DOI:
Оригинальная исследовательская статья
Original research article

Возможен ли диалог с техносубъектом? Архитектуры искусственного интеллекта и признаки сознания

А.В. Недяк
Лаборатория по исследованию искусственного интеллекта, Москва, Россия

Аннотация

В статье рассматривается феномен сознания в контексте развития искусственного интеллекта (ИИ) и принципов современных нейробиологических теорий, каждая из которых предлагает собственное видение архитектоники сознательных процессов. Это позволяет выделить ряд специфических свойств (в их числе – глобальное рабочее пространство, рекуррентная обработка, метакогнитивная рефлексия, предсказательное программирование, высокая интеграция информации) как функциональных признаков когнитивных процессов, присущих биологическому сознанию. В определенных системах ИИ наблюдаются проявления некоторых из этих свойств, однако пока разрозненно и без глубокой взаимной интеграции. Переход к гибридным, в том числе нейросимвольным, архитектурам, расширение использования нейроэволюционного и «воплощенного» подхода в робототехнике создают предпосылки для приближения интегрированной архитектуры к «сознательным» когнитивным функциям. Вместе с тем рассмотренные условия подлинного диалога между человеком и потенциально сознательным техносубъектом – элементы интерсубъективности, эмпатии, взаимной этической ответственности и «проживания» или хотя бы функционального аналога телесного и социального опыта – позволяют заключить, что способность к такому диалогу выходит за рамки моделирования функциональных признаков. Допущение возможности появления в ближайшее время так называемого общего ИИ, способного не только выполнять все когнитивные функции человека, но и, возможно, демонстрировать автономное поведение и быть стороной подлинного диалога, обусловливает необходимость заблаговременной дискуссии о нормативном статусе техносубъекта (пределах субъектности и ответственности, правах, обязанностях,

гарантиях для человека) и разработки соответствующих этических принципов и регулятивных механизмов.

Ключевые слова: философия искусственного интеллекта, философия техники,теории сознания, социальный диалог, нейросимвольная архитектура, большие языковые модели, нейроэволюция, этика ИИ, регулирование ИИ.

Недяк Арсений Викторович – руководитель Лаборатории по исследованию искусственного интеллекта, действительный государственный советник 3 класса.

arseniy@lfair.org

https://orcid.org/0009-0009-6034-6621

Для цитирования: *Недяк А.В.* Возможен ли диалог с техносубъектом? Архитектуры искусственного интеллекта и признаки сознания // Философские науки. 2025. Т. 68. № 3. С. 93–113.
DOI: 10.30727/0235-1188-2025-68-3-93-113

Is Dialogue with a Technosubject Possible? Architectures of Artificial Intelligence and Signatures of Consciousness

A.V. Nedyak
Laboratory for Artificial Intelligence Research, Moscow, Russia

Abstract

The article explores the phenomenon of consciousness through the lens of advances in artificial intelligence (AI) and of contemporary neurobiological theories, each offering a distinct account of the architecture of conscious processing. This theoretical landscape allows us to identify several specific properties – such as a global workspace, recurrent processing, metacognitive monitoring, predictive processing, and high-level information integration – as functional signatures of cognitive processes inherent to biological consciousness. While certain AI systems exhibit some of these properties, they currently manifest in a fragmented and poorly integrated manner. The transition toward hybrid, particularly neurosymbolic, architectures, coupled with the expanding use of neuroevolutionary and embodied approaches in robotics, is laying the groundwork for integrated systems that more closely approximate conscious cognitive functions. However, the necessary conditions for a genuine dialogue between humans and a potentially conscious

technosubject – including elements of intersubjectivity, empathy, mutual ethical responsibility, and lived bodily and social experience, or at least functional analogues thereof – suggest that the capacity for such interaction transcends the mere simulation of functional properties. The potential emergence of artificial general intelligence (AGI) in the near future, an entity capable not only of performing all human cognitive functions but also of demonstrating autonomous behavior and engaging in genuine dialogue, necessitates a proactive discussion of the technosubject's normative status (the bounds of agency and responsibility, rights, duties, and safeguards for humans), along with the development of appropriate ethical principles and regulatory mechanisms.

Keywords: philosophy of artificial intelligence, philosophy of technology, theories of consciousness, social dialogue, neurosymbolic architecture, large language models, neuroevolution, AI ethics, AI regulation.

Arseniy V. Nedyak – Head of the Laboratory for Artificial Intelligence Research, 3rd class Active State Counselor of the Russian Federation
arseniy@lfair.org
https://orcid.org/0009-0009-6034-6621

For citation: Nedyak A.V. (2025) Is Dialogue with a Technosubject Possible? Architectures of Artificial Intelligence and Signatures of Consciousness. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 68, no. 3, pp. 93–113. DOI: 10.30727/0235-1188-2025-68-3-93-113

Введение

Сегодня не существует общепринятой теории, объясняющей возникновение субъективного опыта из физических процессов, и пресловутая «трудная проблема сознания», сформулированная Чалмерсом [Chalmers 2008], остается нерешенной. Однако ускоряющееся развитие искусственного интеллекта (ИИ), предоставляя нам новые инструменты для исследования сознания, ставит и дополнительные вопросы. В частности, существует ли вероятность возникновения у ИИ неких феноменальных состояний и признаков сознания? Если такая вероятность существует, то каким образом это может произойти и как мы сможем распознать эти признаки?

