



Диалог с техносубъектом



DOI: 10.30727/0235-1188-2024-67-3-7-26

Оригинальная исследовательская статья

Original research article

**Может ли искусственный интеллект  
подчинить человека?**

*В.М. Розин*

*Институт философии РАН, Москва, Россия*

**Аннотация**

В статье проведен критический анализ возможностей и ограничений технологий искусственного интеллекта (ИИ) в контексте дискуссий о потенциальных угрозах, связанных с их развитием. Рассматриваются и подвергаются сомнению ключевые аргументы, выдвигаемые в пользу тезиса о возможном подчинении человека системами ИИ. Автор проводит сравнительный анализ возможностей естественного интеллекта и ИИ на примере интерпретации психологического опыта К.Г. Юнга. Демонстрируется, что, несмотря на успехи современных нейросетей в обработке информации, они принципиально ограничены в способности к подлинному пониманию и творческому мышлению. Выделяются три ключевые новации, связанные с развитием ИИ: расширение когнитивных возможностей пользователей, формирование новой психической реальности «цифрового сознания» и возникновение гибридной формы жизни на стыке человеческой деятельности и технических процессов. Автор указывает на фундаментальные ограничения ИИ в сфере эмоционального интеллекта и творческих способностей. Внимание уделяется проблемам развития систем ИИ, включая влияние безличностных социальных структур на принятие решений, разобщенность разработчиков и пользователей, а также психологические эффекты взаимодействия с ИИ. Сделан вывод о необходимости переосмысления проблемы подчинения человека ИИ в более широком контексте влияния современных технологий на общество. Предлагается рассматривать наступающую эпоху как период сосуществования и взаимодействия двух типов интеллекта: естественного и искусственного. Высказывается опасение относительно того, что, не

пройдя серию начавшихся катастроф, человечество не опомнится, не изменит мировоззрения и поведения. В заключение автор призывает к активной работе над минимизацией рисков, связанных с развитием ИИ, при этом подчеркивая нереалистичность полного отказа от использования этих технологий.

**Ключевые слова:** философия искусственного интеллекта, большие языковые модели, нейросети, мировоззрение, социальные институты, творчество, мышление, естественный интеллект, сознание.

**Розин Вадим Маркович** – доктор философских наук, профессор, главный научный сотрудник сектора междисциплинарных проблем научно-технического развития Института философии РАН.

rozinvm@gmail.com

<http://orcid.org/0000-0002-4025-2734>

**Для цитирования:** *Розин В.М. Может ли искусственный интеллект подчинить человека? // Философские науки. 2024. Т. 67. № 3. С. 7–26. DOI: 10.30727/0235-1188-2024-67-3-7-26*

## **Can Artificial Intelligence Subjugate Humans?**

*V. M. Rozin*

*Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia*

### **Abstract**

The article critically examines the capabilities and limitations of artificial intelligence (AI) technologies within the context of ongoing debates surrounding potential threats stemming from their advancement. The study scrutinizes and challenges key arguments positing the possibility of human subjugation by AI systems. The author undertakes a comparative analysis of natural and artificial intelligence, employing the psychological experiences of C.G. Jung as a case study. It is demonstrated that despite the remarkable achievements of contemporary neural networks in information processing, they are fundamentally limited in their capacity for genuine comprehension and creative thought. The paper identifies three key innovations associated with AI development: the enhancement of users' cognitive capabilities, the formation of a novel psychic reality of "digital consciousness," and the emergence of hybrid life forms at the nexus of human activity and technological processes. The author highlights the fundamental limitations of AI in the realms of emotional intelligence and creative capabilities. Attention is drawn to the challenges associated with the development of AI systems, including the influence of impersonal social structures on decision-making, the disconnect between developers and users, and the psychological effects

of interacting with AI. The conclusion reached is that the issue of human subjugation by AI requires a re-evaluation within the broader context of the impact of contemporary technologies on society. It is proposed that the forthcoming era be viewed as a period of coexistence and interaction between two types of intelligence: natural and artificial. Apprehension is expressed that humanity will not adjust its worldview and behavior until after experiencing a series of impending catastrophes. In closing, the author advocates for proactive engagement in mitigating the risks associated with AI development, while simultaneously underscoring the impracticality of complete abstention from these technologies.

**Keywords:** philosophy of artificial intelligence, large language models, neural networks, worldview, social institutions, creativity, thinking, natural intelligence, mind.

**Vadim M. Rozin** – D.Sc. in Philosophy, Professor, Chief Research Fellow at the Department of Interdisciplinary Problems in the Advance of Science and Technology, Institute of Philosophy, Russian Academy of Sciences.

rozinvm@gmail.com

<http://orcid.org/0000-0002-4025-2734>

**For citation:** Rozin V.M. (2024) Can Artificial Intelligence Subjugate Humans? *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 3, pp. 7–26. DOI: 10.30727/0235-1188-2024-67-3-7-26

## Введение

Развитие технологий искусственного интеллекта (ИИ) порождает в научном сообществе и обществе в целом острые дискуссии о возможных угрозах, связанных с этим процессом. Одно из ключевых опасений связано с перспективой потери человеком собственной автономии и свободы вследствие возрастающего могущества ИИ. Данная проблема находит отражение в работах ряда современных исследователей и философов [Баррат 2015; Тегмарк 2019; Бостром 2016; О’Нил 2018; Narari 2024].

