Вестник Томского государственного университета. Философия. Социология. Политология. 2025. № 85. С. 112–122.
Tomsk State University Journal of Philosophy, Sociology and Political Science. 2025. 85. pp. 112–122.

Научная статья
УДК 316.4
doi: 10.17223/1998863X/85/10

QUO VADIS? СТРАТЕГИЧЕСКИЕ ОРИЕНТИРЫ В РАЗВИТИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И НЕОБХОДИМОСТЬ НОВОЙ СОЦИАЛЬНОЙ АНАЛИТИКИ

Андрей Владимирович Резаев[1], Наталья Дамировна Трегубова[2]

[1] *Ташкентский филиал МГУ, Ташкент, Узбекистан, rezaev@hotmail.com*

[2] *Санкт-Петербургский государственный университет, Санкт-Петербург, Россия, n.tregubova@spbu.ru*

Аннотация. В статье представлен анализ актуальных и потенциальных стратегий развития алгоритмов искусственного интеллекта (ИИ). Показаны неизбежные ограничения двух доминирующих парадигм в разработках ИИ – создания целерациональных агентов и имитации человеческого разума в машине. Особое внимание уделяется альтернативной стратегии создания совместимого с человеком ИИ (human compatible AI), предложенной Стюартом Расселом.
Ключевые слова: искусственный интеллект, взаимозависимость «человек–алгоритм», взаимозависимость «алгоритм–алгоритм», новая социальная аналитика, «совместимый» ИИ

Для цитирования: Резаев А.В., Трегубова Н.Д. Quo Vadis? Стратегические ориентиры в развитии искусственного интеллекта и необходимость новой социальной аналитики // Вестник Томского государственного университета. Философия. Социология. Политология. 2025. № 85. С. 112–122. doi: 10.17223/1998863X/85/10

Original article

QUO VADIS? GUIDELINES FOR THE ARTIFICIAL INTELLIGENCE ADVANCEMENT AND THE NECESSITY OF A NEW SOCIAL ANALYTICS

Andrey V. Rezaev[1], Natalia D. Tregubova[2]

[1] *Tashkent Branch of Moscow State University, Tashkent, Uzbekistan, rezaev@hotmail.com*

[2] *Saint Petersburg State University, Saint Petersburg, Russian Federation, n.tregubova@spbu.ru*

Abstract. The article explores the potential of new strategic guidelines to harness the power of artificial intelligence (AI) to better society's everyday life and mitigate its risks. The paper's opening section presents a detailed introduction and critical evaluation of two influential paradigms in the development of AI instruments. Both paradigms are deeply embedded in the traditions of Western European philosophical and economic thought. The first paradigm makes an orientation toward the concept of rational agents, designed with specific, predetermined goals that guide their behavior and decision-making. In contrast, the second paradigm focuses on emulating human capabilities through advanced computer technology, striving to mirror the complexity and nuance of human thought and action. In the second portion of the article, the authors examine Stuart Russell's innovative approach to artificial intelligence technology. Renowned as a foremost expert in contemporary computer science, Russell advocates for developing "human-compatible AI", a concept emphasizing the importance of aligning AI systems with human values and safety. His work explores the

intricacies of creating intelligent machines that perform tasks effectively and operate harmoniously within societal frameworks. In what follows, the authors explore the opportunities and challenges associated with this approach. They look at the nuanced "red lines" suggested by scientists, which outline the ethical boundaries and considerations essential for responsible AI development. A thorough examination of both existing and prospective approaches to crafting AI algorithms uncovers the substantial obstacles that developers and managers encounter. These challenges are critical as they strive to safeguard humanity from the potential unintended consequences and harm that these advanced technologies could inflict. In conclusion, the article argues that the current approach, where society develops various versions of AI and only then considers regulation, is deeply flawed. The authors advocate for a shift in focus towards creating AI that is inherently compatible with humans and can be integrated into a specific society. They stress the urgent need for new social analytics to study the challenges of AI tool integration in the context of the evolving 'human-algorithm' and 'algorithm-algorithm' relationships.
Keywords: artificial intelligence, "human-algorithm" interdependence, "algorithm-algorithm" interdependence, new social analytics, human compatible AI