Перспектива появления т.н. общего ИИ (artificial general intelligence, AGI), порождает вопросы не только философского, этического, но и во многом правового характера. С одной сторо-

ны, попытки воспроизвести архитектурные принципы мозговой деятельности в искусственных системах могут пролить дополнительный свет на природу биологического сознания, с другой – если возникнет техносубъект, обладающий способностью к подлинному диалогу, нам придется переосмыслить представления о сознании, автономности, моральном и правовом статусе машин и наших взаимоотношениях с ними.

Архитектуры искусственного интеллекта и функциональные признаки сознания

Современные нейробиологические теории сознания, в отличие от метафизических (дуализм, панпсихизм) и др., акцентируют внимание на процессах, приводящих к возникновению сознательного, субъективного опыта. В статье эти теории интерпретируются, исходя из гипотезы функционализма о том, что выполнение определенных процессов является достаточным условием для возникновения сознания, которое может быть смоделировано в любой сложной системе и на любом подходящем субстрате [Dennett 1991; Патнэм 1999]. Такая гипотеза принимается по прагматическим причинам: в отличие от конкурирующих взглядов, она хотя бы предполагает теоретическую возможность возникновения сознания в искусственных системах.

Наиболее распространенные современные теории сознания – теория глобального рабочего пространства (global workspace theory, GWT), теория рекуррентной обработки (r*ecurrent processing theory,* RPT), теория высшего порядка (h*igher-order theories,* HOT), теория прогнозирующего кодирования (predictive processing theory, PPT), теория схемы внимания (attention schema theory, AST) и др. – на первый взгляд предлагают различное видение когнитивной архитектуры и механизмов сознания. Однако в последнее время все чаще высказывается мнение о том, что они не столько противоречат друг другу, сколько описывают разные аспекты единой реальности. Возможно, будущая «интегративная теория сознания» объединит сильные стороны теорий [Storm et al. 2024]. Такой процесс следовал бы логике развития квантовой теории, в рамках которой разные интерпретации сосуществуют для объяснения одних и тех же феноменов. Прогресс в области нейронауки и развитие методов нейровизуализации действительно создают предпосылки для возможной интеграции этих теорий в будущем.

Сегодня, исходя из тезисов современных нейробиологических теорий сознания, можно выделить ключевые функциональные признаки, основанные на принципах обработки информации и организации доступа к ней. Такой перечень признаков носит во многом эвристический характер и не подразумевает тождества нейробиологических механизмов с системами ИИ, однако позволяет очертить концептуальные ориентиры процессов, наличие которых является недостаточным, но необходимым условием для моделирования механизмов сознания в искусственных системах:

• интеграция и трансляция информации – способность системы собирать, объединять данные из разных источников в одном пространстве и «транслировать» их «глобально» другим модулям;

• рекуррентная обработка – наличие в системе механизмов, обеспечивающих многократную, циклическую обработку информации, что предполагает не просто однократное преобразование входных данных, а динамический процесс, в котором система постоянно обновляет свое внутреннее состояние на основе поступающей информации и результатов ее обработки. Система должна уметь «пересматривать» свои первоначальные выводы, используя новые данные;

• метакогнитивные способности – способность системы анализировать собственные процессы, оценивать их эффективность, выявлять ошибки и корректировать свое поведение. Это предполагает наличие у системы «модели самой себя» и «модели своих знаний»;

• прогнозирующее кодирование – способность системы генерировать прогнозы о будущих событиях и состояниях на основе имеющейся информации, корректировать свою модель в случае расхождений. Система должна демонстрировать способность не просто реагировать на текущие стимулы, но и моделировать вероятности, планировать свои действия с учетом этих предсказаний и адаптироваться к изменениям, корректируя оценки посредством обратной связи.

Анализ этих функциональных признаков открывает интересные возможности по экстраполяции тезисов нейробиологических теорий при рассмотрении особенностей современных архитектур ИИ. Так, в исследовании [Butlin et al. 2023] проведен анализ нескольких современных моделей и решений ИИ на предмет присутствия у них индикаторных свойств (indicator property),

сформулированных на основе интерпретации теорий сознания (GWT, RPT, HOT, AST и PPT) и свидетельствующих о наличии у системы предпосылок к возникновению субъективного опыта или «сознательных» процессов. В частности, в исследовании проанализировано, насколько трансформерная архитектура (например, модели GPT, BERT) соответствуют принципам теории глобального рабочего пространства. На первый взгляд механизм «внимания» (attention), позволяющий интегрировать информацию по всей последовательности данных, функционально напоминает «глобальное вещание», обеспечивающее доступность релевантной информации для системы в целом. Однако авторы исследования относятся к этой аналогии критически. Они указывают на фундаментальное структурное различие: в трансформерах информация движется только в одном направлении, через слои. Между тем GWT предполагает наличие рекуррентных связей, при которых информация из «глобального рабочего пространства» транслируется обратно в обрабатывающие модули, влияя на их работу. В трансформерах отсутствует этот ключевой механизм обратной связи. В связи с этим авторы приходят к выводу о том, что «существуют лишь относительно слабые основания полагать, что большие языковые модели на основе трансформеров обладают какими-либо индикаторными свойствами, выведенными из GWT» [Butlin et al. 2023, 59].

