с помощью таких простых методов ИИ, как классификация и регрессия. Более сложные методы применяют пока редко. Кроме того, возникают этические проблемы, связанные с тем, что предоставление или ограничение доступа к услугам для различных социальных групп находится в зависимости от алгоритмов ИИ. Это говорит о том, что использование технологий ИИ в государственном управлении находится еще далеко от зрелого уровня, предполагающего наличие обширной база знаний лучших практик.

Отдельно стоит упомянуть об использовании ИИ в муниципальном управлении совместно с технологией интернета вещей (IoT) [Kankanhalli et al. 2019]. Именно такое сочетание находится в основе концепции «умного» города, которая предполагает сбор и обработку информации с различных устройств, размещенных на городских улицах (с уличных веб-камер, GPS-датчиков на транспорте, светофоров, датчиков погоды и загрязнений окружающей среды и т.п.). Собираемые данные позволяют и предоставить различные уникальные сервисы населению (данные о прибытии транспорта, предупреждения о погодных аномалиях и др.), и сделать город более безопасным с точки зрения борьбы с преступностью, предотвращения экологических катастроф и т.п.

Еще одной технологией, включающей в себя ИИ, которую сегодня планируют использовать при создании государственных сервисов, являются чат-боты [Androutsopoulou et al. 2019]. Обычные чат-боты имеют ограниченное применение, если нужно быстро получить информацию, находящуюся в базе знаний того или иного сервиса. В более сложных ситуациях чат-боты неэффективны, а в ряде случаев даже имеют негативный эффект, если слишком долго пытаются ответить на вопрос, которого нет в алгоритме общения с клиентом, и не переключают на оператора. Чат-боты, использующие генеративный ИИ и технологии поддержки естественного языка (natural language processing – NLP), позволяют имитировать человеческое общение существенно лучше, а также помогают найти ответ на поставленный вопрос, который сформулирован не так, как в базе знаний. Пока такие решения используются лишь в качестве пилотных проектов, однако ожидается расширение их применения.

Как было показано на примере США, наибольшее количество случаев внедрения ИИ после правоохранительных органов наблюдается в системе здравоохранения. В этой связи интересен опыт Китая, в котором, в отличие от США, область здравоохранения во многом остается государственной функцией. В одной из работ [Sun, Medaglia 2019] проанализированы проблемы внедрения ИИ в сфере здравоохранения Китая в семи измерениях (социальном, экономическом, этическом, политико-правовом, организационно-управленческом, связанном с данными, технологическом), причем в разрезе трех групп заинтересованных сторон (руководителей органов власти, врачей и руководителей больниц, менеджеров IT-компаний). Проблемы внедрения раскрыты на примере исполь-

зования в Китае в области врачебной диагностики сервиса Watson, разработанного американской компанией IBM. Исследования показали, что все три группы стейкхолдеров имеют различные позиции по всем измерениям. Это, безусловно, снижает эффективность внедрения ИИ в государственном здравоохранении.

Рассматривая опыт использования ИИ в Китае, нельзя не упомянуть о китайской системе социального кредита (Social Credit Systems – SCS). Она соответствует китайской стратегии развития и использования ИИ, о которой речь шла ранее в настоящей статье. Хотя SCS все еще не внедрена в полном объеме во всех провинциях Китая, она аккумулировала огромный объем информации о личном, финансовом, общественном и даже политическом поведении граждан для построения их социальных показателей, полученных с помощью анализа поведения гражданам (с использованием веб-камер, мониторинга социальных сетей и др.). Гражданам с низким баллом, например, запрещено летать, ездить на поездах, останавливаться в отелях, учиться в престижных школах, получать социальные льготы, работать в правительстве [Xu et al. 2022]. SCS подвергается активной критике в странах Запада (прежде всего ввиду возможности политических репрессий со стороны власти). Тем не менее граждане Китая относятся к этой системе благосклонно, понимая ее как необходимый инструмент регулирования и наведения порядка. Системы типа SCS позволяют реализовать индивидуальный подход во взаимоотношениях власти и общества, поскольку способствуют получению информации о каждом гражданине. Однако сама по себе система социального кредита не является ни плохой, ни хорошей: все зависит от применения данных, полученных с ее помощью. Либо такое использование будет связано только с наказанием (или лишением чего-либо), либо персональные данные будут также в большой мере использовать для оказания помощи людям.

Наверное, чаще всего используют для разработки стратегий в области государственного управления такую технологию, как агентное моделирование, позволяющее имитировать те или иные социальные явления. В частности, Банк Англии моделирует изменение рынка жилья Великобритании в зависимости от различных политических мер. Правительство США оценивает с использованием агентных вычислительных моделей последствия потенциальных катастроф, а правительство Мексики – приоритеты для достижения целей устойчивого развития в рамках программы ООН.