В обоснование реальности угрозы подчинения или даже порабощения человека со стороны систем ИИ приводят различные аргументы. Во-первых, современный уровень развития нейросетей интерпретируется как свидетельство того, что ИИ в будущем якобы станет равным или даже превзойдет естественный интеллект. В свое время Дж. Серль в знаменитом мысленном эксперименте «Китайская комната» показал, что внешне осмысленное поведение машины, дающей ответы на

вопросы на естественном языке, не означает наличия у нее подлинного понимания [Серль 1998]. Однако успехи современных нейросетевых моделей в имитации человеческого мышления и коммуникации у исследователей рождают ряд вопросов. Не выходят ли возможности ИИ за рамки простой имитации? Не приближаются ли машины к «подлинному» мышлению и пониманию? Например, известный популяризатор науки С. Карелов излагает точку зрения, в соответствии с которой интеллект по своей сути представляет собой вычисления, производимые сетью нейронов без заранее заданного алгоритма<sup>1</sup>. Согласно его позиции, существуют эксперименты, демонстрирующие возможность сведения понимания, вроде бы исключительно человеческого феномена, к вычислительным операциям. Еще одним направлением, в котором ИИ демонстрирует прогресс при воспроизведении человеческих способностей, считают т.н. эмоциональный интеллект. Более того, в последние несколько лет появляются разработки систем, умеющих распознавать и имитировать человеческие эмоциональные реакции. Системы типа Empathic Voice Interface и ChatGPT-4o демонстрируют способность, анализируя голос собеседника, идентифицировать его эмоции и использовать эту информацию в коммуникации. И, хотя аутентичность подобных эмоций вызывает сомнение, возможность их технической имитации многих впечатляет<sup>2</sup>.

Во-вторых, история развития техники говорит о том, что человечество склонно к созданию технологий, которые впоследствии не способно использовать разумно и взвешенно. Многие считают, что в случае ИИ эта проблема приобретает критический характер ввиду его потенциальной автономности и способности к видам интеллектуальной деятельности, которые ранее были привилегией человека. Порабощение в данном контексте может быть выражено в различных формах тотального контроля и манипуляции на базе ИИ [Зубофф 2022].

---

<sup>1</sup> Карелов С., Гурова Т. Хакнуть человечество. Интервью с Сергеем Кареловым // Монокль. 2019. 3 марта. – URL: <https://monocle.ru/expert/2019/10/haknut-chelovechestvo/?ysclid=m4fw0svkiy906705237>.

<sup>2</sup> См.: Карелов С. Началось обрушение фронта обороны от социохаккинга // Medium. 2024. 5 апреля. – URL: <https://sergey-57776.medium.com/d5a9ac89c441>.

В качестве примера приводят использование систем слежки и рейтингования граждан на основе анализа их поведения и физиологических реакций (известная практика в современном Китае). В дальнейшем совершенствование системы социального рейтинга с использованием новых технологий, как некоторые опасаются, откроет путь к т.н. цифровому тоталитаризму, в условиях которого ИИ станет инструментом предугадывания мыслей каждого индивида, подавления личной свободы<sup>3</sup>. Ряд исследователей полагают, что перспектива тотального порабощения человека посредством ИИ перестанет выглядеть футуристической антиутопией и приобретет вполне реальные очертания. Траектория развития ИИ определена не только научно-техническими факторами, но и социально-экономическими и политическими интересами. Ключевые решения о применении ИИ принимают не столько разработчики, сколько безличные институты: политические и экономические структуры. Их приоритетами служат не всеобщее благо и права человека, а извлечение прибыли, захват ресурсов, достижение стратегического преимущества, реализация властных и идеологических амбиций. Возникает закономерный вопрос о том, как будут разрешать этические дилеммы, связанные с применением потенциально опасных интеллектуальных систем. Развитие ИИ практически неизбежно порождает конфликт интересов между его создателями, владельцами и обществом в целом. Высказывается опасение и относительно того, что в условиях конфликта интересов решения будут принимать, исходя не из общественного блага и общечеловеческой морали, а частной выгоды тех, кто владеет и использует системы, построенные на технологиях ИИ.

Осознание этих угроз приводит некоторых исследователей к пессимистическим оценкам общего вектора развития цивилизации. Так, С. Карелов высказывает радикальное предположение о том, что, возможно, развитие человечества, начиная с Нового времени, – это «некий вывих эволюции», «тупиковая ветвь», которую «нужно прикрыть... и пойти более правильным путем»<sup>4</sup>.

<sup>3</sup> См., например, анализ этой проблемы: [Лекторский 2024, 19–20].