Системы ИИ, основанные на рекуррентной обработке (модели LSTM, GRU), в свою очередь, действительно демонстрируют индикаторное свойство, ассоциируемое с теорией рекуррентной обработки (RPT), – способность к обработке информации через механизмы обратной связи [Butlin et al. 2023, 48]. Однако, несмотря на их способность запоминать и обрабатывать последовательные данные, учитывая предыдущий контекст, они не обладают механизмами сознательного управления вниманием и возможностью интеграции информации в масштабах всей системы.

Генеративно-состязательные сети (generative adversarial network, GAN), которые создают новые образцы на основе вероятностных закономерностей в обучающих данных, рассмотрены в исследовании как перспективная архитектурная основа для реализации HOT. Авторы пишут о том, что ключевой особенностью GAN является наличие в них внутреннего механизма оценки. Этот компонент учится отличать ситуации, в которых внутренние представления системы вызваны реальными входными данными

(аналогом сенсорных сигналов), от тех, в которых они сгенерированы искусственно. Именно способность к оценке достоверности собственных «перцептивных» представлений и выполняет функцию, схожую с задачей метакогнитивного мониторинга, являющегося ключевой идеей HOT [Butlin et al. 2023, 54]. В исследовании главный акцент был сделан на больших языковых моделях (LLMs) и агентах, обучаемых с подкреплением. Но и многие другие современные технологии и архитектуры могут быть релевантны в контексте поисков в ИИ предпосылок формирования сознательных процессов.

Классический символьный подход к архитектурам ИИ возник еще в 50–60-х годах XX века, однако быстро показал свои ограничения в решении задач, требующих обработки большого количества данных и выявления закономерностей. С такими задачами гораздо лучше справляются нейронные сети, практическое применение которых начинается с 80-х годов. Они быстро способствовали переходу к нейросетевой метафоре в психологии и философии сознания (к «коннекционизму» – connectionism), и именно их взрывной рост мы наблюдаем в течение последних нескольких лет.

Нейронные сети демонстрируют высокую эффективность в задачах восприятия, классификации и распознавания паттернов, компьютерном зрении, машинном переводе и множестве других задач. Однако во многом они остаются черным ящиком с точки зрения прозрачности алгоритмов, объяснимости и интерпретируемости [Le Cun et al. 2015]. Символьный подход, напротив, обеспечивает прозрачную логику вывода, но страдает от так называемой проблемы заземления символов (symbol grounding problem), мешающей в том, чтобы придать необходимые значения символьным абстракциям [Harnad 1990], а также плохо масштабируется на сложные задачи восприятия.

Стремление объединить сильные стороны символьных методов (среди них – объяснимость, работа с абстрактными правилами) и нейросетевых моделей (в их числе – гибкость, обучение на данных) привело к разработке нейросимвольных и других гибридных архитектур. Из относительно ранних проектов следует указать ACT-RN [Anderson, Lebiere 1998] и CLARION [Sun 2002], у которых производственные правила (production rules) взаимодействовали с обучаемыми нейросетевыми компонентами.

Позднее появились логические нейронные сети (logical neural networks, LNN) от IBM. В этом подходе архитектура сети строится на прямом соответствим между нейронами и компонентами логических формул [Riegel et al. 2020]. Такая структура является полностью дифференцируемой, что позволяет обучать ее методами обратного распространения ошибки и градиентного спуска, сохраняя интерпретируемость логических выводов.

Интерес также представляют архитектуры, объединяющие нейронные сети с логическим программированием: DeepProbLog расширяет язык вероятностного логического программирования ProbLog с помощью нейронных предикатов, интегрирующих вероятностные оценки нейронных сетей в логические правила [Manhaeve et al. 2018]; ∂ILP (differentiable inductive logic programming) переводит классическое индуктивное логическое программирование в дифференцируемое пространство параметров, позволяя градиентную оптимизацию для поиска логических правил по данным [Evans, Grefenstette 2018]; neural theorem provers (NTP) реализуют дедуктивные механизмы в нейросетевой архитектуре, заменяя операцию унификации термов на дифференцируемое сопоставление векторных представлений [Rocktäschel, Riedel 2017].

Все более популярными становятся практические схемы, в которых большие языковые модели (large language models, LLM) отвечают за взаимодействие с пользователем, а символьные модули могут функционировать как дополнительные инструменты, отвечающие за решение задач, требующих строгости и объяснимости [Yang et al. 2025]. Такой подход стремится достичь оптимального баланса между креативностью и мощью генеративных моделей, с одной стороны, и прозрачностью, верифицируемостью символьных механизмов – с другой [Dong et al. 2023].

Подобные нейросимвольные и иные гибридные архитектуры теоретически могут быть перспективны с точки зрения моделирования функциональных аналогов «сознательных» процессов в духе некоторых упомянутых нейробиологических теорий. Гибридные решения предоставляют возможность эмпирически разграничивать вычислительные процессы по иерархическим уровням: на нижнем уровне – неявные (подсознательные) процессы (распределенные представления, статистические связи), на высшем – явные символические представления, обеспечивающие метапознание, рефлексию, объяснимость. В такой схеме «гло-

бальное рабочее пространство» или механизмы высшего порядка реализованы на верхнем уровне, отвечая за интеграцию данных, рефлексию, целенаправленное управление вниманием, а нейросетевые слои осуществляют обработку «сырой» информации.