Одна из особенностей технологии ИИ заключается в ее влиянии на большое количество граждан. Сегодня возможности персонализированного подхода существенно возросли в связи с распространением больших языковых моделей. Если говорить об использовании технологий ИИ в политике, то следует указать, что вскоре ИИ сможет не только определять ожидания того или иного избирателя от претендента на выборный пост, но и генерировать нужное сообщение. Через несколько лет, возможно, мы увидим сражения на выборах, в которых генеративный ИИ станет одним из основных видов оружия, как это сегодня происходит с применением беспилотных аппаратов в боевых действиях.

Риски для общества и государства, порождаемые системами искусственного интеллекта

Генеративный ИИ, как и любая технология, имеет не только позитивные возможности (персонализация сообщений, генерация качественного текста и иллюстраций), но и возможности для деструктивных действий. В контексте проблемы нормативного регулирования при использовании ИИ в бизнесе и государственном управлении чаще всего возникают вопросы о правах человека. Не нарушаем ли мы права человека, если без его ведома снимаем на камеру видеонаблюдения или сканируем его социальные сети? По отношению к бизнесу практически во всех странах позиция одинакова: коммерческие организации должны получать согласие у клиента на то, чтобы иметь возможность обрабатывать его данные. Так, в Европе с 2018 года действует специальный Регламент по работе с персональными данными (General Data Protection Regulation – GDPR), позволяющий гражданам отказаться от предоставления персональных данных для маркетингового таргетирования или потребовать удалить информацию о себе (использовать т.н. право на забвение). Аналогичные инструменты регулирования существуют и в России[4].

Однако не только персональные данные оказываются под угрозой. В отчете ОЭСР за 2021 год[5], посвященном использованию

[4] Поправки к Федеральному закону № 149-ФЗ «Об информации, информационных технологиях и о защите информации», Федеральному закону № 152-ФЗ «О персональных данных».

[5] Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges and Implications for Policy Makers // OECD. 2021. August 11. – URL: https://www.oecd.org/en/publications/artificial-intelligence-machine-learning-and-big-data-in-finance_98e761e7-en.html.

технологий ИИ, машинного обучения и больших данных в финансовой области, говорится о рисках применения новых технологий в деятельности банков и бирж. Известны прецеденты, когда сбои в работе программных биржевых роботов, использующих в т.ч. ИИ, приводили к дестабилизации торгов. В отчете приведен целый ряд рисков, связанных с использованием ИИ: недостаточная проработанность моделей; низкое качество данных, на которых проводится обучение нейросетей и т.д. Все это требует внимания со стороны органов, регулирующих финансовую деятельность.

Еще один риск использования ИИ – информационная безопасность. Директор Института системного программирования РАН академик А.И. Аветисян, выступая на заседании президиума РАН с докладом «Кибербезопасность в контексте искусственного интеллекта» [Аветисян 2022], предупредил о том, что в эпоху ИИ риски информационной безопасности многократно возрастают, поскольку вредоносный код может быть встроен в модель ИИ. Это позволит влиять на решения, которые формируются с применением ИИ. Дополнительная возможность для киберпреступников – влиять на ИИ посредством манипуляции данных, на которых обучаются нейронные сети. Скорее всего, у регулирующих органов и служб информационной безопасности возрастет объем работы по мере расширения использования ИИ в управлении.

В одной из работ [Gao et al. 2023] авторами выделено несколько этапов в историческом развитии правительственных открытых данных (open government data – OGD) в мире. На первом этапе (примерно первое десятилетие этого века) разрабатывают инициативы по сбору открытых данных, созданию инструментов доступа к ним. Второй этап (начало второго десятилетия) характеризуется созданием экосистем, позволяющих интегрировать различные данные, как государственные, так и коммерческие. С середины второго десятилетия начинается переосмысление безопасности данных. Во многом это связано с успехами ИИ, позволившего за счет интеллектуального анализа данных получать информацию, которая напрямую в них не хранилась, но могла быть конфиденциальной. Четвертый этап (до конца второго десятилетия) связан с обменом опытом между странами, в т.ч. с развивающимися странами. Новый этап, по мнению ряда авторов, будет заключаться в развитии устойчивости OGD, более широком использовании интернета вещей и ИИ.