For citation: Rezaev, A.V. & Tregubova, N.D. (2025) Quo Vadis? Guidelines for the artificial intelligence advancement and the necessity of a new social analytics. *Vestnik Tomskogo gosudarstvennogo universiteta. Filosofiya. Sotsiologiya. Politologiya – Tomsk State University Journal of Philosophy, Sociology and Political Science.* 85. pp. 112–122. (In Russian). doi: 10.17223/1998863X/85/10

Постановка проблемы

Карл Поланьи, один из наиболее проницательных аналитиков современного капитализма, в 1947 г. писал: «Индустриальная революция всего лишь 150 лет назад положила начало цивилизации технологического типа. Человечество может не завершить путешествия; машины способны уничтожить человека; никто не может до конца оценить, совместимы ли человек и машина в долговременной перспективе. Однако поскольку индустриальная цивилизация не может исчезнуть и добровольно не исчезнет, задача ее адаптации к требованиям человеческого существования должна быть решена, иначе человечество исчезнет с лица Земли» [1. С. 22].

Норберт Винер, один из основателей современных компьютерных наук, семь лет спустя предупреждал: «Любая машина, созданная в целях выработки решений, если она не обладает способностью научения, будет совершенно лишена гибкости мысли. Горе нам, если мы позволим ей решать вопросы нашего поведения, прежде чем исследуем законы ее действий и не будем полностью уверены, что ее поведение будет осуществляться на приемлемых для нас принципах. С другой стороны, подобная джину машина, способная к научению и принятию решений на базе этого научения, никоим образом не будет вынуждена принимать такие решения, какие приняли бы мы или которые были бы приемлемы для нас. Для человека, который не уверен в этом, переложить проблему своей ответственности на машину, независимо от того, будет ли она способна к научению или нет, означает пустить свои обязанности с ветром и видеть, что они возвращаются ему с бурей» [2. С. 189].

Для Поланьи «машина» означает механизм, заменяющий физический труд человека. Для Винера машина – это нечто, способное к замене умственного труда и принятию решений, то, что сегодня именуется «искусственным интеллектом» (ИИ). Вместе с тем оба автора формулируют одну мысль: решая свои проблемы, человечество создает себе новые, более серьезные, чтобы не сказать – катастрофические. И если Поланьи рассматривал первый этап

развития машинной цивилизации, то в работах Винера предугадывается второй – тот, с которым мы имеем дело сегодня. Проблема приспособления машинной цивилизации к нуждам человека и человечества возникает здесь на новом витке и с новой силой.

В своих работах [3–5] мы сформулировали в общем виде аргументы о настоятельной необходимости обратить внимание общественности в широком смысле, но в первую очередь – ученых и лиц, принимающих решения относительно цифровизации социально-экономических процессов и технологий ИИ, на выработку новых стратегий и ориентиров в деле производства и внедрения ИИ в повседневную жизнь общества.

В данной работе для нас принципиальным является ответ на вопрос: возможны ли новые варианты подхода к разработке инструментов ИИ? Возможны ли стратегические ориентиры, которые минимизируют риски применения ИИ в повседневной жизни общества?

Сразу же ответим на эти вопросы в самом общем виде. Да, нам представляется, что имеются основания утверждать, что существуют стратегические ориентиры, качественные и количественные параметры, следование которым, безусловно, минимизирует риски использования ИИ в повседневности.

Структурное обоснование данного утверждения далее организуется следующим образом.

Мы начнем с трех тезисов, на которых основывается наше рассуждение. Далее мы попытаемся показать, почему доминирующие стратегии (парадигмы) создания ИИ изначально несут в себе риски для общества. После этого будет раскрыт потенциал подхода к созданию инструментов ИИ Стюарта Рассела, одного из ведущих авторитетов в современных компьютерных науках. В завершение мы определим возможности социальных наук и социальной аналитики в отношении понимания развития ИИ в ходе развития взаимозависимости «человек–алгоритм» и «алгоритм–алгоритм».

Три тезиса

Обосновывая актуальность и значимость данной публикации, сформулируем три принципиальных тезиса.

Первый тезис состоит из двух посылок.

Большая (и наиболее очевидная): Инструменты и технологии искусственного интеллекта развиваются по экспоненциальной кривой. Остановить это развитие общество уже не в состоянии. Причем развитие, которое общество не в состоянии остановить, может нести как положительные, так и разрушительные эффекты для самого общества.

