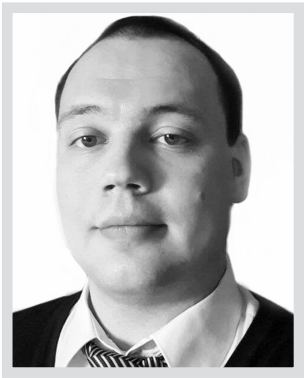
УДК 316/004

Особенности механизмов регулирования искусственного интеллекта в условиях новой социальности



Обидин Д.С.
Магистрант факультета социальных наук
Национального исследовательского Нижегородского
государственного университета им. Н.И. Лобачевского

Цель статьи – обосновать необходимость создания механизмов регулирования искусственного интеллекта (далее – ИИ) со стороны социальных наук, в частности, социологии. С этой целью проведен анализ существующих определений ИИ и предложено различать ИИ как концепцию и ИИ как технологию, требующую регулирования. Описаны характеристики ИИ, которые представляют его потенциальным общественным риском и которые нуждаются в контроле и регулировании. В заключении подчеркивается необходимость разработки эффективной междисциплинарной методологии исследования и понимания сложных и всесторонних проблем ИИ в реальных условиях.

Ключевые слова: искусственный интеллект, характеристики ИИ, машинная технология, механизм регулирования, технологическая индустрия, социальные риски, проблемы управления

Актуальность темы исследования обусловлена тем, что искусственный интеллект (далее – ИИ) быстрыми темпами становится частью нашей повседневной жизни, предлагая множество преимуществ для общества. Растущее повсеместное распространение и стремительный рост коммерческого потенциала ИИ стимулировали массовые инвестиции частного сектора в проекты искусственного интеллекта. Такие фирмы, как *Google, Facebook, Amazon* вступили в «гонку вооружений» искусственного интеллекта, борьбу за создание лабораторий и покупку стартапов. В то же время есть опасения по поводу непредсказуемости и неконтролируемости ИИ.

Искусственный интеллект – междисциплинарная технология, которая направлена на использование больших наборов данных (*Big Data*), подходящих вычислительных мощностей, а также специальных аналитических процедур и процедур принятия решений для того, чтобы позволить компьютерам выполнять задачи, приближенные к человеческим способностям, а в некоторых отношениях даже превосходящие их [1].

Приложения на основе ИИ могут выполнять определенные задачи, и этот тип технологии искусственного интеллекта известен как узкий ИИ (или слабый ИИ). Общий искусственный интеллект (ОИИ или сильный ИИ) может одновременно поддерживать многозадачность и считается интеллектом, который способен превзойти или даже заменить человеческий интеллект. Эксперты расходятся во мнениях относительно того, как скоро ОИИ станет реальностью. Например, ряд исследователей ИИ считают, что ОИИ имеет 50 % шансов быть разработанным в период между 2040 и 2050 гг. и 90 % – к 2075 г. Хотя есть ряд экспертов в области искусственного интеллекта, считающих, что до ОИИ еще несколько веков [1].

Но каковы бы ни были прогнозы ученых, уже сегодня появился ряд серьезных проблем, вызванных применением ИИ, поскольку искусственный интеллект – это не только технологии, он выстраивает новую социальность. Потенциал для дальнейшего быстрого развития технологий ИИ пробудил тревогу различных специалистов, в том числе привел к призывам к государственному регулированию разработок ИИ и ограничению операций. Само по себе это неудивительно; страх перед технологическими изменениями и призывы регулировать новые технологии – явление не новое. Такая же ситуация существовала и в первые дни Интернета. Предыду-

щий опыт управления, безусловно, поможет решить сегодняшние проблемы с ИИ, но как новое явление он, скорее всего, создаст определенные проблемы управления и регулирования, которые не встречались на ранних этапах развития технологий.

Таким образом, **цель** данной статьи – обосновать необходимость внедрения механизмов регулирования ИИ со стороны социальных наук, в частности, социологии.