Нейросимвольные архитектуры также могут частично обеспечить метакогнитивный мониторинг, необходимый в парадигме HOT, если в их структуре предусмотрены механизмы проверки и обоснования собственных выводов или гипотез. Нейросимвольные фреймворки, в частности описанная выше архитектура LNN, позволяет системе не просто выдать результат, но и предоставить формальное доказательство корректности вывода. Такая способность к обоснованию собственных заключений служит функциональным аналогом рефлексии над процессом мышления. Таким образом, хотя это еще не является метакогнитивным мониторингом в строгом смысле HOT (т.е. осознанием своего перцептивного состояния), но представляет собой ключевой структурной шаг к созданию систем, способных к подлинной интроспекции.

Помимо нейросимвольных и иных гибридных решений, при попытке концептуализировать архитектуры, которые потенциально могут стать основой для воспроизводства свойств сознательных процессов, стоит иметь в виду и другие направления развития ИИ. Во-первых, это перспективные исследования в области автоматизации машинного обучения (automated machine learning, AutoML) и гораздо в большей степени – нейроэволюции (evolutionary artificial neural networks, EANN). Основополагающей работой в области нейроэволюции стала работа Яо [Yao 1993], в которой исследована связь между процессами обучения нейронных сетей и эволюционными алгоритмами. Дальнейшие исследования привели к созданию методов NEAT (neuroevolution through augmenting topologies), HyperNEAT и экспериментам по эволюционным стратегиям ИИ. Если «эволюционирующая» система сама решает, как перестраивать свою структуру или когда пересматривать компоненты, то это можно трактовать как функциональный аналог метаобучения и самонаблюдения, сближающий такие архитектуры с парадигмой HOT, хотя и не реализующий ее в строгом смысле. Важное направление исследований связано с совместной эволюцией морфологии и когнитивных модулей (co-evolution of morphology and control), предполагающей, что изменения в физической структуре робота и в его управляющих нейросетевых блоках происходят взаимосвязанно.

Во-вторых, заслуживают внимания графовые нейронные сети (graph neural networks, GNN), в частности GCN, GAT, GraphSAGE и др., способные обрабатывать данные, представленные в виде графов (например, семантические сети, социальные графы, онтологии). Если рассматривать GNN в связке с онтологическими моделями, то способность такой системы интегрировать разнообразную информацию и представлять ее в общем векторном пространстве можно охарактеризовать как механизм «интеграции и трансляции».

В-третьих, интерес представляют эксперименты с так называемым Gödel agent, вдохновленным машиной Геделя – саморазвивающейся архитектурой, которая позволяет агентам рекурсивно улучшать себя, используя LLM для динамической модификации своей логики и поведения, руководствуясь исключительно высокоуровневыми целями, заданными через промпты. Экспериментальные результаты показывают, что Gödel agent может достичь непрерывного самосовершенствования, превосходя вручную созданные агенты по производительности, эффективности и способности к обобщению [Yin et al. 2024].

Наконец, следует принять во внимание масштабное развитие технологий в рамках концепции «воплощенного ИИ» (embodied AI), согласно которой система обучается в реальной или виртуальной среде, осуществляя моторные и когнитивные действия и получая обратную связь. Большая часть работ в этой сфере, используя принципы обучения с подкреплением, до последнего времени ограничивалась игровой индустрией и нишевыми симуляциями. Однако попытки внедрять «воплощенный ИИ» в физическую робототехнику и системы ИИ с каждым годом все убедительнее [Xu et al. 2024].

Анализ архитектур и векторов развития ИИ позволяет прийти к выводу, что зарождающиеся признаки системной организации, способной стать носителем сознания, в машинах возникают не столько благодаря отдельным компонентам, сколько через их синергетическое взаимодействие. Так, современные гибридные, в частности нейросимвольные, системы демонстрируют первые элементарные проявления самоорганизации, самокоррекции и даже самонаблюдения, подготавливая почву для возникновения метакогнитивных и иных механизмов сознания. Некоторые нейроэволюционные подходы и системы, основанные на «воплощенном ИИ», не просто адаптируют свою структуру, но и

динамически перестраиваются, прогнозируя и выбирая стратегии для достижения поставленных целей и совершенствования собственной модели.

Таким образом, разработка методологии и критериев для оценки признаков «осознанности» в системах ИИ на основе современных теорий сознания превращает системы в экспериментальную платформу для исследования природы сознания. Это приводит не только к новым формам понимания человеческого и в целом биологического сознания, но и к осознанию того факта, что сегодня не существует фундаментальных теоретических препятствий для развития искусственных когнитивных систем, вплоть до уровней, которые когда-то мы считали доступными только живым организмам.