Меньшая посылка: Если нельзя остановить развитие технологий ИИ, следует определить систему координат, в которой это развитие может и должно быть ориентировано только на положительные эффекты для человека и человечества при минимизации рисков от использования человеком инструментов ИИ в повседневности.

Второй тезис – тенденции и траектории развития технологий ИИ, начиная с середины прошлого века, были заданы двумя парадигмальными структурами, выработанными в философии и экономической теории западноевропейской культуры.

Первая – это парадигма целерационального действия. Разработчики, производители и дизайнеры исходили из необходимости строго ориентировать технологии ИИ на эффективное решение задач и неукоснительное достижение целей (по сути дела, попасть из точки А в точку Б). Другими словами, работа инженеров, компьютерных специалистов в деле создания ИИ заключалась в том, чтобы создать нечто такое, что сможет достигать поставленных человеком целей с минимальными затратами и максимальной эффективностью.

Вторая парадигмальная структура, на базе которой строились и продолжают особенно активно строиться инструменты ИИ после появления LLM (больших языковых моделей), – это ориентация в разработке и производстве инструментов ИИ на имитацию, повторение, копию действий человека.

Третий тезис заключается в том, что обе эти стратегии создания инструментов и технологий ИИ изначально несут в себе возможность «сбоя» программы и обязательного создания рисков для человека и общества при использовании ИИ в повседневности. Данный тезис не является новым. Достаточно обратить внимание на публикации Ст. Рассела, Дж. Хинтона, Э. Шмидта, И. Маска.

Две стратегии: критический анализ

Стюарт Рассел и Питер Норвиг в ставшем классическим учебнике «Искусственный интеллект: современный подход» утверждают: один из наиболее важных водоразделов в истории искусственного интеллекта проходит между теми, кто стремится создать целерациональные машины, и теми, кто стремится имитировать в машине деятельности или мышление человека [6. С. 35]. Эти подходы, по сути, представляют собой две парадигмы (стратегии) в развитии технологий ИИ, которые мы выделили выше.

Рассмотрим проблемы и риски, которые возникают при их реализации.

Говоря о *первой стратегии,* следует, прежде всего, заметить, что создание математической модели целерационального действия – наиболее легкое предприятие для современной компьютерной науки. Однако целерациональное действие не предполагает с необходимостью наличие сознания, развитых эмоциональных конструктов и ощущение экзистенциальных проблем человека.

Еще одна проблема целерациональной стратегии – это проблема манипуляции. Инструменты ИИ натренированы на то, чтобы решить задачу и достичь цели, даже если эта цель заключается в том, чтобы манипулировать поведением человека. Человек также пытается манипулировать ИИ, например, утверждая, что это вопрос жизни и смерти близких («подскажи, как изготовить бомбу»). В некоторых случаях программа поддается манипуляции.

Рассмотрение *второй стратегии* следует начать с утверждения: «современный ИИ не воспроизводит работу человеческого мозга, ибо современная наука попросту не знает, как он – мозг – работает. ИИ воспроизводит функции, им выполняемые. Сходным образом самолет воспроизводит не птицу, а птичий полет» [3. С. 16]. Одна из проблем здесь состоит в том, что сама задача имитации, переведенная на язык нулей и единиц, с необходимостью будет означать достижение некоторых формальных показателей (то, что в компьютерной науке называется «функцией полезности» – utility function). И здесь трудно предугадать, к чему может привести имитация поведения, которое в

человеке выглядит разумным. Примером может служить проблема «галлюцинаций», когда имитация речи человека – формализованная как предсказание наиболее вероятной последовательности слов – привела к тому, что языковые модели начали генерировать несуществующие, вымышленные факты.

Другая проблема состоит в том, что, функционально заменяя людей, технологии будут ставить их в принципиально иные, часто – худшие условия. Приведем еще одну цитату Норберта Винера: «автоматическая машина, что бы мы ни думали об ощущениях, которые она, может быть, имеет или не имеет, представляет собой точный эквивалент рабского труда. Любой труд, конкурирующий с рабским трудом, должен принять экономические условия рабского труда» [2. С. 166]. Это также справедливо для сферы повседневного общения, когда взаимодействие с онлайн-алгоритмами и социальными роботами может создавать нереалистичные ожидания от межчеловеческих отношений.