<sup>4</sup> Карелов С., Гурова Т. Хакнуть человечество. Интервью с Сергеем Кареловым // Монокль. 2019. 3 марта. – URL: <https://monocle.ru/expert/2019/10/haknut-chelovechestvo/?ysclid=m4fw0svkiy906705237>.

Насколько правомерен такой тревожный вывод? Действительно ли человечество обречено пасть жертвой порожденного им «цифрового Левиафана»? Чтобы ответить на эти вопросы, нам нужно понять возможности ИИ в сравнении с естественным интеллектом и определить направления влияния технологий ИИ на человека и общество.

### **Обладает ли искусственный интеллект способностью к пониманию?**

Приведенные выше рассуждения, обосновывающие обреченность человечества, могут показаться убедительными. Однако попытаемся в процессе анализа установить, не переоцениваем ли мы способности ИИ. Не сомневаюсь в том, что исследователь может организовать такую коммуникацию человека с нейронным компьютером, например, не требуя от компьютера реконструкции личности и ситуации (стандартный текст культуры); что компьютер сможет такой текст прочитать. Это и будет предъявлено как понимание. Но будет ли это пониманием для человека, создавшего данный текст? Проведем соответствующий эксперимент. В качестве примера рассмотрим фрагмент воспоминаний из книги К.Г. Юнга как стандартный текст, который можно предложить нейросетевому ИИ. Юнг описывает подростковые размышления, пришедшие ему в один из солнечных летних дней 1887 года: «Мир прекрасен и церковь прекрасна, и Бог, который создал все это, сидит далеко-далеко в голубом небе на золотом троне и... Здесь мысли мои оборвались, и я почувствовал удушье. Я оцепенел и помнил только одно: Сейчас не думать! Наступает что-то ужасное» [Юнг 1994, 46]. Через три дня Юнг все-таки решился вернуть явившуюся ранее картину: «Я собрал всю свою храбрость, как если бы вдруг решился немедленно прыгнуть в адское пламя, и дал мысли возможность появиться. Я увидел перед собой кафедральный собор, голубое небо. Бог сидит на своем золотом троне, высоко над миром – и из-под трона кусок кала падает на сверкающую новую крышу собора, пробивает ее, все рушится, стены собора разламываются на куски» [Юнг 1994, 49].

При объяснении данного текста следует учитывать ситуацию, в которой Юнг находился в тот период, и его состояние. Судя

по воспоминаниям, в эти годы Юнг оказался во власти проблем: он критикует, отрицает отца (пастора, потомственного священнослужителя) и церковь. Однако, поскольку верит в Бога, не может примириться с собственными переживаниями. Юнг пишет: «В этой религии я больше не находил Бога. Я знал, что больше никогда не смогу принимать участие в этой церемонии. Церковь – это такое место, куда я больше не пойду. Там все мертво, там нет жизни. Меня охватила жалость к отцу. Я осознал весь трагизм его профессии и жизни. Он боролся со смертью, существование которой не мог признать. Между ним и мной открылась пропасть, она была безгранична, и я не видел возможность когда-либо преодолеть ее» [Юнг 1994, 64].

Подобную ситуацию можно разрешить по-разному. Решение Юнга видится парадоксальным: он отрицает и церковь, и отца, но не прямо, а посредством нового Бога-революционера, которого Юнг сочиняет применительно к случаю, «здесь и сейчас». Это становится очевидным из текста. Восстановив возникшую картину, Юнг вспоминает: «Вот оно что! Я почувствовал неслышимое облегчение. Вместо ожидаемого проклятия благодать снизошла на меня, а с нею невыразимое блаженство, которого я никогда не знал... Я понял многое, чего не понимал раньше, я понял то, чего так и не понял мой отец, – волю Бога... Отец принял библейские заповеди как путеводитель, он верил в Бога, как предписывала Библия и как его учил его отец. Но он не знал живого Бога, который стоит, свободный и всемогущий, стоит над Библией и над Церковью, который призывает людей стать столь же свободным. Бог, ради исполнения Своей Воли, может заставить отца оставить все его взгляды и убеждения. Испытывая человеческую храбрость, Бог заставляет отказываться от традиций, сколь бы священными они ни были» [Юнг 1994, 45, 50].

Обратим внимание на предложение: «Я понял многое, чего не понимал раньше, я понял то, чего так и не понял мой отец, – волю Бога». Это понимание не могло бы состояться без двух вещей: реализации личности Юнга и его творчества (сочинить такое – задача непростая). В отличие от Юнга, мною предложено понимание того, что произошло с будущим психологом,

через реконструкцию рассмотренной ситуации в логике гуманитарной науки [Розин 2018].

Сакраментальным представляется вопрос о том, могла ли нейронная сеть, во-первых, занять такую позицию, т.е. антропологическую, исследовательскую, герменевтическую; во-вторых, осуществить сходную гуманитарную реконструкцию. Думаю, что не могла в принципе: компьютер не человек и не исследователь, не гуманитарий и методолог, каким является автор настоящей статьи. Проследим нашу позицию на примерах интерпретаций, предложенных реальными нейросетями.