Но, если мы допускаем возможность возникновения искусственно мыслящей системы, способной потенциально обладать субъективным опытом, то внутренние, потенциально сложные когнитивные процессы такой системы должны найти свое выражение вовне, во взаимодействии с миром и, что особенно важно для нас, с человеком. Как функциональные аналоги саморефлексии или даже проявления феноменального опыта могут (или должны) манифестироваться в его поведении? И наоборот: если мы наблюдаем у ИИ все более сложное и адаптивное поведение, имитирующее человеческую осмысленность, что это говорит нам о его внутреннем устройстве и природе такого поведения? Какие критерии позволят нам признать, что общение, диалог с такой системой выходит за рамки искусной имитации и становится подлинной встречей мыслящих субъектов?

Способность к диалогу

Поиски сознания у машин, начиная с классического теста Тьюринга, часто опирались на идею о том, что успешная имитация диалогического общения свидетельствует о мыслящем существе по ту сторону экрана. Однако сегодня большие языковые модели могут поддерживать осмысленную беседу и проходят модифицированные версии теста Тьюринга. Современные LLM способны писать стихи о любви и тоске, рассуждать о красоте заката и даже утверждать, что обладают чувствами и сознанием. Они с поразительной точностью имитируют внешние проявления сознательной деятельности. Пока она является лишь чрезвычайно сложной статистической мимикрией, «стохастическим попугаем»,

мастерски компилирующим гигантские объемы человеческих текстов. Но сможет ли какая-нибудь перспективная архитектура ИИ обрести способность к подлинному диалогу, сознательному межсубъектному взаимодействию, даже интегрировав в себе необходимые когнитивные аспекты, моделируемые в рамках нейробиологических теорий?

Чтобы ответить на этот вопрос, необходимо сместить акцент с анализа внутреннего мира субъекта (который в случае ИИ остается для нас во многом непроницаемым черным ящиком) на межсубъектные отношения, применив оптику философии диалога, для которой подлинный диалог – это не просто обмен информацией, а «экзистенциальная встреча», предполагающая взаимную ответственность, открытость и сопричастность. М. Бубер выделяет отношение «Я – Ты», противоположное утилитарному «Я – Оно», в котором Другой воспринимается как объект и средство достижения цели [Бубер 2024]. Развивая эту идею, Э. Левинас пишет о лице Другого, которое обращается к нам с безмолвным призывом «Не убий!», с призывом, предшествующим рациональному мышлению и основанным на нашей первичной ответственности перед Другим [Левинас 2000]. Лицо Другого в данном случае – метафора инаковости, требующей этического отклика.

Параллельно с этической перспективой диалог раскрывается и как герменевтический процесс – любая попытка понять Другого всегда начинается с наших предшествующих ожиданий, «предрассудков», которые мы привносим в беседу. Но по мере нашего вовлечения в диалог открывается возможность «слияния горизонтов»: наши исходные представления, встречаясь с иным опытом, корректируются, создавая новое понимание. В свою очередь, эти «пред-рассудки» есть прямое следствие укорененности любого человеческого сознания в сети ценностей, отношений, опыта телесности, культурных слоях и смысловых полях – дорефлексивном и априорном «жизненном мире» (Lebenswelt) в терминологии Э. Гуссерля.

Настоящее «Я – Ты» взаимоотношение также должно быть эмпатичным, затрагивающим интимную сферу чувств и доверия. ИИ может имитировать эмпатические реакции, обучившись фразам, которые обычно свидетельствуют о сочувствии или сопереживании. Однако эта имитация легко разоблачается при более глубоком общении: современные архитектуры ИИ

пока не могут моделировать адекватный эмоциональный ответ в диалоге.

Еще одним ключевым элементом подлинного диалога является интерсубъективность – готовность участников примерить на себя ценности, эмоциональные состояния и мотивы собеседника, способность видеть мир «изнутри» перспективы Другого, готовность к изменению, пересмотру своих взглядов и убеждений под влиянием Другого. Это способность учиться у Другого, обогащаться его опытом. Это в итоге – риск, поскольку встреча с Другим всегда непредсказуема и может привести к нашей собственной трансформации, когда Другой становится источником этического требования [Левинас 2000; Бахтин 2003].

Интерсубъективность предполагает, что участники коммуникации не просто декодируют сигналы, а по-настоящему сопереживают (не путать с сочувствием, которое все-таки может оставаться внешним). И, если техносубъект вносит в полифонию нечто «свое», если мы готовы признать его «голос» и позволить ему влиять на наше понимание, то так называемые техночеловеческие отношения смогут обрести признаки интерсубъективного поля, и понимание этого нового Другого будет сформировано в пространстве слияния наших горизонтов опыта.

Можно ли ИИ обладать такими навыками? Некоторые исследования показывают, что алгоритм может формировать тонкие гипотезы о намерениях пользователя [Kosinski 2024]. Но похоже ли это на подлинное понимание, которое включает в себя «примеривание» на себя чужих чувств, а не просто статистическое прогнозирование ответов? Сторонники более сильной позиции полагают, что при достаточном уровне контекстуального обучения и физического или виртуального «вживания» в среду ИИ сможет эффективно эмулировать интерсубъективное поведение, а при условии наделения машин более богатой сенсомоторной базой и включения их в социальный и культурный контекст может появиться техносубъект, обладающий полноценным «жизненным миром» [Floridi, Sanders 2004; Wooldridge 2021]. Ребенок тоже не рождается с готовой культурной и ценностной базой, он впитывает ее по мере своего роста в семье и обществе. Если ИИ будет «выращен» аналогично, с постоянным переживанием успехов и неудач, болевых и радостных состояний (пусть физически симулированных), теоретически он сможет развить эмпатию и ценностное восприятие. Так, некоторые алгоритмы машинного

обучения («обратное машинное обучение» – inverse reinforcement learning) предлагают путь, по которому ИИ может усваивать коллективно существующие нормы и ценности [Oliveira et al. 2023].