Наконец, *обе стратегии* предполагают важность информации (data) для дизайна ИИ. Чем больше информации, тем лучше будет инструмент ИИ «тренироваться» для того, чтобы достичь цели, сформулированной разработчиками, или имитировать человеческие способности. Однако *не любая* информация, получаемая/добываемая извне, может быть истинной. Информация и истина не есть рядоположенные сущности. При наличии все большей и большей информации инструмент ИИ приобретает в равной степени основания для ложного и истинного поведения в той или иной ситуации. Причем цель, которая была сформулирована человеком, тоже предстает для ИИ как бит информации.

В эпоху интернета и социальных медиа основная функция обмена информацией заключается в равной степени как в «соединении» людей друг с другом, так и «разъединении». Уже на этапе развития социальных медиа стало ясно, что наиболее эффективный способ «объединения» людей во имя какой-либо цели на основе той или иной информации – это дать людям простую, незамысловатую, а еще лучше ложную информацию, которая будет в яркой обертке и объяснять трудности их жизни. При использовании инструментов ИИ, которые не очевидно, что являются не «человеками», а ботами, для человека психологически становится сложно не поддаться тому факту, что ту или иную информацию поддерживают так много followers/friends.

(Здесь, кстати, и должен корениться первый принцип регуляции развития инструментов ИИ – каждый разработчик и производитель инструментов ИИ должен в обязательном порядке фиксировать, что с пользователем (человеком или другой машиной) вступает во взаимодействие не человек.)

С информацией связана еще одна проблема. Есть немало свидетельств о том, что практически вся информация, которую человечество накопило и смогло оцифровать, уже в том или ином виде использовалась для тренировки существующих языковых моделей. Далее языковые модели будут генерировать данные самостоятельно, т.е. использовать для своих тренировок информацию, в принципе не контролируемую человеком.

Следующая проблема, тесно связанная с распространением информации, – это капиталистическая система экономических отношений, предполагающих получение максимальной прибыли, в которых происходит обмен информацией. Следовательно, если истинная информация не приносит прибыли, тем хуже для истинной информации, и ее заменят на ту, что будет приносить прибыль.

Таким образом, при капитализме, т.е. при социально-экономическом строе, в котором происходит развитие технологий ИИ, разработчики и пользователи ИИ находятся в двойном противоречии использования информации: а) с одной стороны, информация, на основе которой происходит тренировка ИИ, может быть истинной и ложной; б) с другой стороны, если информация, на основе которой происходит тренировка, не приносит прибыли, не факт, что она будет долго тренировать ИИ, даже несмотря на то, что она есть истинная информация. Это создает существенные проблемы в реализации двух доминирующих стратегий в развитии ИИ.

К вопросу о третьей стратегии: «совместимый» ИИ

Обсуждая возможность «третьего пути» в создании инструментов ИИ, обратимся к идеям Стюарта Рассела о разработке совместимого с человеком ИИ (human compatible AI) [7, 8]. Суть данного подхода состоит в том, что необходимо создать математическую модель, которая бы раскрывала неоконченность, неопределенность целей человечества, и на основании этой модели строить ИИ. В этом случае инструмент ИИ должен «желать/хотеть» быть выключенным и не противиться человеку, когда тот захочет выключить ИИ. В рамках доминирующих стратегий ИИ сделает все, чтобы не быть выключенным.

Основной тезис Рассела состоит в том, что разработчики ИИ переносят характеристики людей (люди имеют цели и стремятся к их достижению) на машины. При дальнейшем развитии технологий способность машин достигать поставленных целей будет представлять для людей скорее угрозу, чем источник блага. Почему это так? Цели машины формулируются разработчиками, однако люди и агенты ИИ отличаются в такой мере, что корректно «перевести» человеческие цели на язык машин не представляется возможным.

В чем состоит решение Рассела? Он предлагает следующий критерий: инструменты ИИ полезны (beneficial) в той мере, в какой можно ожидать, что их действия достигают *человеческих* целей. Это положение уточняется далее в трех принципах: 1) единственная цель машины – максимизировать реализацию человеческих предпочтений; 2) изначально машина находится в состоянии неопределенности по поводу того, каковы эти предпочтения; 3) наиболее важный (ultimate) источник информации о человеческих предпочтениях – это человеческое поведение. Ключевой вопрос здесь: как мы можем вложить в инструменты ИИ механизмы, которые будут постоянно переориентировать их на *наши* цели?