*Ответ Claude 3 Opus:* Этот образ можно интерпретировать как мощный внутренний конфликт Юнга, столкновение его религиозного воспитания и бессознательных сомнений, еретических мыслей. Кал, разрушающий собор, это – очень яркая метафора падения веры, крушения привычной картины мира под давлением вытесненных кощунственных идей и импульсов.

То, что Юнг первоначально боялся даже допустить эту мысль, но затем решил посмотреть своему видению в лицо – важный момент. Это отражает его будущий подход как психолога: стремление не избегать тревожного и иррационального содержания психики, а исследовать и интегрировать его.

В целом этот отрывок дает интересное окно в мир подростковых переживаний Юнга и намекает на развитие его будущих психологических идей о необходимости принятия своей «тени» – бессознательной, пугающей, социально неодобряемой части личности. Преодоление своего страха и встреча со своим видением, каким бы шокирующим оно ни было, это – важный шаг личностного роста.

*Ответ Llama 3 70b:* Этот текст является отрывком из автобиографии Карла Юнга, где он описывает яркое детское воспоминание, которое произвело на него глубокое впечатление...

Этот отрывок важен, потому что он раскрывает ранние формы борьбы Юнга с понятием Бога и природой добра и зла. Видение представляет собой конфликт между его идеализированным взглядом на мир и суровой реальностью существования. Образ Бога на

золотом троне, окруженного красотой, вдруг нарушается падением кала, символизируя вторжение тьмы и хаоса в идеальный мир.

Этот опыт произвел глубокое впечатление на позднейшую работу Юнга как психоаналитика, повлияв на его теории о коллективном бессознательном, процессе индивидуации и интеграции противоположностей.

Попытаемся прокомментировать ответы нейросети. Последний из ответов – эпизод, произошедший с Юнгом, – близок к интерпретации, приведенной мною ранее в одной из работ [Розин 2011, 66–67]. Но я писал не о борьбе добра и зла, а о том, каким образом Юнг разрешает свою проблемную ситуацию и создает для этой цели Бога-революционера. Ответ ИИ основан на самой распространенной в интернете интерпретации ситуации, подобной той, которую описал Юнг в собственных воспоминаниях. И в остальных ответах мы видим распространенные в интернете версии данной ситуации. Можно ли такие ответы считать мышлением и творчеством? Скорее, нельзя. Но, если все-таки признать творчеством, то только на уровне эффективного поиска в интернете нужных ответов и адаптации их для предлагаемого случая (и это тоже непростая задача). Не отрицаю тот факт, что подобное «творчество» является необходимым для решения многих современных задач. Неслучайно возникло мнение о том, что если компьютер победил в шахматной игре человека, то ему никогда не победить человека, вооруженного компьютером. Тем не менее существует множество проблем и задач, которые может разрешить только естественный интеллект, в т.ч. использующий ИИ. Например, как было показано, ИИ не способен самостоятельно осуществить полноценную новаторскую герменевтическую реконструкцию философских текстов.

Поэтому, как предполагает ряд исследователей, в настоящее время одним из перспективных направлений развития ИИ служит разработка мультиагентных систем, в рамках которых создают сообщества ИИ-агентов, способных взаимодействовать друг с другом и с окружающей средой. Развивается также концепция «смертных» нейросетей, предполагающая установление

временной ограниченности существования моделей, чтобы стимулировать развитие более сложных форм поведения и рассуждения у ИИ [Славин 2024, 99]. Таким образом, в какой-то мере разработки, осуществляемые в контексте создания искусственной социальности, соответствуют изложенной нами ранее концепции о социальной природе мышления, о том, что процесс мышления разворачивается «под влиянием коммуникации и общения, культуры и социальности, личности и других проблемных ситуаций и социальных норм, форм осознания (концептуализации)» [Розин 2023, 17]. Обратим внимание на два момента: настоящую революцию, происходящую в сфере ИИ, и на то, что эта революция осуществляется на основе нового функционирования естественного интеллекта, который за счет рефлексии мышления, а также новых концепций и моделей становится двигателем эффективных разработок в сфере ИИ.

### **Агентность искусственного интеллекта**

Выше в статье речь шла о способностях, отсутствующих у нейросетей. Но каковы их возможности? Этот вопрос сегодня в гораздо большей степени интересует пользователей. Прослеживаются по меньшей мере три новации. Во-первых, одна из них обусловлена тем, что системы ИИ, основанные на нейросетях, – это новое «социальное тело» человека, т.е. семиотика и техника, расширяющая его биологические и антропологические возможности и способности. Поскольку на несколько порядков увеличиваются быстрдействие, память и информация, что дополняется извлекаемым из интернета опытом мышления, постольку средний пользователь с точки зрения социальных критериев выходит на уровень развитой личности, который в силу своего бэкграунда он не мог бы достичь.