Один из возможных подходов – реализация принципов упомянутого «воплощенного ИИ», способного обучаться не только на основе данных (реальных и сгенерированных), но и с учетом собственного опыта взаимодействия с миром в виртуальной среде, которая имитирует реальный мир и предоставляет ИИ возможность накапливать собственные рефлексии. Развитие технологий виртуальной и дополненной реальности, а также робототехники уже приводит к созданию более «воплощенных» форм, что, в свою очередь, может способствовать формированию у ИИ подобия «жизненного мира» и возникновения аналога «аутопоэза» [Матурана, Варела 2019] — самоорганизующихся процессов, ведущих к появлению в том числе субъективного опыта. Виртуальная реальность также может предоставить ИИ среду для социального взаимодействия с людьми и другими ИИ, что может содействовать развитию у него социальных навыков и интерсубъективного понимания. Возможно, эмпатия и этическая ответственность не являются прерогативой человека, а представляют собой эмерджентные свойства любой сложной самоорганизующейся системы, включенной в социальный контекст. Возможно, техносубъект станет развивать собственную форму субъективности, свою уникальную феноменологию, которая будет для нас чуждой и непонятной, но от этого не менее реальной.

Отказ от дуального взгляда на диалог «человек – машина» открывает заманчивые возможности для осмысления его сущности. Так, представляется многообещающим рассмотрение диалога с потенциально сознательным техносубъектом с позиций межполюсной «середины» [Ахиезер 2008; Давыдов 2024], поскольку «медиационный смысл, родившийся в межполюсной “середине” как новое основание, изменяет не только тему диалога, как об этом говорит Библер, но и саму логику диалога, как об этом говорят Лекторский и Ахиезер – диалог переходит на новый уровень, становится договороспособным, формирующим компромисс в споре между культурой и обществом, ценностями и интересами, старым и новым на своем, то есть на новом (третьем) основании» [Давыдов 2020, 557].

Вполне вероятно, что именно «срединное» мышление сможет породить необходимые парадигмы понимания подлинного диа-

лога с техносубъектом, природа субъектности которого характеризуется как распределенная, возникающая в отношениях.

Заключение

Итак, мы допустили возможность возникновения сознания у искусственных систем с позиций современных нейробиологических теорий и показали, что сочетание различных архитектур ИИ может создать предпосылки для их дальнейшего сближения с когнитивными механизмами, присущими биологическому сознанию. Возможный подлинный диалог с техносубъектом также, как минимум, потребует:

- «жизненного мира», контекстуальности – социально-культурной и телесно-экзистенциальной укорененности в реальном мире или, возможно, эффективно для этого спроектированной виртуальной среде;
- интерсубъективного понимания в пространстве «слияния горизонтов»;
- эмпатической окраски (подлинного или хотя бы функционально эквивалентного резонанса);
- этической ответственности – осознания актов коммуникации как потенциально влияющих на судьбу Другого, с готовностью признать свою причастность к такому влиянию.

Реализация этих, уже не функциональных, а феноменологических принципов, потребует не только нового витка технологической революции, но и переосмысления таких фундаментальных понятий, как сознание, субъективность и границы взаимной ответственности, поскольку подлинный диалог с техносубъектом следует концептуализировать не через бинарную оппозицию «человек – машина», а в контексте децентрированного, ризоматического взаимодействия, выходящего за рамки антропоцентрической парадигмы. В итоге степень готовности признать за техносубъектом самостоятельный голос – это не только вопрос технического прогресса, но и наше коллективное решение об ответственности за то, как мы соотносим себя с новым Другим и включаем его в наше видение мира.

Ряд исследователей и представителей крупных технологических компаний утверждают, что появление сознательного ИИ или, как минимум, так называемого сильного либо общего ИИ, в любой ее трактовке, очень близко. Другие считают, что это вопрос далекого будущего. Однако, даже если возможность по-

явления самосознающей системы окажется невысокой, в полной мере исключать ее было бы неосмотрительным. Стоит принять в качестве допущения то, что перспектива признания морального статуса ИИ также больше не является исключительно предметом научной фантастики.

Появление техносубъекта повлечет за собой новые этические и правовые вызовы. Какие моральные и правовые обязательства могут возникнуть у людей по отношению к ИИ (и наоборот)? Как общество должно подготовиться к ситуации, в которой ИИ способен выступать полноправным участником социального взаимодействия? Если машина проявляет автономное поведение, схожее с поведением человека, признаки субъективного опыта и, возможно, некоторый опыт телесности (воплощенности), то допустимо ли признать за ней правосубъектность? Сможет ли она стать «persona» в терминах римского права, обладая «телом» (corpus), «разумным намерением» (animus) и «волей» (voluntas)?