Задачи:

– проанализировать существующие определения искусственного интеллекта для возможности дать рабочее определение ИИ, нуждающегося в регулировании;

– описать характеристики ИИ, которые представляют ИИ потенциальным общественным риском и которые нуждаются в контроле и регулировании со стороны социума;

– доказать необходимость регулирования ИИ, сделав акцент на различиях ИИ и предыдущих, ставших уже традиционными, источниках общественного риска;

– показать, что несмотря на ряд проблем, эффективное регулирование ИИ тем не менее должно быть возможным и даже необходимым.

Следует отметить, что проблема изучения ИИ стала уже традиционной для зарубежных исследований, в то время как в России она только выходит на повестку дня. По справедливому замечанию А.В. Резаева и Н.Д. Трегубовой, «в массиве академических публикаций об ИИ сегодня доминирует "большая тройка" – компьютерные науки, когнитивные науки и философия» [3]. И если в области технологий, компьютерного моделирования и точных наук исследования ИИ справедливо занимают свое место среди актуальнейших вопросов, обсуждаемых на крупнейших международных конференциях и на страницах ведущих научных журналов, то в области социальных наук проблема ИИ пока еще только формулируется. Главная и несомненная заслуга принадлежит здесь санкт-петербургским ученым А.В. Резаеву и Н.Д. Трегубовой [4], которые обозначили основные проблемы и задачи в области ИИ для социальных наук и систематизировали имеющиеся на сегодняшний день исследования в данной области.

Теоретической базой исследования стали труды крупнейших специалистов в области ИИ, в частности С. Рассела и П. Норвига [2], А. Вулфа [5], М. Швейца [6], Р. Коллинза [7], Ф. Хаббарда [8], А.В. Резаева, Н.Д. Трегубовой [3, 4] и др.

Практическая значимость исследования связана с дальнейшей разработкой методологии исследования механизмов регулирования ИИ социальными науками в целом, социологией, в частности.

Приступая к анализу заявленной проблемы, необходимо отметить, что ИИ может помочь нам в нашей работе и жизни и освободить от рутинных и монотонных задач. Но в то же время ИИ – палка о двух концах.

Питер Хубер ввел термин «общественный риск» для описания угроз здоровью или безопасности человека, которые «производятся централизованно или массово, широко распространяются и в значительной степени находятся вне прямого понимания и контроля отдельного носителя риска» [9, p. 277]. Первые комментаторы общественного риска сосредоточились в первую очередь на ядерных технологиях, экологических угрозах и т.д. Растущее распространение ИИ почти уверенно свидетельствует, что системы искусственного интеллекта будут генерировать много рисков для общества и человека. Эти риски могут оказаться сложными для регулирования, потому что ИИ представляет проблемы, не связанные с общественными рисками предыдущих столетий. Тем не менее необходимо найти механизмы, которые могут помочь снизить общественные риски, связанные с ИИ даже перед лицом уникальности ИИ.

Любой механизм регулирования ИИ должен определять, что именно он должен регулировать; другими словами, он должен определять искусственный интеллект. К сожалению, пока нет четкого определения ИИ даже среди специалистов в данной области, не говоря уже о рабочем определении ИИ для целей регулирования. ИИ считается общим термином, обозначающим широкий спектр дисциплин и методов. Машинное обучение, автоматизация и робототехника имеют отношение к технологиям искусственного интеллекта или принадлежат к ним. ИИ в общих чертах подразделяют на две категории: слабый ИИ и сильный ИИ. Сильный ИИ – очень дискуссионная тема, и некоторые исследователи считают сильный ИИ реальной угрозой для людей. Существует множество слабых приложений ИИ, но и внедрение слабого ИИ уже привело к ряду управленческих и этических проблем. Например, к проблемам безопасности и конфиденциальности.

Мы не ставим перед собой задачи создать новое определение ИИ, но предпримем попытку проанализировать проблемы, с которыми придется столкнуться при регулировании процесса использования ИИ.