Во-вторых, с развитием систем ИИ начинается процесс формирования новой психической реальности – «цифрового сознания». Общение с чат-ботами, основанными на больших языковых моделях, напоминает коммуникацию (поскольку происходит процесс обмена вопросами и ответами). Однако для коммуникации необходим субъект, потому что без него человек не представляет себе коммуникацию. Например, если

нам снятся действия или поступки, вплоть до необычных, например полетов во сне, наша психика, как я показываю в своей концепции сновидения, конструирует «сновидческого субъекта», которого мы чаще всего ощущаем как свое «Я» [Розин 2011, 240–241]. Это начинает происходить и с ИИ: мы невольно конструируем субъекта, представляющего послед-ного. Именно «цифровой субъект» нам отвечает и, как многие считают, мыслит не хуже нас.

Третья новация обусловлена объективным процессом активизации в пространстве-среде систем ИИ и интернета процессов и событий, инициированных нашими заданиями и ответами нейросетей. Во время одного из обсуждений в Институте философии РАН В.А. Лекторский заявил, что ИИ, конечно, является разумом, но чуждым человеку. Я бы уточнил: не разумом, а новой формой жизни, которую не задумывали, но она возникла самостоятельно. На эту мысль меня впервые навел интернет. На него, действительно, можно посмотреть как на особую форму жизни: пакетная и облачная передача информации, программы-паразиты и программы-защитники, разработчики и пользователи, изменяющие программное и аппаратное обеспечение, поскольку находятся в сложных взаимоотношениях (конкуренции, сотрудничества, заботы о людях, борьбы с ними и др.). Интернет развивается, и не только потому, что это развитие программируется. Нельзя ли аналогично взглянуть и на ИИ? С одной стороны, это электрические процессы, создание и передача информации. С другой – процессы и события новой, гибридной формы жизни, которую составляют, хотя и программируемые, но неизвестные процессы внутри устройства ИИ, связь с интернетом, концепции вычислений и решений задач, созданных под влиянием естественного интеллекта, действия разработчиков и пользователей, которые учитывают разработчики. Такая жизнь, конечно, необычна: вспыхивает и исчезает (возможно, сон нейросети). Но очевидны и основные характеристики жизни: рождение (создание нейросети, ее настройка и обучение), питание (электричеством и информацией), размножение, воспроизведение и развитие (над этим неутомимо работают ученые, программисты, техники), конкуренция с

другими видами компьютеров и форм жизни (уже происходит). Пока еще новая форма жизни (цифровая, техническая и антропологическая) не обладает субъектностью (агентностью). Тем не менее процесс не останавливается. В перспективе вырисовывается и фигура подобного «цифрового субъекта», в данном случае не только психологического, но и реального, живущего в пространстве-среде больших языковых моделей и интернета.

В связи с этим стоит уточнить и определение понятия ИИ. Чаще всего в определениях обращают внимание на новые возможности компьютеров: ИИ понимается как способность вычислительных машин решать задачи, доступные человеческому разуму [Sheikh et al. 2023, 15]. Но существуют и другие определения, трактующих разработку систем ИИ как сферу деятельности по исследованию и созданию совокупности методов и программных решений, позволяющих машинам анализировать окружающую среду, обучаться и принимать решения для достижения заданных целей наиболее эффективным способом [Sheikh et al. 2023, 16].

Во втором из приведенных выше определений указан не автономный машинный продукт (характеристика), а сфера деятельности человека (область компьютерных наук). Если ИИ – особая машина, способная соревноваться с человеком, то это одно понимание проблемы, но если приписывать ИИ способность опираться на возможности естественного интеллекта и развиваться, тогда деятельность ИИ включает сферу деятельности человека. Действительно, программное и машинное обеспечение ИИ постоянно совершенствуется, к ИИ предъявляют все новые и новые требования, которые заставляют разработчиков изменять и совершенствовать данное обеспечение. Требования формулируются пользователями ИИ, которые находятся между собой в сложных отношениях (конкуренции, кооперации и др.).

В результате становится очевидным, что определение понятия ИИ должно включать в себя характеристики (образы), во-первых, изменения и развития, во-вторых, антропологических сущностей (естественного интеллекта, разработчиков и пользователей, определенных аспектов социальности). ИИ –

развивающееся, антропосоциальное образование. В этом случае снимаются проблемы относительно того, может ли ИИ мыслить как человек, но вместе с тем возникают другие. Рассмотрим далее проблематику эмоций.

### **Ограничения искусственного интеллекта**

Современные разработчики утверждают, что они смогли покорить эмоции. Например, А.Р. Ефимов пишет: «Робот “ЭЛЕНА” является виртуальным гуманоидом, который обладает виртуальным гуманоидным телом, наделенным *мимикой*, может использовать естественный язык, *полностью поддерживая русский язык* (виземы, фонемы), является автономным (не требующим действий оператора при выполнении базовых функций), *обладающим собственной персоной* (в зависимости от персоны чат-бота) и имеющим первичную возможность для обучения изменениям внешнего мира (в работе реализовано распознавание лиц собеседников робота). Впервые в научной литературе представлена архитектура подобного робота. Уникальность ее заключается в том, что при ее построении использовался не только ряд существенных характеристик *субъективной реальности человека (опирающихся на современные исследования феноменологии сознания)*, но и сложный комплекс внешних проявлений сознательной деятельности человека, исполняющего роль ведущего на телевидении: мимика, выражение глаз, жесты, интонации голоса и т.д. (курсив мой. – В. Р.)» [Ефимов 2021, 20].