Именно поэтому одним из самых громких критиков ИИ сегодня выступает Совет ООН по правам человека. В конце 2021 года Советом подготовлен отчет под названием «Право на неприкосновенность частной жизни в эпоху цифровых технологий»[6]. Верховный комиссар ООН по правам человека М. Бачелет, возглавляющая указанный Совет, прокомментировала появление этого отчета следующим образом: «Искусственный интеллект может быть силой добра, помогая обществу преодолевать некоторые из серьезных проблем нашего времени. Но технологии искусственного интеллекта могут иметь негативные, даже катастрофические последствия, если они используются без должного учета того, как они влияют на права человека»[7]. На рубеже 2022–2023 годов возникли новые риски, связанные уже не непосредственно с правами человека, а с возможностями генеративного ИИ.

Более узкий, технический, подход к снижению рисков от внедрения ИИ заключается в жестком аудите используемых моделей машинного обучения. Такой подход получил название «доверенный ИИ» (trusted AI, или trustworthy AI). Один из инструментов создания доверенного ИИ включает в себя комбинацию ИИ с технологией блокчейн, которая защищает ИИ от возможного взлома и вредоносного кода с использованием криптографии [Sarpatwar et al. 2019]. К сожалению, внедрение новых информационных технологий, в т.ч. и ИИ, приводит к увеличению и киберугроз. Это плата за цифровые удобства. Но альтернативы цифровизации, в частности использованию технологий ИИ, нет.

Помимо риска введения в заблуждение пользователей и напрасно потраченных инвестиций, разработка больших языковых моделей без должного документирования и выверки данных, на которых они обучаются, может приводить к созданию экстремистского или уничижительного для граждан контента. Еще одним риском является возможность обхода авторского права. Достаточно попросить языковую модель изложить чужую идею, и можно ее выдавать за свою, поскольку сгенерированный текст бу-

[6] Annual report of the United Nations High Commissioner for Human Rights. 2021 // United Nations. Office of the High Commissioner for Human Rights. 2022. June 2. – URL: https://www.ohchr.org/en/publications/annual-report/ohchr-report-2021.

[7] Artificial intelligence risks to privacy demand urgent action – Bachelet // United Nations. Office of the High Commissioner for Human Rights. 2021. September 15. – URL: https://www.ohchr.org/en/2021/09/artificial-intelligence-risks-privacy-demand-urgent-action-bachelet.

дет обладать высокой оригинальностью. Наконец, присутствуют риски кибербезопасности. Языковые модели могут генерировать не только связанные тексты, но и коды программ, в т.ч. вирусов. Возможность быстро изменять структуру кода вирусов станет большой проблемой для антивирусных программ, которые выявляют вирусы на основании схожести кода.

В действительности проблема регулирования систем с ИИ видится частью более общей проблемы рисков использования цифровых технологий в обществе. Цифровые технологии ускоряют все процессы в обществе, как позитивные, так и негативные. Появление сети Интернет, с одной стороны, сняло ограничения на коммуникации между людьми, независимо от их местонахождения, с другой – привело к появлению массового фейкового контента. Мобильные телефоны дали возможность человеку быть на связи практически всегда. Вместе с тем они стали удобным инструментом для мошенников, обманывающих граждан. Как и в регулировании ИИ, в создании безопасной цифровой среды необходимо дополнить государственный контроль, который должен быть минимальным, сочетанием профессиональной и общественной экспертизы. Только самоорганизация общества сможет выработать действенные инструменты контроля над новыми технологиями.

Принципы развития технологий искусственного интеллекта в современных условиях и долгосрочной перспективе

Результатом правового регулирования должна быть не приостановка использования ИИ, а ускоренная разработка стандартов такой деятельности. Это снизит риски и сделает более безопасной работу с новыми технологиями. В данном контексте интерес представляют принципы работы с ИИ, сформулированные в 2017 году на конференции, состоявшейся в Асиломаре[8] (Калифорнии), и получившие название асиломарских принципов. В них содержится основная цель работы с ИИ: создание пользы для человечества. К тому же работа должна учитывать возможные риски от внедрения ИИ, предполагать конструктивный диалог между разработчиками и политиками, учитывать права человека. Асиломарские принципы проповедуют общечеловеческий подход,

[8] Принципы работы с ИИ, разработанные на Асиломарской конференции // Future of Life Institute. 2017. 13 сентября. – URL: https://futureoflife.org/open-letter/ai-principles-russian/.

и в этом аспекте они противоречат ряду национальных стратегий в области развития ИИ, в которых предполагается жесткая конкуренция между странами. Кроме того, принципы носят и пацифистский характер. В частности, в них утверждается, что «стоит избегать гонки вооружений в области автономного летального оружия». Становится понятным, что такого рода принципы обречены остаться благими пожеланиями.