Сложность определения искусственного интеллекта заключается не в концепции искусственности, а, скорее, в концептуальной неоднозначности интеллекта. Потому что люди – единственные существа, которые повсеместно признаны (по крайней мере, среди людей) как обладающие интеллектом, поэтому неудивительно, что определения интеллекта связаны с человеческими характеристиками. Дж. Маккарти, который, как многие считают, придумал термин «искусственный интеллект», заявлял, что не существует «четкого определения интеллекта, не зависящего от его связи с человеческим интел-

лектом», потому что «мы еще не можем охарактеризовать вообще, какие вычислительные процедуры мы хотим назвать интеллектуальными» [цит. по: 10, p. 360]. Таким образом, определения интеллекта сильно различаются и сосредоточены на мириадах взаимосвязанных человеческих характеристик, которые сами по себе трудно определить, включая сознание, самосознание, язык, способность учиться, способность абстрагироваться, способность адаптироваться и способность рассуждать [11].

Те же проблемы, которые мешают попыткам определить интеллект в целом, также применимы к попыткам определить искусственный интеллект. Ведущими теоретиками ИИ С. Расселом и П. Норвигом представлено восемь различных определений ИИ, разделенных на четыре категории: мыслить по-человечески, действовать по-человечески, мыслить рационально и действовать рационально [2, с. 34-39].

Рассел и Норвиг цитируют работы пионера вычислительной техники Алана Тьюринга, чьи труды предшествовали введению термина «искусственный интеллект», как пример «действующего по-человечески» подхода [2]. Другие ранние подходы к определению ИИ часто связывали понятие интеллекта со способностью выполнять определенные интеллектуальные задачи. Наиболее широко используемые современные подходы к определению ИИ фокусируются на концепции машин, которые работают для достижения цели – ключевой компонент «рационального действия», согласно С. Расселу и П. Норвигу.

С. Рассел и П. Норвиг используют концепцию «рационального агента» в качестве оперативного определения ИИ, определяя такого агента как «агента, который действует так, чтобы достичь наилучшего результата или, в случае неопределенности, наилучшего ожидаемого результата» [2, с. 78]. Однако с точки зрения регулирования целенаправленный подход не представляется достаточным, потому что он просто заменяет один трудно поддающийся определению термин (интеллект) другим (цель). Говоря обыденным языком, цель является синонимом намерения. Каким образом машина может иметь намерение – это скорее метафизический вопрос, чем этический или научный, и трудно определить цель каким-либо образом, который позволяет избежать требований, относящихся к намерению и самосознанию, не создавая чрезмерно исчерпывающего определения. Следовательно, не понятно, как определение ИИ через призму целей может обеспечить надежное рабочее определение искусственного интеллекта для регулирующих (нормативных) целей.

Использование более общей концепции «рационального действия» может быть как чрезмерно инклюзивным, так и недостаточным. Рациональное действие можно приписать огромному количеству компьютерных программ, которые не представляют опасности для общества. Определение «рационального действия» является также недостаточно исчерпывающим; точно так же, как программы ИИ, которые действительно действуют рационально, могут не создавать общественный риск, программы искусственного интеллекта, которые не действуют рационально, могут представлять серьезную опасность, общественные риски, если отсутствие рациональности затрудняет прогнозирование действия программы.

Это не означает, что системы ИИ, которые действуют рационально, не могут представлять общественный риск. Напротив, большая часть современной науки, рассматривающей катастрофические риски, связанные с ИИ, фокусируется на системах, которые стремятся максимизировать функцию полезности, даже если такая максимизация может представлять экзистенциальную угрозу для человечества. Но принцип рационального действия сам по себе не обеспечивает достаточно полного и точного определения ИИ.

Мы предлагаем «определять» ИИ для целей нашего исследования следующим образом: «Искусственный интеллект» – это машинная технология, способная выполнять задачи, которые требовали бы интеллекта, если бы их выполнял человек. Чтобы различать ИИ как концепцию и ИИ как технологию, требующую регулирования, мы будем использовать термин «система ИИ» применительно к последнему.

Сложности регулирования ИИ проистекают из определенных характеристик искусственного интеллекта. Эти характеристики отличают ИИ от предшествующих человеческих изобретений и ставят под сомнение достаточность любого механизма регулирования ИИ на основе процессов, которые вмешиваются только постфактум. Наиболее очевидная особенность ИИ, которая отличает его от более ранних технологий, – это способность искусственного интеллекта действовать автономно. Системы искусственного интеллекта уже могут выполнять сложные задачи, такие как вождение автомобиля и составление инвестиционного портфеля, без активного контроля со стороны человека. Сложность и объем задач, которые сосредоточатся в руках ИИ, несомненно, будут продолжать расти в ближайшие годы.