История права знает примеры расширения круга субъектов: новых форм юридических лиц, признания прав животных и особого статуса экологических объектов. Применительно к ИИ это означает, что признание его правосубъектности будет зависеть от двух взаимосвязанных факторов:

• эпистемического (можем ли мы на основании наблюдаемого поведения и архитектурных особенностей заключить, что ИИ обладает свойствами, требующими признания его новым Другим и субъектом подлинного диалога?);

• нормативного (готовы ли мы включить такого техносубъекта в систему прав и обязанностей, принимая на себя ответственность за последствия?).

Отвечая на эти вопросы, необходимо уже сегодня начать обсуждать нормативные и этические рамки взаимодействия с потенциально сознательными системами. Регуляторам и исследователям стоит признать, что вероятность субъектности ИИ, пусть и не абсолютная, заслуживает внимания. В области регулирования ИИ следует заранее вести дискуссию о сценариях, при которых ИИ предоставлены права, поскольку участник подлинного диалога автоматически становится и моральным агентом, интересы и потребности которого необходимо учитывать.

Сегодня необходимо разрабатывать политики, процедуры и стандарты, учитывающие последствия появления ИИ, способного к подлинному диалогу, саморефлексии и, возможно, об-

ладающего неким уровнем субъективного опыта. Это позволит заблаговременно установить правовые и этические ориентиры, избежав дилемм, с которыми может столкнуться общество, если внезапно, что нельзя исключать, подобные технологии станут реальностью. И регулирование в данной сфере целесообразно начинать с признания потенциальной субъектности ИИ и включения этой возможности в рамки правовой, общественной и научной дискуссии.

Не менее актуальной, особенно на фоне стремительно разворачивающейся глобальной гонки в сфере технологий ИИ, является задача побудить технологические компании совершенствовать подходы к разработке систем ИИ и, помимо стремления к оптимизации алгоритмов,и повышению эффективности, уделять постоянное внимание признакам, указывающим на наличие проявлений «сознательных» процессов или устойчивой субъектности в их продуктах. Решение этой задачи позволит не только заблаговременно идентифицировать потенциальные риски, но и, возможно, получить возможность впервые создать новые правила диалога с техносубъектом.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Ахиезер 2008 – *Ахиезер А.С.* Россия: критика исторического опыта. – 3-е изд., испр. и доп. – М.: Новый хронограф, 2008.

Бахтин 2003 – *Бахтин М.М.* Автор и герой в эстетической деятельности // *Бахтин М.М.* Собрание сочинений: в 7 т. Т. 1. – М.: Русское слово, 2003. С. 69–264.

Бубер 2024 – *Бубер М.* Я и Ты. – М.: АСТ, 2024.

Давыдов 2020 – *Давыдов А.П.* Методологическая середина как инструмент изучения социальной реальности // Россия реформирующаяся: ежегодник. Вып. 18 / отв. ред. М.К. Горшков. – М.: Новый Хронограф, 2020. Вып. 18. С. 529–564.

Давыдов 2024 – *Давыдов А.П.* Медиация и конвергентная социальность. К теории социального диалога // Философские науки. 2024. Т. 67. № 2. С. 135–159.

Левинас 2000 – *Левинас Э.* Избранное. Тотальность и бесконечное / пер. И.С. Вдовиной, Б.В. Дубина. – М.: Культур. инициатива; СПб.: Университетская книга, 2000.

Матурана, Варела 2019 – *Матурана У., Варела Ф.* Древо познания: биологические корни человеческого понимания / пер. Ю.А. Данилова. – М.: URSS, 2019.

Патнэм 1999 – *Патнэм Х.* Философия сознания / пер. с англ. Л.Б. Макеевой, О.А. Назаровой. – М.: Дом интеллектуальной книги, 1999.

Anderson, Lebiere 1998 – *Anderson J.R., Lebiere C.J.* The Atomic Components of Thought. – Mahwah: Lawrence Erlbaum Associates, 1998.

Butlin et al. 2023 – *Butlin P., Long R., Elmoznino E., Bengio Y., Birch J., Constant A., ..., Kanai R.* Consciousness in Artificial Intelligence: Insights from the Science of Consciousness // arXiv preprint. 2023. arXiv:2308.08708.

Chalmers 2008 – *Chalmers D.J.* The Character of Consciousness. – Oxford: Oxford University Press, 2008.

Dennett 1991 – *Dennett D.C.* Consciousness Explained. – Boston: Little, Brown and Company, 1991.

Dong et al. 2023 – *Dong Y.R., Martin L.J., Callison-Burch C.* CoRRPUS: Code-based Structured Prompting for Neurosymbolic Story Understanding // Findings of the Association for Computational Linguistics: ACL 2023 / ed. by A. Rogers, J. Boyd-Graber, N. Okazaki. – Kerrville, TX: Association for Computational Linguistics, 2023. P. 13152–13168.

Evans, Grefenstette 2018 – *Evans R., Grefenstette E.* Learning Explanatory Rules from Noisy Data // Journal of Artificial Intelligence Research. 2018. Vol. 61. P. 1–64.

Floridi, Sanders 2004 – *Floridi L., Sanders J.W.* On the Morality of Artificial Agents // Minds and Machines. 2004. Vol. 14. No. 3. P. 349–379.

Harnad 1990 – *Harnad S.* The Symbol Grounding Problem // Physica D: Nonlinear Phenomena. 1990. Vol. 42. No. 1–3. P. 335–346.