Видимо, единственной формой контроля над современными и будущими моделями ИИ, которые многие сегодня называют сильным ИИ, служит привлечение экспертов. При этом эксперты и разработчики должны опираться на интересы обычных граждан, их пользовательский опыт (user experience – UX), что также соответствует асиломарским принципам. Ни экспертный (или меритократический) контроль, ни демократический контроль по отдельности не смогут снизить риски от использования продвинутых моделей ИИ. Обычные люди, не имеющие глубоких технических знаний, не в состоянии сформулировать требования, необходимые для ограничения применения ИИ. Эксперты, наоборот, понимают технические проблемы и способны находить в этом общий язык с коллегами, но не могут и не должны определять ценность тех или иных приложений ИИ, используемых людьми.

Например, для минимизации рисков, связанных с развитием крупных языковых моделей, исследователи [Bender et al. 2021] предлагают ряд мер. Во-первых, необходимо переориентировать исследовательские усилия на тщательное планирование перед началом создания как наборов данных, так и самих систем. Исследовательское время должно рассматриваться как ценный ресурс, который следует направлять на проекты, обеспечивающие более равномерное распределение преимуществ от технологий или, что еще лучше, приносящие пользу исторически маргинализированным группам. Во-вторых, предлагается уделять особое внимание документированию данных, включая мотивацию их отбора и процессы сбора, а также указывать цели, ценности и мотивы исследователей при создании моделей. В-третьих, авторы призывают к переосмыслению исследовательских целей: вместо погони за увеличением размеров моделей и достижением высоких показателей в рейтингах, часто основанных на искусственных задачах, следует сосредоточиться на понимании того, как именно машины решают поставленные задачи и как они будут функционировать

в социально-технических системах. Важным инструментом для этого может стать методология pre-mortem анализа, при которой команда заранее рассматривает возможные причины потенциальных неудач проекта. Кроме того, рекомендуется использовать методы ценностно-ориентированного проектирования (value sensitive design), позволяющие идентифицировать заинтересованные стороны, их ценности и разрабатывать системы с учетом этих ценностей. Особо подчеркивается, что все эти подходы требуют времени и наиболее эффективны на ранних этапах разработки, до того, как исследователи становятся слишком привязанными к своим идеям и менее склонными изменять курс при обнаружении потенциального вреда.

Одним из направлений разработки стандартов работы с ИИ, обеспечивающих безопасность и эффективность за счет экспертизы и UX, служит т.н. человекоцентричный подход к ИИ (human-centered AI – HCAI), изложенный в книге известного американского исследователя в области компьютерных наук Б. Шнейдермана [Shneiderman 2022]. Подход HCAI ставит человека в центр и внимания, и управления. Это связано с тем, что системы с ИИ, использующие машинное обучение, не только помогают решить новые задачи, но и «затрудняют определение возможных точек сбоя». Выходом из такой ситуации является разработка понятных интерфейсов как для пользователей, так и для создателей алгоритмов. Шнейдерман выделяет четыре сферы профессионального и гражданского контроля за ИИ: надежные системы и методы разработки программного обеспечения; стратегическое управление, обеспечивающее безопасность; доверенная и независимая внешняя сертификация; регулирование со стороны государственных органов.

Следует учитывать, что с точки зрения потенциальных возможностей ИИ пока находится на начальном этапе развития. Именно поэтому существующие технологии интеллектуального распознавания, предсказания и поиска получили название слабого ИИ. И наоборот: технологии, которые в будущем смогут полноценно моделировать человеческий интеллект, называют сильным или общим ИИ (artificial general intelligence – AGI) [Advances… 2007]. Ряд исследователей относятся пессимистично к тому, что такие технологии будут созданы в ближайшее время [Fjelland 2020]. Некоторые полагают, что сильный ИИ послужит лишь инструментом в руках человека [Korteling et al. 2021], но не станет субъектом,

поскольку, в отличие от человека, не может нести ответственность за свои решения. Основными аргументами относительно того, почему общий ИИ не может быть создан в обозримом будущем, со времен Х. Дрейфуса [Dreyfus 1972] считали неспособность алгоритмизировать неявные человеческие знания, а также то, что для социализации сильного ИИ потребуется множество моделей. Вместе с тем в последнее время в связи с успехами генеративных моделей ИИ все чаще наделяют субъектностью, фетишизируя его возможности [Дубровский и др. 2022]. В любом случае существует общее мнение о том, что для разработки AGI недостаточно увеличить вычислительные мощности и количество параметров, а необходимо расширить спектр функциональности систем ИИ.