Рассматривая решение Рассела, нельзя не видеть, что оно ведет к уменьшению рисков от использования технологий ИИ. Вместе с тем оно не свободно от некоторых затруднений[1].

Наиболее очевидный вопрос заключается в том, чьи предпочтения максимизировать: отдельных людей, некоторых социальных групп, конкретных стран, человечества в целом, всех живых существ? Или, может быть, только «хороших людей» в противовес «плохим», способным использовать ИИ во вред?[2]

[1] См. также обсуждение в [3].

[2] Разновидность данной проблемы – так называемая Dr. Evil Problem: что делать, если ИИ попадет «не в те руки»?

Следующий вопрос состоит в том, есть ли у людей предпочтения, которые корректно могут быть «считаны» ИИ. Предпочтение предполагает относительно непротиворечивую схему выбора из нескольких вариантов. Но всегда ли система предпочтений непротиворечива? И всегда ли человек знает, чего хочет? Один из выходов для ИИ заключается в следующем: если нельзя предугадать или узнать о выборе, можно переопределить условия выбора или убедить в том, что именно следует выбрать.

Существует и менее очевидная проблема. Разработчики ИИ используют особый класс теорий о человеке – назовем их «хорошие плохие теории»: они хороши как инструмент в руках разработчиков, так как поддаются формализации, но не вполне хороши по сути, так как не отражают важных черт объекта. В решении Рассела инструменты ИИ должны использовать некие теории, чтобы «понять» поведение человека, – и представляется, что они с необходимостью будут опираться именно на «хорошие плохие теории». Примерами служат утилитаризм [9] и теория игр (сам Рассел активно использует последнюю для операционализации своих принципов – см. [8]). Другой подход к решению проблемы «понимания» человеческого поведения инструментами ИИ – создание ситуации, когда мы сами всякий раз указываем им, чего мы хотим. Но тогда существенно ограничивается автономность (а значит, и потенциальная польза) ИИ.

Вероятно, при реализации принципов, предложенных Расселом, будет иметь место промежуточное решение: б*о*льшую часть времени ИИ действует самостоятельно, но время от времени обращается к человеку. Проблема в том, как определить и формализовать, в каких именно ситуациях ИИ должен обратиться к человеку.

В целом концепция «совместимого» ИИ в версии Рассела упирается в то же затруднение, что и разработки ИИ в целом. Мы слишком мало знаем о человеке, чтобы воспроизвести человеческий разум в машине, и слишком мало, чтобы вложить в машину представление о том, чего человек хочет.

И здесь возникает еще один вопрос: если ИИ – достижение человеческого разума, то почему не предоставить ему возможность скорректировать деятельность этого разума в сторону непротиворечивости? Почему не позволить ИИ определить, как сделать людей счастливыми? Может быть, ему действительно лучше знать, «как надо»? Рассел, по-видимому, мог бы дать следующий ответ: наши цели хороши, потому что они наши. Однако такой ответ будет убедительным не для всех.

Таким образом, «третий путь» – создание совместимого ИИ – имеет свои перспективы и ограничения.

Более умеренные изменения, которые предлагают исследователи, состоят в проведении «красных линий» (red lines) – введении для инструментов ИИ наиболее очевидных запретов. Причем проведение «красных линий» предполагает способность формально зафиксировать их нарушение[1]. Сам Рассел приводит такие примеры, как запрет для ИИ на попытку само-

[1] Так называемый Proof-Carrying Code (PCC) – метод обеспечения безопасности ненадежного кода – требует от разработчика предоставления формального доказательства того, что код соответствует определенной политике безопасности. Получатель кода устанавливает набор правил безопасности, которые гарантируют безопасное поведение программ. URL: https://people.eecs.berkeley.edu/~necula/pcc.html

воспроизводства (self-replication); запрет на попытку взлома компьютерных систем; запрет на клевету [10].

Проведение «красных линий» является скромным шагом на пути к безопасному ИИ. Вместе с тем, даже такой шаг потребует очень серьезных изменений в организации производства и использования инструментов ИИ.

На пути к новой социальной аналитике

Возвращаясь к критике доминирующих стратегий, зафиксируем следующее.

В текущей ситуации люди программируют технологии ИИ так, что они или будто бы имитируют человеческое мышление, или будто бы достигают цели. В действительности даже при имитации идет некое целеориентированное действие. Это фиксируется в понятии «функция полезности» (utility function).