Однако относится ли все это к эмоциям? Исследования показывают, что эмоции тесно связаны с двумя важными антропологическими планами, в частности соматикой и семиотикой человека, которые вряд ли присущи компьютеру. Сильные эмоциональные реакции человека (смех, страх, гнев или горе) возникают в особых ситуациях, когда поведение становится неоднозначным из-за столкновения разных или противоположных мотивов, реальностей и установок. В таких случаях человек не может объединить эти противоречивые импульсы в единое целостное поведение. Этот принцип столкновения разных реальностей хорошо иллюстрирует, например детское

стихотворение К.И. Чуковского, в котором намеренно создаются абсурдные ситуации:

Мыши кошку изловили,  
В мышеловку посадили.  
Рыбы по полю гуляют,  
Жабы по небу летают.  
А лисички взяли спички,  
К морю синему пошли,  
Море синее зажгли...

При столкновении подобных реальностей, если человек не способен интегрировать их в некую новую, третью реальность, его поведение становится дезорганизованным. В зависимости от характера ситуации и скорости развития событий это может проявляться в форме ступора, защитных реакций или колебаний между разными реальностями, сопровождающихся возникновением эмоций. Эмоциональные реакции как телесные гештальты характеризуются рядом существенных особенностей. Прежде всего, они обладают способностью временно приостанавливать текущую деятельность, изолировать от нее и перенаправлять психическую энергию. Эмоции также формируют самостоятельную реальность с собственной системой отношений к происходящему, которые в культурном контексте определяются через понятия страха, радости, гнева, печали. Немаловажно и то, что эмоция как телесный гештальт представляет собой особую форму выражения, требующую освоения определенной техники владения собой. Как указывалось мной, «эти качества телесного гештальта эмоций позволяют человеку *прервать или затормозить текущую жизнедеятельность* (определяемую противоположными и несовпадающими мотивами); *перейти в особый режим* (его можно назвать “эмоциональным состоянием”, именно здесь эмоции выступают как значащие выражения, символы; последние не только указывают на определенные эмоциональные состояния человека – радость, горе, печаль и т.д., но и отчасти образуют эти состояния, помогают им состояться, выявиться в коммуникации); *перестроиться, “найти” новую реальность* (сформировать новый мотив,

установку), отчасти интегрирующую, снимающую предыдущие. В этом отношении эмоции действительно выполняют интегрирующую роль; так, горе позволяет нам не только забыть, но и пережить неприятные события и нащупать новые стимулы к жизни» [Розин 2002, 72].

О каких эмоциях можно говорить в случае нейросети? Большая языковая модель ChatGPT-4 на запрос пользователя сочинить стихотворение в таком же стиле, что и приведенное выше стихотворение К.И. Чуковского, сгенерировала стихи:

Воробы взяли шапки,  
По лесу играют в лапти.  
Зайцы на лугу танцуют,  
Совы днем глаз не закрывают.  
Кроты в небо улетели,  
Облака копать захотели...  
Караси в сапогах бродят,  
Гуси в поле зерно сеют.  
Свиньи в озере купаются,  
Коровы на деревья забрались.  
Лошади книги читают,  
В шахматы меж собой играют...

Эти стихи далеки от поэзии и эмоций в человеческом понимании. Почему же эмоциональные структуры и процессы нельзя моделировать, алгоритмизировать, а затем изготовить и реализовать на аппаратном уровне? Потому, что очень сложно. Да и зачем повторять биологическую и социальную эволюцию? Это потребует множество времени и ресурсов и вряд ли приведет к положительному для человечества итогу. Перейдем к рассмотрению вопроса о том, кто в сфере ИИ принимает решения и почему не удастся гуманизировать многообещающую новую технику.

### **Риски развития систем искусственного интеллекта**

Исследователи и разработчики ИИ, если бы были независимыми от заказчиков и ориентировались только на общественное благо, то думали бы о безопасности пользователей и других

последствиях, в т.ч. негативных. Но не они принимают решения, а безличностные институты (государство, монополии, корпорации, финансовые структуры). Последние заботятся не о благе граждан, а о прибыли, ресурсах, выигрыше в конкуренции, реализации «больших идей» (национальных, религиозных, идеологических, военных). Это первый круг проблем.

Второй круг проблем связан с разобщенностью разработчиков ИИ и пользователей. Мы не знаем о том, какие алгоритмы в нейронные сети заложили разработчики, каким образом они настроили программное и аппаратное обеспечение, кто еще участвовал в программировании и разработках (секретные службы, хакеры-«белые шляпы» или, что еще хуже, хакеры-«черные шляпы»). В результате мы думаем, что ИИ решает наши задачи, но часто он параллельно решает задачи других субъектов или в интересах этих субъектов. Не стоит думать о том, что, дообучая нейронную сеть, как это сегодня предлагается, мы разрешаем указанный круг проблем.