В литературе нет однозначного определения AGI [Artificial… 2007]. Мы будем придерживаться определения общего ИИ как вычислительной системы, реализующей полностью интеллектуальную человеческую деятельность, т.е. равной по когнитивным способностям человеку. Хотя термин «сильный ИИ» появился как противопоставление термину «слабый ИИ», некоторые исследователи [Wang 2019] считают, что неправильно определять AGI как «ИИ человеческого уровня» или как «сильный ИИ». По их мнению, использование термина «AI» в определении AGI сужает подходы к AGI, поскольку AGI может оказаться полностью отличным от AI. Иногда вместо AI используется аббревиатура [Kuusi, Heinonen 2022] ANI (т.е. «искусственный узкий интеллект») для обозначения различий. Однако подобные споры об определении AGI носят в значительной степени схоластический характер. Все так или иначе понимают, что AGI, в отличие от ИИ, должен реализовывать «человеческий уровень» в целом, в полной мере. В литературе обсуждают возможность того, что ИИ способен со временем превзойти человеческий интеллект в целом [Bostrom 2014], и для него даже зарезервирована аббревиатура ASI (т.е. «искусственный сверхинтеллект»). Но это, скорее, удел футурологии, чем практической науки.

Учитывая, что слабый ИИ сегодня демонстрирует уникальные возможности, перспективы сильного ИИ кажутся фантастическими. Гораздо больше возникает этических и философских проблем. Если AGI будет обладать таким же разумом, как и обычный человек, он должен будет обладать и субъектностью, отвечать за свои поступки, а значит, он должен быть встроен в социальную среду, в общество. Вместе с тем вычислительные ресурсы, которые

использует ИИ, превосходят возможности человека, а человек к тому же и смертен. Не появятся ли с возникновением AGI и бессмертные монстры, которые полностью сломают существующее социальное устройство? Пока в отношении этого наблюдается больше гипотез. Например, в одной из работ [Slavin 2023] обосновывается, что сильный ИИ будет гораздо больше похож на человека, чем слабый ИИ: он будет также смертен и будет находиться в равных с человеком условиях с точки зрения времени коммуникаций.

Для обеспечения безопасности и интегрированности в человеческое сообщество общего ИИ предлагается многоуровневая архитектура, в которой высшим уровнем выступает уровень актуализации, играющий ключевую роль в формировании подлинной субъектности системы [Slavin 2023]. На этом уровне реализуются четыре взаимосвязанных функции: стратегическое целеполагание, этика, знания и самоидентификация. Принципиально значимым видится то, что эти функции не просто надстраиваются над находящимися ниже уровнями, а формируются через социальное взаимодействие и языковую коммуникацию. Знания не являются простым накоплением информации в памяти системы, они должны быть получены через социальные коммуникации и верифицированы практикой. Самоидентификация, в свою очередь, связана с осознанием системой собственного места в мире и, что особенно важно, пониманием конечности существования. Именно ограниченность жизненного цикла служит необходимым условием для формирования осмысленной стратегии существования ИИ и его этических принципов, подобно тому, как осознание смертности является важнейшим фактором формирования человеческого сознания и морали.

Заключение

ИИ представляется не просто новой технологией, но фактором, радикально трансформирующим социальную реальность. Его внедрение приводит к изменению характера взаимоотношений между человеком и государством и самими системами ИИ, формируя новую конфигурацию социальных связей и институциональных механизмов. Особенно актуальной эта трансформация становится в контексте развития генеративного ИИ и больших языковых моделей, которые открывают беспрецедентные возможности для автоматизации многих процессов работы различных

государственных служб. В связи с этим особую значимость приобретает профессиональная и общественная экспертиза, как на этапе формирования принципов регулирования ИИ, так и в процессе их практической реализации. Значимы разработка и конкретизация принципов человекоцентричного подхода к ИИ, поскольку без понимания того, что служит основным интересом человека и общества на современном этапе, невозможно выстроить адекватное правовое регулирование использования технологий ИИ государственными органами и в государственном секторе экономики.

Важно понимать, что даже самое совершенное регулирование не способно полностью заменить этические нормы, ценностные установки и моральные императивы, в соответствии с которыми социальные субъекты выстраивают свои взаимоотношения. По мере того, как все более мощные технологические инструменты оказываются в распоряжении государственных структур, коммерческих организаций и отдельных индивидов, возрастает степень ответственности всех участников социальных отношений. Совершенствование технологий должно сопровождаться социально-этическим развитием [Назаретян 2004].