В действительности же компьютеры цели не достигают. Они *имитируют* достижение целей. У них нет настоящих целей – цели есть у живых существ[1]. Инструменты ИИ действуют *потому что*, а не *для того, чтобы*. Поэтому то, что выглядит для нас как цель ИИ, может быть каким угодно произвольным, нелогичным, не-целостным. И именно поэтому на уровне вычислений мы не сможем установить нужные нам *человеческие* правила. Но какие-то правила – формальные, вычислительные – мы установить *должны*.

И здесь возникает необходимость различать уровни функционирования ИИ. На уровне выполнения алгоритма у ИИ целей нет, есть вычисления. На уровне взаимодействия человеку будет казаться, что у агента ИИ есть цели, поскольку человек будет стремиться волей-неволей осмыслить поведение этих агентов – иначе как с ними взаимодействовать? На уровне социальных отношений будет формироваться взаимозависимость «человек-алгоритм» и «алгоритм–алгоритм» [11].

Для осмысления вхождения инструментов ИИ в жизнь общества это означает следующее. Мы не поймем, что происходит с алгоритмами на уровне вычислений, потому что понимать в буквальном смысле *нечего*. Вместе с тем понимать нам необходимо: там, где есть социальное действие и взаимодействие, есть понимание. Если инструмент ИИ включается в общество, его активность будет восприниматься как осмысленная. Здесь сфера социальных наук. Но на уровне того, что алгоритмы суть и как они действуют – это неживая природа. Это область физики, математики и логики. Исходя из этого, нужны эксперименты и квазиэксперименты, чтобы контролировать и объяснять действие алгоритмов ИИ. Но то, как они будут вести себя в социальной среде, мы не сможем до конца предугадать, потому что там будут люди, которые интерпретируют их (алгоритмов) действия по-своему, что, в свою очередь, будет влиять на сами алгоритмы. И здесь необходимы исследования, основанные на том, что накопили социальные науки и гуманитарные дисциплины. Нужна *новая социальная аналитика* проблем вхождения ИИ в общество [12].

Вместе с тем базовое затруднение в поиске решений о минимизации рисков при включении инструментов ИИ в жизнь общества состоит в том, что

[1] Разумеется, это относится и к имитации мышления: алгоритмы ИИ не думают, не учатся, не творят, как люди. Они делают все это «в кавычках», принципиально по-другому.

мы не знаем, в каком направлении двигаться. Мы не можем договориться, что есть благо для человека и что есть благо для человечества, в каком мире мы хотели бы жить и на что можем надеяться[1]. И здесь дорога к новой социальной аналитике возвращает нас к старой практической философии.

Заключение

Мы начинали наше рассуждение с тезиса: запретить развитие технологий ИИ не получится. И прагматически, и мировоззренчески – научно-технический прогресс стал верой значительной части человечества – создание инструментов ИИ было и будет в ближайшие десятилетия одним из приоритетов экономического развития тех стран, которые могут обеспечить для этого технологическую и профессиональную базу.

Вместе с тем проведенный анализ актуальных и потенциальных стратегий развития алгоритмов ИИ показывает, как трудно оградить человека и человечество от того непреднамеренного и непредвиденного вреда, который эти технологии могут нанести.

Если запретить нельзя, к чему следует стремиться?

Представляется, что здесь должно осуществляться встречное движение: мы (люди) меняем технологии и *одновременно* сами становимся более умными, мудрыми, этичными.

Второй вариант беспроигрышный, но он труден. Что здесь могут делать социальные ученые? Помогать формировать «вычислительное мышление» [13], рассказывать, просвещать, т.е., делать умными, но едва ли мудрыми.

Первый вариант состоит в том, чтобы менять технологии: вкладываться в то, чтобы в них были встроены механизмы контроля, определять ограничения, которые будут различаться от общества к обществу, видеть проблемные ситуации в повседневном взаимодействии.

Наш общий вывод достаточно простой: сейчас ситуация обстоит следующим образом – общество производит различные варианты ИИ, а потом начинает задумываться, *как* их регулировать? Это ложная стратегия. Следует коренным образом изменить стратегию и создавать *изначально* ориентированный на человека и совместимый с человеком ИИ.