Если все-таки пользователь хочет контролировать ИИ, не должен ли он в какой-то форме участвовать в создании последнего? Следует понимать, какими компетенциями необходимо обладать, чтобы такой контроль осуществлять. С одной стороны, пользователь должен быть знаком с общей логикой проектирования и создания ИИ, с другой – логикой разрешения проблем и задач, с которыми ему предстоит справляться. Из этого следуют новые задачи образования, возрастание значения методологии (изучение мышления, творчества, решения проблем и задач), а также философии техники (исследование проектирования, программирования, робототехники, технической среды, образованной интернетом и ИИ).

Третий круг проблем и задач – влияние ИИ на психику разработчиков, пользователей и рядовых граждан. Является ли нормальным, с одной стороны, желание передать ИИ основные нерешаемые человечеством проблемы и заменить ИИ человека, с другой – настоящий страх перед ИИ? ИИ будет трансформировать сознание и видение человека. Это нормально. Важно избежать фобий и шизофрении, вызванных отставанием в осознании феномена ИИ в период технологической революции.

### **Заключение**

На мой взгляд, ответ на вопрос о возможности ИИ мыслить и заменить в этом аспекте человека постепенно становится более очевидным. В мышлении, как полагают современные философы, можно выделить две стороны: «создание нового» согласно логике естественного интеллекта, неотделимое от творчества, и «трансляцию мышления», предполагающую его рефлексивность и нормирование (образцы мышления, логику, онтологию, методологию). Например, Ж. Делез к мышлению относил только первую сторону, а вторую считал «узнаванием»: «В мире есть нечто, заставляющее мыслить. Это нечто – объект встречи, а не узнавания <...> “Человек может мыслить тогда, поскольку имеет такую возможность, но возможное еще не гарантирует того, что мы будем на это способны”; мышление мыслит лишь насильно, вынуждено встречая то, что “заставляет задуматься”, того, что следует обдумать – а обдумать нужно и немислимое или не-мысль, т.е. тот постоянный факт, что “мы еще не мыслим”» [Делез 1998, 135, 181].

ИИ вполне конкурентен в аспекте второй стороны мышления (это становится очевидным в т.ч. на примерах ответов нейросети), но с точки зрения создания в мышлении нового ограничен возможностями, которые в него заложили программисты. Эти возможности иные, чем в реальном творчестве, обусловленном активностью и взаимоотношениями людей, исторической филиацией идей и социальности, вызовами времени, случайными факторами. Новое, создаваемое в ИИ, ориентировано на другие задачи (социального контроля, скорости и стандартизации решений, выигрыша в конкурентной борьбе) и другие факторы, обусловленные обезличенностью социальных образований (о проблеме субъектов принятия решений писалось выше).

Изложенное означает, что начинается эпоха, в которой будут сосуществовать и взаимодействовать два типа интеллекта (естественный и искусственный), что возникнут (и уже складываются) гибридные способы мышления и творчества, что потребуются непросто философская и научная рефлексия их взаимоотношений, а также как положительных, так и негативных последствий [Розин 2023].

Вопрос о подчинения человека системой ИИ, поставленный нами вначале, нужно переформулировать в проблему негативных влияний («рисков», как писал М. Хайдеггер) техники в целом. В этом контексте, например, интернет или роботы в малой степени отличаются от ИИ. Но вряд ли Карелов прав, утверждая, что решение состоит в том, чтобы не использовать ИИ, а соответственно, интернет или роботов. Нереалистичным представляется и предложение Хайдеггера о том, чтобы человек «опомнился, снова ощутил широту своего сущностного пространства» [Хайдеггер 1993, 254]. Думаю, не пройдя серию катастроф, которые уже начались, человек не опомнится, не изменит мировоззрения и поведения. Проблема, однако, заключается в том, что современная техника (атомные станции и оружие), к тому же управляемая системами ИИ, может выйти из-под контроля и уничтожить цивилизованную жизнь на земле. Будем надеяться на лучшее, а главное, продолжать работать в этом направлении.

#### ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Баррат 2015 – *Баррат Дж.* Последнее изобретение человечества: искусственный интеллект и конец эры Homo sapiens / пер. с англ. Н. Лисовой. – М.: АНФ, 2015.

Бостром 2016 – *Бостром Н.* Искусственный интеллект: этапы, угрозы, стратегии / пер. с англ. С. Филина. – М.: Манн, Иванов и Фербер, 2016.

Делез 1998 – *Делез Ж.* Различение и повторение / пер. с фр. Н.Б. Маньковской, Э.П. Юровской. – СПб.: Петрополис, 1998.

Зубофф 2022 – *Зубофф Ш.* Эпоха надзорного капитализма. Битва за человеческое будущее на новых рубежах власти / пер. с англ. А.Ф. Васильева. – М.: Изд-во Института Гайдара, 2022.

Ефимов 2021 – *Ефимов А.Р.* Философско-методологические основы постьюринговой интеллектуальной робототехники: автореф. дис. ... канд. филос. наук. – М., 2021.