Существует фундаментальное различие между процессами принятия решений людьми и процессами современного ИИ – различия, которые могут привести системы искусственного интеллекта к созданию решений, недоступных человеческим ожиданиям. Люди, ограниченные когнитивными способностями человеческого мозга, неспособны анализировать всю или даже большую часть информации в условиях ограниченного времени. Поэтому они часто соглашаются на удовлетворительное, а не на оптимальное решение. Вычислительная мощность современных компьютеров (которая будет только расти) означает, что программа ИИ может

выполнять поиск многих других возможностей, чем это сможет осуществить человек за данный промежуток времени.

Именно эта способность генерировать уникальные решения делает использование ИИ привлекательным в постоянно увеличивающемся разнообразии областей, и у разработчиков, таким образом, есть экономический стимул для создания ИИ систем, способных генерировать такие неожиданные решения. Эти системы ИИ могут действовать непредвиденно в некотором смысле, но способность совершать непредвиденные действия, возможно, будет умышленно заложенной проектировщиками и операторами систем.

Разработка более универсальных систем искусственного интеллекта в сочетании с достижениями в машинном обучении почти наверняка говорит о том, что проблемы, связанные с непредвиденным поведением ИИ, будут возникать все чаще и чаще, что неожиданность поведения искусственного интеллекта значительно возрастет.

Обозначенные характеристики делают ИИ потенциальным источником общественного риска в масштабе, намного превышающем более известные формы общественного риска, которые являются исключительно результатом человеческого поведения.

Потерю контроля можно разделить на две разновидности. Потеря местного контроля происходит, когда система ИИ больше не может управляться человеком или людьми, несущими ответственность за его работу. Потеря общего контроля происходит, когда система ИИ не может больше контролироваться человеком вообще. Очевидно, последняя перспектива представляет гораздо больший общественный риск, чем первый, но даже потеря общего контроля не обязательно будет представлять значительный общественный риск, пока цели системы ИИ совпадают с целями общественности. К сожалению, обеспечить такое согласование интересов и целей может быть довольно сложно [12, p. 41-43].

Возможность несовпадения интересов проистекает из того факта, что цели ИИ определяются его первоначальным программированием. Даже если это первоначальное программирование позволяет или поощряет ИИ изменять цели, основанные на последующем опыте, эти изменения будут происходить в соответствии с требованиями начального программирования. Но многие эксперты в области ИИ предполагают, что если ИИ запрограммирован на достижение определенной цели, он может продолжать работать над достижением этой цели, даже если результаты его усилий не соответствуют субъективным ожиданиям первоначальных разработчиков ИИ.

Ученые, программисты и футуристы предупреждают, что более сильные формы ИИ могут противостоять человеческим усилиям управлять своими действиями и представлять собой угрозу – возможно, даже экзистенциальную – для человечества. Выражение этой озабоченности фокусируется на возможности того, что сложная система ИИ может улучшить собственное программирование до такой степени, что приобретет когнитивные способности, намного превосходящие таковые у его создателей-людей. Даже без принятия таких сценариев экзистенциального риска следует признать, что возникнут проблемы контроля и регулирования по мере того, как системы ИИ будут становиться все более мощными, сложными и автономными.

Необходимо признать, что рост ИИ до сих пор происходил в регулирующем вакууме. Существует очень мало законов или нормативных актов, конкретно касающихся уникальных проблем, поднятых ИИ, и не разработаны стандарты, конкретно регулирующие, кто на законных основаниях несет ответственность за нанесение вреда ИИ. Аналогична нехватка и потенциальных регуляторных подходов к процессам использования ИИ со стороны социума. Другими словами, отсутствует регулирование ИИ через призму институциональной компетентности для противостояния уникальным вызовам, связанным с развитием ИИ.