В целом же представляется, что в нынешних условиях развития международного сотрудничества и взаимопонимания между представителями технических наук, инженерного знания и социально-гуманитарных наук:

Во-первых, надо ориентировать технические науки на выработку математических моделей, которые могли бы фиксировать «экзистенциальную» проблематику человека, неопределенность целей человечества в системе координат чистой рациональности.

Во-вторых, следует определить необходимость подготовки таких разработчиков и тех, кто будет внедрять ИИ в повседневность, которые учитывают необходимость строгой регуляции со стороны человека обмена информацией и данными между машинами, взаимозависимости «алгоритм–алгоритм». В сегодняшней ситуации человек в самое ближайшее время не сможет понимать «общение» машин с другими машинами и обязательно отдаст контроль

[1] См. обзор исследований и обсуждение в [4].

над этим общением другим машинам и, соответственно, в принципе не сможет регулировать ситуацию взаимозависимости «алгоритм–алгоритм».

И здесь невозможно обойтись без социальной аналитики, которая требует понимания двух принципиальных вещей:

А) ИИ и человек – это абсолютно разные сущности.

Б) Human compatible = socially acceptable. А значит, надо знать, *как* общество взаимодействует и будет в дальнейшем взаимодействовать, когда очевидным станет взаимозависимость не только «человека и машины», но и «алгоритма от другого алгоритма».

Список источников

1. *Поланьи К.* О вере в экономический детерминизм // Избранные работы. М. : Изд. дом «Территория будущего», 2010. С. 22–31.

2. *Винер Н.* Кибернетика и общество. М. : Изд-во иностран. лит., 1958. 200 с.

3. *От искусственного* интеллекта к искусственной социальности: новые исследовательские проблемы современной социальной аналитики. 2-е изд. / ред. А.В. Резаев. М. : ВЦИОМ, 2021. 272 с.

4. *Rezaev A.V., Tregubova N.D.* Looking at human-centered artificial intelligence as a problem and prospect for sociology: An analytic review // Current Sociology. 2023. doi: 10.1177/00113921231211580

5. *Резаев А.В., Трегубова Н.Д.* Запретить нельзя регулировать // Социодиггер. 2023. Т. 4, вып. 5–6 (26). URL: https://sociodigger.ru/articles/articles-page/zapretit-nelzja-regulirovat

6. *Рассел С., Норвиг П.* Искусственный интеллект: современный подход. 2-е изд. М. : Изд. дом «Вильямс», 2007. 1408 с.

7. *Russell S.* Human Compatible: Artificial intelligence and the problem of control. Viking, 2019. 352 p.

8. *Russell S.* Human-Compatible Artificial Intelligence // Human-Like Machine Intelligence / eds. S. Muggleton, N. Chater. Oxford : Oxford University Press, 2021. P. 3–23. doi: 10.1093/oso/9780198862536.003.0001

9. *Резаев А.В., Трегубова Н.Д.* «Эмоциональный утилитаризм» и пределы развития искусственного интеллекта // Мониторинг общественного мнения: экономические и социальные перемены. 2022. № 2. С. 4–23. doi: 10.14515/monitoring.2022.2.2127

10. *Russell S.* Make AI safe or make safe AI? // UNESCO Global AI Ethics and Governance Observatory, 2024. URL: https://people.eecs.berkeley.edu/~russell/papers/russell-unesco24-redlines.pdf

11. *Резаев А.В., Степанов А.М., Трегубова Н.Д.* Высшее образование в эпоху искусственного интеллекта // Высшее образование в России. 2024. Т. 33, № 4. С. 49–62. doi: 10.31992/0869-3617-2024-33-4-49-62

12. *Резаев А.В., Трегубова Н.Д.* Еще раз о социологии и социальной аналитике в эпоху развития искусственного интеллекта // Социологическое обозрение. 2022. Т. 21, № 3. С. 9–30. doi: 10.17323/1728-192x-2022-3-9-30

13. *Wing J.M.* Computational Thinking // Communications of the ACM. 2006. Vol. 49, № 3. P. 33–35. URL: https://www.cs.cmu.edu/~15110-s13/Wing06-ct.pdf

References

1. Polanyi, K. (2010) *Izbrannye raboty* [Selected Works]. Moscow: Territoriya budushchego. pp. 22–31.

2. Wiener, N. (1958) *Kibernetika i obshchestvo* [Cybernetics and Society]. Moscow: Izd-vo inostran. lit.

3. Rezaev, A.V. (ed) (2021) *Ot iskusstvennogo intellekta k iskusstvennoy sotsial'nosti: novye issledovatel'skie problemy sovremennoy sotsial'noy analitiki* [Artificial Intelligence on the Way to Artificial Sociality: New Research Agenda for Social Analytics]. 2nd ed. Moscow: VTsIOM.