Лекторский 2024 – *Лекторский В.А.* Индивидуализация, коллективность, диалог // *Философские науки.* 2024. Т. 67. № 2. С. 13–25.

О’Нил 2018 – *О’Нил К.* Убийственные большие данные / пер. с англ. В. Дегтяревой. – М.: АСТ, 2018.

Розин 2002 – *Розин В.М.* Эмоции в искусстве, искусство – психотехника эмоций // *Мир психологии.* 2002. № 4. С. 71–85.

Розин 2011 – *Розин В.М.* Введение в схемологию: схемы в философии, культуре, науке, проектировании. – М.: URSS, 2011.

Розин 2018 – *Розин В.М.* Особенности дискурса и образцы исследования в гуманитарной науке. – М.: URSS, 2018.

Розин 2023 – *Розин В.М.* Концепции и понятия искусственного и естественного интеллекта (методологический анализ) // *Философские науки*. 2023. Т. 66. № 4. С. 7–25.

Серль 1998 – *Серль Дж.Р.* Сознание, мозг и программы // *Аналитическая философия: становление и развитие: антология / общ. ред. и сост. А.Ф. Грязнов*. – М.: Дом интеллектуальной книги; Прогресс-традиция, 1998.

Славин 2024 – *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.

Тегмарк 2019 – *Тегмарк М.* Жизнь 3.0. Быть человеком в эпоху искусственного интеллекта / пер. с англ. Д. Баюка. – М.: АСТ, 2019.

Хайдеггер 1993 – *Хайдеггер М.* Вопрос о технике // *Время и бытие: статьи и выступления*. – М.: Республика, 1993. С. 221–238.

Юнг 1994 – *Юнг К.Г.* Воспоминания, сновидения, размышления / пер. с нем. И. Булкиной. – Киев: AirLand, 1994.

Harari 2024 – *Harari Y.N.* Nexus: A Brief History of Information Networks from the Stone Age to AI. – New York: Random House, 2024.

Sheikh et al. 2023 – *Sheikh H., Prins C., Schrijvers E.* Mission AI: The New System Technology. – Cham: Springer, 2023.

#### REFERENCES

Barrat J. (2015) *Our Final Invention: Artificial Intelligence and the End of the Human Era* (N. Lisova, Trans.). Moscow: ANF (Russian translation).

Bostrom N. (2016) *Superintelligence: Paths, Dangers, Strategies* (S. Filin, Trans.). Moscow: Mann, Ivanov i Ferber (Russian translation).

Deleuze G. (1998) *Difference and Repetition* (N.B. Mankovskaya & E.P. Yurovskaya, Trans.). Saint Petersburg: Petropolis (Russian translation).

Efimov A.R. (2021) *Philosophical and Methodological Foundations of Post-Turing Intelligent Robotics* (dissertation). Moscow (in Russian).

Harari Y.N. (2024) *Nexus: A Brief History of Information Networks from the Stone Age to AI*. New York: Random House.

Heidegger M. (1993) The Question Concerning Technology. In: Heidegger M. *Time and Being: Articles and Speeches* (pp. 221–238). Moscow: Respublika (Russian translation).

Jung C.G. (1994) *Memories, Dreams, Reflections* (I. Bulkina, Trans.). Kiev: AirLand (Russian translation).

Lektorsky V.A. (2024) Individualization, Collectivity, Dialogue. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 2, pp. 13–25 (in Russian).

O'Neil C. (2018) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (V. Degtyareva, Trans.). Moscow: AST (Russian translation).

Rozin V.M. (2002) Emotions in Art, Art as a Psychotechnics of Emotions. *Mir psikhologii*. No. 4, pp. 71–85 (in Russian).

Rozin V.M. (2011) *Introduction to Schemology: Schemes in Philosophy, Culture, Science, Design*. Moscow: URSS (in Russian).

Rozin V.M. (2018) *Features of Discourse and Research Models in the Humanities*. Moscow: URSS (in Russian).

Rozin V.M. (2023) Concepts and Definitions of Artificial and Natural Intelligence: A Methodological Analysis. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 66, no. 4, pp. 7–25 (in Russian).

Searle J.R. (1998) Minds, Brains, and Programs. In: Gryaznov A.F. (Comp.) *Analytic Philosophy: Formation and Development: An Anthology*. Moscow: Dom intellektualnoy knigi; Progress-Traditsiya (Russian translation).

Sheikh H., Prins C., & Schrijvers E. (2023) *Mission AI: The New System Technology*. Cham: Springer.

Slavin B.B. (2024) Artificial Intelligence as a Factor of Transformation of State and Society: The Search for a Balance Between Managerial Efficiency and Human-Centeredness. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 3, pp. 99–122 (in Russian).

Tegmark M. (2019) *Life 3.0: Being Human in the Age of Artificial Intelligence* (D. Bayuk, Trans.). Moscow: AST (Russian translation).

Zuboff S. (2022) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (A.F. Vasilyev, Trans.). Moscow: Gaidar Institute Press (Russian translation).