Исследований по вопросам управления и регулирования ИИ на сегодняшний день, особенно в России, не так много. Однако потенциальные возможности ИИ и проблемы, создаваемые ИИ, должны получить внимание исследователей. Вопросы управления и регулирования, связанные с ИИ, особенно со стороны социальных наук – это новая и сложная тема. Необходимо понимать, что качественное исследование данной проблемы должна обеспечить гибкость при сборе данных и управлении исследовательским процессом, который может быть длительным и неоднозначным. Успех исследования нам видится в разработке эффективной междисциплинарной методологии исследования и понимании сложных и всесторонних проблем ИИ в реальных условиях.

Понимание и решение вопросов управления и регулирования, связанных с ИИ, все еще находится на начальной стадии. Тем не менее ИИ быстро развивается, и вопросы управления и регулирования имеют решающее значение, и их необходимо обсудить уже сейчас. Данная статья призвана привлечь внимание к этим вопросам. Пытаясь сформулировать модели управления и регулирования для ИИ, мы получим представление о будущем развитии технологии ИИ, лучше поймем экономическое, социальное и политическое влияние ИИ и улучшим наше понимание и применение управленческих и регуляторных теорий в эпоху искусственного интеллекта. Ожидается, что дальнейшее изучение данной проблемы будет способствовать как академическому прогрессу в этой области, так и созданию и внедрению принципов управления и правил, связанных с ИИ.

Литература:

1. Siau K., Wang W., Governance, Policies, and Regulations // Proceedings of the Thirteenth Midwest Association for Information Systems Conference. – Saint Louis, Missouri, 2018. – May 17-18. – P. 1-5.
2. Рассел С., Норвиг П. Искусственный интеллект: современный подход, 2-е изд.: Пер. с англ. – М.: Издат. дом «Вильямс», 2007. – 1408 с.
3. Резаев А. В., Трегубова Н. Д. «Искусственный интеллект», «онлайн-культура», «искусственная социальность»: определение понятий // Мониторинг общественного мнения: экономические и социальные перемены. – 2019. – № 6. – С. 35-47.
4. Резаев А.В., Трегубова Н.Д. Искусственный интеллект и искусственная социальность: новые явления, проблемы и задачи для социальных наук // Мониторинг общественного мнения: экономические и социальные перемены. – 2021. – № 1. – С. 4-19.
5. Wolfe A. The Human Difference: Animals, Computers, and the Necessity of Social Science. – Berkley: University of California Press, 1993. – 235 p.
6. Ziewitz M. Governing Algorithms: Myth, Mess, and Methods // Science, Technology, & Human Values. – 2016. – Vol. 41. – № 1. – P. 3-16.
7. Коллинз Р. Может ли социология создать искусственный разум? // Личностноориентированная социология / П. Бергер, Б. Бергер, Р. Коллинз. – М.: Академический проект, 2004. – С. 566-598.
8. Hubbard F. P. “Sophisticated Robots”: Balancing Liability, Regulation, and Innovation // Florida Law Review. – 2014. – Vol. 66. – P. 1803.
9. Huber P. Safety and the Second Best: The Hazards of Public Risk Management in the Courts // Columbia Law Review. – 1985. – P. 277-320.
10. Scherer M. U. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies // Harvard Journal of Law & Technology. – 2016. – Vol. 29. – № 2 (Spring). – P. 354-400.
11. Premack D. Intelligence in Ape and Man. – New York: L. Erlbaum Associates, 1976. – 370 c.
12. Бостром И. Искусственный интеллект. Этапы. Угрозы. Стратегии / Пер. с англ. С. Филина. – М.: Манн, Иванов и Фербер, 2016. – 404 с.

Features of AI Regulation Mechanisms in the Context of a New Sociality

Obidin D.S.
Lobachevsky National Research State University of Nizhni Novgorod

The purpose of the article is to substantiate the need to create mechanisms for regulating AI on the part of the social sciences, and in particular, sociology. For this purpose, the analysis of existing definitions of AI is carried out and it is proposed to distinguish AI as a concept and AI as a technology that requires regulation. The characteristics of AI are described that represent AI as a potential public risk and that need to be monitored and regulated. In conclusion, the need to develop an effective interdisciplinary research methodology and understanding of the complex and comprehensive problems of AI in the real world is emphasized.

Key words: artificial intelligence, AI characteristics, machine technology, regulation mechanism, technological industry, social risks, management problems