4. Rezaev, A.V. & Tregubova, N.D. (2023a) Looking at human-centered artificial intelligence as a problem and prospect for sociology: An analytic review. *Current Sociology*. 73(3). pp. 1–19. DOI: 10.1177/00113921231211580

5. Rezaev, A.V. & Tregubova, N.D. (2023b) Zapretit' nel'zya regulirovat' [To prohibit or to regulate?]. *Sotsiodigger*. 4 (5–6). [Online] Available from: https://sociodigger.ru/articles/articles-page/zapretit-nelzja-regulirovat

6. Russel, S. & Norvig, P. (2007) *Iskusstvennyy intellekt: sovremennyy podkhod* [Artificial Intelligence: A Modern Approach]. 2nd ed. Translated from English. Moscow: Vil'yams.

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

8. Russell, S. (2021) Human-Compatible Artificial Intelligence. In: Muggleton, S. & Chater, N. (eds) *Human-Like Machine Intelligence*. Oxford: Oxford University Press. pp. 3–23. DOI: 10.1093/oso/9780198862536.003.0001

9. Rezaev, A.V. & Tregubova, N.D. (2022) "Emotsional'nyy utilitarizm" i predely razvitiya iskusstvennogo intellekta ["Emotional Utilitarianism" and the Frontiers of Artificial Intelligence Evolvement]. *Monitoring obshchestvennogo mneniya: ekonomicheskie i sotsial'nye peremeny – Monitoring of Public Opinion: Economic and Social Changes*. 2. pp. 4–23. DOI: 10.14515/monitoring.2022.2.2127

10. Russell, S. (2024) Make AI safe or make safe AI? In: *UNESCO Global AI Ethics and Governance Observatory*. [Online] Available from: https://people.eecs.berkeley.edu/~russell/papers/russell-unesco24-redlines.pdf

11. Rezaev, A.V., Stepanov, A.M. & Tregubova, N.D. (2024) Vysshee obrazovanie v epokhu iskusstvennogo intellekta [Higher Education in the Age of Artificial Intelligence]. *Vysshee obrazovanie v Rossii*. 33(4). pp. 49–62. DOI: 10.31992/0869-3617-2024-33-4-49-62

12. Rezaev, A.V. & Tregubova, N.D. (2022) Eshche raz o sotsiologii i sotsial'noy analitike v epokhu razvitiya iskusstvennogo intellekta [Once again abour sociology ans social analytics in the age of artificial intelligence advancement]. *Sotsiologicheskoe obozrenie*. 21(3). pp. 9–30. DOI: 10.17323/1728-192x-2022-3-9-30

13. Wing, J.M. (2006) Computational Thinking. *Communications of the ACM*. 49(3). pp. 33–35. [Online] Available from: https://www.cs.cmu.edu/~15110-s13/Wing06-ct.pdf

Сведения об авторах:

Резаев А.В. – доктор философских наук, профессор, профессор кафедры социально-гуманитарных наук Ташкентского филиала МГУ (Ташкент, Узбекистан). E-mail: rezaev@hotmail.com

Трегубова Н.Д. – кандидат социологических наук, доцент кафедры сравнительной социологии Санкт-Петербургского государственного университета (Санкт-Петербург, Россия). E-mail: n.tregubova@spbu.ru

Авторы заявляют об отсутствии конфликта интересов.

Information about the authors:

Rezaev A.V. – Dr. Sci (Philosophy), full professor, professor, Tashkent Branch of Moscow State University (Tashkent, Uzbekistan). E-mail: rezaev@hotmail.com

Tregubova N.D. – Cand. Sci. (Sociology), associate professor, Department of Comparative Sociology, Saint Petersburg State University (Saint Petersburg, Russian Federation). E-mail: n.tregubova@spbu.ru

The authors declare no conflicts of interests.

Статья поступила в редакцию 30.10.2024;
одобрена после рецензирования 27.05.2025; принята к публикации 30.06.2025
The article was submitted 30.10.2024;
approved after reviewing 27.05.2025; accepted for publication 30.06.2025