В этом контексте фактор доверия между государством и гражданами приобретает критическое значение. Системы ИИ предоставляют государству беспрецедентные возможности для осуществления эффективного контроля над обществом. Однако ключевой вопрос заключается в том, сможет ли государство использовать эти возможности не только для повышения эффективности управления, но и для повышения уровня доверия к государственным институтам со стороны общества, для создания условий для всестороннего интеллектуального и нравственного развития граждан. На этот вопрос должен быть выработан профессионально обоснованный и социально приемлемый ответ уже сегодня.

Следует учитывать, что если на текущем этапе мы говорим о стратегиях развития и регулирования ИИ как частного направления государственной политики, то в будущем, когда создание общего ИИ станет технически достижимым, потребуется комплексное переосмысление системы государственного и социального планирования в целом. Ключевым фактором этого переосмысления станет необходимость учитывать появление принципиально новых субъектов социальных отношений: искусственных агентных систем, обладающих определенным уровнем автономности и способности к принятию решений. Это потребует

и внедрения новых правовых механизмов, и пересмотра базовых принципов социального взаимодействия, переопределения понятий ответственности, прав и обязанностей в условиях, в которых участниками общественных отношений будут не только люди, но и искусственные субъекты. Человечеству фактически придется взять ответственность не только за себя, но и за ту форму искусственной жизни и искусственного разума, которую оно создаст.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Аветисян 2022 – *Аветисян А.И.* Кибербезопасность в контексте искусственного интеллекта // Вестник Российской академии наук. 2022. Т. 92. № 12. С. 1119–1123.

Дубровский и др. 2022 – *Дубровский Д.И., Ефимов А.Р., Лепский В.Е., Славин Б.Б.* Фетиш искусственного интеллекта // Философские науки. 2022. Т. 65. № 1. С. 44–71.

Йонас 2004 – *Йонас Г.* Принцип ответственности. Опыт этики для технологической цивилизации / пер. с нем., предисл., прим. И.И. Маханькова. – М.: Айрис-пресс, 2004.

Назаретян 2004 – *Назаретян А.П.* Антропогенные кризисы: гипотеза техно-гуманитарного баланса // Вестник Российской академии наук. 2004. Т. 74. № 4. С. 319–330.

Advances… 2007 – Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms / ed. by B. Goertzel, P. Wang. – Amsterdam: IOS Press, 2007.

Amadei 2024 – *Amadei D.* Machines of Loving Grace: How AI Could Transform the World for the Better. – URL: https://darioamodei.com/machines-of-loving-grace.

Androutsopoulou et al. 2019 – *Androutsopoulou A., Karacapilidis N., Loukis E., Charalabidis Y.* Transforming the Communication between Citizens and Government through AI-Guided Chatbots // Government Information Quarterly. 2019. Vol. 36. No. 2. P. 358–367.

Artificial… 2007 – Artificial General Intelligence / ed. by B. Goertzel, C. Pennachin. – Berlin: Springer, 2007.

Bareis, Katzenbach 2022 – *Bareis J., Katzenbach C.* Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics // Science, Technology & Human Values. 2022. Vol. 47. No. 5. P. 855–881.

Bender et al. 2021 – *Bender E.M., Gebru T., McMillan-Major A., Shmitchell S.* On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? // FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. – New York: Association for Computing Machinery, 2021. P. 610–623.

Bostrom 2014 – *Bostrom N.* Superintelligence: Paths, Dangers, Strategies. – Oxford: Oxford University Press, 2014.

Dreyfus 1972 – *Dreyfus H.L.* What Computers Can't Do: A Critique of Artificial Reason. – New York: Harper, 1972.

Engstrom et al. 2020 – *Engstrom D., Ho D., Sharkey C., Cuéllar M.* Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. – New York: NYU School of Law, 2020.

Fjelland 2020 – *Fjelland R.* Why General Artificial Intelligence Will Not Be Realized // Humanities and Social Sciences Communications. 2020. Vol. 7. No. 10. P. 1–9.

Floridi 2014 – *Floridi L.* The Fourth Revolution: How the Infosphere is Reshaping Human Reality. – Oxford: Oxford University Press, 2014.

Gao et al. 2023 – *Gao Y., Janssen M., Zhang C.* Understanding the Evolution of Open Government Data Research: Towards Open Data Sustainability and Smartness // International Review of Administrative Sciences. 2023. Vol. 89. No. 1. P. 59–75.

Hine, Floridi 2024 – *Hine E., Floridi L.* Artificial Intelligence with American Values and Chinese Characteristics: A Comparative Analysis of American and Chinese Governmental AI Policies // AI & Society. 2024. Vol. 39. P. 257–278.

Kankanhalli et al. 2019 – *Kankanhalli A., Charalabidis Y., Mellouli S.* IoT and AI for Smart Government: A Research Agenda // Government Information Quarterly. 2019. Vol. 36. No. 2. P. 304–309.

Korteling et al. 2021 – *Korteling J., Boer-Visschedijk G., Blankendaal R., Boonekamp R., Eikelboom A.* Human- versus Artificial Intelligence // Frontiers in Artificial Intelligence. 2021. Vol. 4. Article 622364.

Kurzweil 2005 – *Kurzweil R.* The Singularity Is Near: When Humans Transcend Biology. – New York: Viking Press, 2005.

Kuusi, Heinonen 2022 – *Kuusi O., Heinonen S.* Scenarios From Artificial Narrow Intelligence to Artificial General Intelligence – Reviewing the Results of the International Work / Technology 2050 Study // World Futures Review. 2022. Vol. 14. No. 11. P. 65–79.

Margetts 2022 – *Margetts H.* Rethinking AI for Good Governance // Daedalus. 2022. Vol. 151. No. 2. P. 360–371.

Russell 2019 – *Russell S.* Human Compatible: Artificial Intelligence and the Problem of Control. – New York: Viking Press, 2019.

Sarpatwar et al. 2019 – *Sarpatwar K., Vaculin R., Min H., Su G., Heath T., Ganapavarapu G., Dillenberger D.* Towards Enabling Trusted Artificial Intelligence via Blockchain // Policy-Based Autonomic Data Governance / ed. by S. Calo, E. Bertino, D. Verma. – Cham: Springer, 2019. P. 137–153.

Shneiderman 2022 – *Shneiderman B.* Human-Centered AI. – Oxford: Oxford University Press, 2022.

Slavin 2023 – *Slavin B.* An Architectural Approach to Modeling Artificial General Intelligence // Heliyon. 2023. Vol. 9. No. 3. Article e14443.

Sun, Medaglia 2019 – *Sun T., Medaglia R.* Mapping the Challenges of Artificial Intelligence in the Public Sector: Evidence from Public Healthcare // Government Information Quarterly. 2019. Vol. 36. No. 2. P. 368–383.

Vogl et al. 2020 – *Vogl T., Seidelin C., Ganesh B., Bright J.* Smart Technology and the Emergence of Algorithmic Bureaucracy: Artificial Intelligence in UK Local Authorities // Public Administration Review. 2020. Vol. 80. No. 6. P. 946–961.

Wang 2019 – *Wang P.* On Defining Artificial Intelligence // Journal of Artificial General Intelligence. 2019. Vol. 10. No. 2. P. 1–37.

Xu et al. 2022 – *Xu X., Kostka G., Cao X.* Information Control and Public Support for Social Credit Systems in China // The Journal of Politics. 2022. Vol. 84. No. 4. P. 2230–2245.

Yigitcanlar et al. 2021 – *Yigitcanlar T., Corchado J., Mehmood R., Li R., Mossberger K., Desouza K.* Responsible Urban Innovation with Local Government Artificial Intelligence (AI): A Conceptual Framework and Research Agenda // Journal of Open Innovation: Technology, Market, and Complexity. 2021. Vol. 7. No. 1. Article 71.

Zuboff 2019 – *Zuboff S.* The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. – New York: Public Affairs, 2019.

REFERENCES

Amadei D. (2024) *Machines of Loving Grace: How AI Could Transform the World for the Better.* Retrieved from https://darioamodei.com/machines-of-loving-grace

Androutsopoulou A., Karacapilidis N., Loukis E., & Charalabidis Y. (2019) Transforming the Communication between Citizens and Government through AI-Guided Chatbots. *Government Information Quarterly.* Vol. 36, no. 2, pp. 358–367.

Avetisyan A.I. (2022) Cybersecurity in the Context of Artificial Intelligence. *Herald of the Russian Academy of Sciences = Vestnik Rossiyskoy akademii nauk.* Vol. 92, no. 12, pp. 1119–1123 (in Russian).

Bareis J. & Katzenbach C. (2022) Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science, Technology & Human Values.* Vol. 47, no. 5, pp. 855–881.

Bender E.M., Gebru T., McMillan-Major A., & Shmitchell S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). New York: Association for Computing Machinery.

Bostrom N. (2014) *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Dreyfus H.L. (1972) *What Computers Can't Do: A Critique of Artificial Reason.* New York: Harper & Row.

Dubrovsky D.I., Efimov A.R., Lepsky V.E., & Slavin B.B. (2022) The Fetish of Artificial Intelligence. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 65, no. 1, pp. 44–71.

Engstrom D., Ho D., Sharkey C., & Cuéllar M. (2020) *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies.* New York: NYU School of Law.

Fjelland R. (2020) Why General Artificial Intelligence Will Not Be Realized. *Humanities and Social Sciences Communications.* Vol. 7, no. 10, pp. 1–9.

Floridi L. (2014) *The Fourth Revolution: How the Infosphere is Reshaping Human Reality.* Oxford: Oxford University Press.

Gao Y., Janssen M., & Zhang C. (2023) Understanding the Evolution of Open Government Data Research: Towards Open Data Sustainability and Smartness. *International Review of Administrative Sciences.* Vol. 89, no. 1, pp. 59–75.

Goertzel B. & Pennachin C. (Eds.) (2007) *Artificial General Intelligence.* Berlin: Springer.

Goertzel B. & Wang P. (Eds.) (2007) *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms.* Amsterdam: IOS Press.

Hine E. & Floridi L. (2024) Artificial Intelligence with American Values and Chinese Characteristics: A Comparative Analysis of American and Chinese Governmental AI Policies. *AI & Society.* Vol. 39, pp. 257–278.

Jonas H. (2004) *The Imperative of Responsibility: In Search of an Ethics for the Technological Age* (I.I. Makhankov, Trans.). Moscow: Airis-press (Russian translation).

Kankanhalli A., Charalabidis Y., & Mellouli S. (2019) IoT and AI for Smart Government: A Research Agenda. *Government Information Quarterly.* Vol. 36, no. 2, pp. 304–309.

Korteling J., Boer-Visschedijk G., Blankendaal R., Boonekamp R., & Eikelboom A. (2021) Human- versus Artificial Intelligence. *Frontiers in Artificial Intelligence.* Vol. 4, article 622364.

Kurzweil R. (2005) *The Singularity Is Near: When Humans Transcend Biology.* New York: Viking.

Kuusi O. & Heinonen S. (2022) Scenarios From Artificial Narrow Intelligence to Artificial General Intelligence – Reviewing the Results of the International Work / Technology 2050 Study. *World Futures Review.* Vol. 14, no. 11, pp. 65–79.

Margetts H. (2022) Rethinking AI for Good Governance. *Daedalus.* Vol. 151, no. 2, pp. 360–371.

Nazaretyan A.P. (2004) Anthropogenic Crises: The Hypothesis of Techno-Humanitarian Balance. *Herald of the Russian Academy of Sciences = Vestnik Rossiyskoy akademii nauk.* Vol. 74, no. 4, pp. 319–330 (in Russian).

Russell S. (2019) *Human Compatible: Artificial Intelligence and the Problem of Control.* New York: Viking.

Sarpatwar K., Vaculin R., Min H., Su G., Heath T., Ganapavarapu G., & Dillenberger D. (2019) Towards Enabling Trusted Artificial Intelligence via Blockchain. In: Calo S., Bertino E., & Verma D. (Eds.) *Policy-Based Autonomic Data Governance* (pp. 137–153). Cham: Springer.

Shneiderman B. (2022) *Human-Centered AI.* Oxford: Oxford University Press.

Slavin B. (2023) An Architectural Approach to Modeling Artificial General Intelligence. *Heliyon.* Vol. 9, no. 3, article e14443.

Sun T. & Medaglia R. (2019) Mapping the Challenges of Artificial Intelligence in the Public Sector: Evidence from Public Healthcare. *Government Information Quarterly.* Vol. 36, no. 2, pp. 368–383.

Vogl T., Seidelin C., Ganesh B., & Bright J. (2020) Smart Technology and the Emergence of Algorithmic Bureaucracy: Artificial Intelligence in UK Local Authorities. *Public Administration Review.* Vol. 80, no. 6, pp. 946–961.

Wang P. (2019) On Defining Artificial Intelligence. *Journal of Artificial General Intelligence.* Vol. 10, no. 2, pp. 1–37.

Xu X., Kostka G., & Cao X. (2022) Information Control and Public Support for Social Credit Systems in China. *The Journal of Politics.* Vol. 84, no. 4, pp. 2230–2245.

Yigitcanlar T., Corchado J., Mehmood R., Li R., Mossberger K., & Desouza K. (2021) Responsible Urban Innovation with Local Government Artificial Intelligence (AI): A Conceptual Framework and Research Agenda. *Journal of Open Innovation: Technology, Market, and Complexity.* Vol. 7, no. 1, article 71.

Zuboff S. (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* New York: Public Affairs.