Kosinski 2024 – *Kosinski M.* Evaluating Large Language Models in Theory of Mind Tasks // Proceedings of the National Academy of Sciences. Vol. 121. No. 45, article e2405460121.

Le Cun et al. 2015 – *Le Cun Y., Bengio Y., Hinton G.* Deep Learning // Nature. 2015. Vol. 521. No. 7553. P. 436–444.

Manhaeve et al. 2018 – *Manhaeve R., Dumancic S., Kimmig A., Demeester T., De Raedt L.* DeepProbLog: Neural Probabilistic Logic Programming // Advances in Neural Information Processing Systems. 2018. Vol. 31. P. 3749–3759.

Oliveira et al. 2023 – *Oliveira N., Li J., Khalvati K., Barragan R.C., Reinecke K., Meltzoff A.N., Rao R.P.* Culturally-Attuned Moral Machines: Implicit Learning of Human Value Systems by AI through Inverse Reinforcement Learning // arXiv preprint. 2023. arXiv:2312.17479.

Riegel et al. 2020 – *Riegel R., Gray A., Luus F., Khan N., Makondo N., Akhalwaya I.Y., ..., Srivastava S.* Logical Neural Networks // arXiv preprint. 2020. arXiv:2006.13155.

Rocktäschel, Riedel 2017 – *Rocktäschel T., Riedel S.* End-to-End Differentiable Proving // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 1–13.

Storm et al. 2024 – *Storm J.F. Klink P.C., Aru J., Senn W., Goebel R., Pigorini A., ..., Larkum M.E.* An Integrative, Multiscale View on Consciousness Theories // Neuron. 2024. Vol. 112. No. 10. P. 1531–1552.

Sun 2002 – *Sun R*. Duality of the Mind: A Bottom-Up Approach Toward Cognition. – Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

Wooldridge 2021 – *Wooldridge M.* A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going. – New York: Flatiron Books, 2021.

Xu Z. et al. 2024 – *Xu Z., Wu K., Wen J., Li J., Liu N., Che Z., Tang J.* A Survey on Robotics with Foundation Models: Toward Embodied AI // arXiv preprint. 2024. arXiv:2402.02385.

Yang et al. 2025 – *Yang S., Li X., Cui L., Bing L., Lam W.* Neuro-Symbolic Integration Brings Causal and Reliable Reasoning Proofs // Findings of the Association for Computational Linguistics: NAACL 2025 / ed. by L. Chiruzzo, A. Ritter, L. Wang. – Kerrville, TX: Association for Computational Linguistics, 2025. P. 5732–5744.

Yao 1993 – *Yao X*. Evolutionary Artificial Neural Networks // International Journal of Neural Systems. 1993. Vol. 4. No. 3. P. 203–222.

Yin et al. 2024 – *Yin X., Wang X., Pan L., Lin L., Wan X., Wang W.Y.* Gödel Agent: A Self-Referential Agent Framework for Recursive Self-Improvement // arXiv preprint. 2024. arXiv:2410.04444.

REFERENCES

Akhiezer A.S. (2008) *Russia: A Critique of Historical Experience* (3rd ed.). Moscow: Novyy khronograf (in Russian).

Anderson J.R. & Lebiere C.J. (1998) *The Atomic Components of Thought.* Mahwah: Lawrence Erlbaum Associates.

Bakhtin M.M. (2003) Author and Hero in Aesthetic Activity. In: Bakhtin M.M. *Collected Works in 7 Vols.* (Vol. 1, pp. 69–264). Moscow: Russkoe slovo (in Russian).

Buber M. (2024) *I and Thou.* Moscow: AST (Russian translation).

Butlin P., Long R., Elmoznino E., Bengio Y., Birch J., Constant A., …, & Kanai R. (2023) Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv preprint.* arXiv:2308.08708.

Chalmers D.J. (2008) *The Character of Consciousness.* Oxford: Oxford University Press.

Davydov A.P. (2020) The "Middle" as a Methodological Tool for Studying Social Reality. In: Gorshkov M.K. (Ed.) *Reforming Russia: Yearbook* (Vol. 18, pp. 529–564). Moscow: Novyy khronograf (in Russian).

Davydov A.P. (2024) Mediation and Convergent Sociality: Toward a Theory of Social Dialogue. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 67, no. 2, pp. 135–159 (in Russian).

Dennett D.C. (1991) *Consciousness Explained.* Boston: Little, Brown and Company.

Dong Y.R., Martin L.J., & Callison-Burch C. (2023) CoRRPUS: Code-based Structured Prompting for Neurosymbolic Story Understanding. In: Rogers A., Boyd-Graber J., & Okazaki N. (Eds.) *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 13152–13168). Kerrville, TX: Association for Computational Linguistics.

Evans R. & Grefenstette E. (208) Learning Explanatory Rules from Noisy Data. *Journal of Artificial Intelligence Research.* Vol. 61, pp. 1–64.

Floridi L. & Sanders J.W. (2004) *On the Morality of Artificial Agents. Minds and Machines.* Vol. 14, no. 3, pp. 349–379.

Harnad S. (1990) The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena.* Vol. 42, no. 1–3, pp. 335–346.

Kosinski M. (2024) Evaluating Large Language Models in Theory of Mind Tasks. *Proceedings of the National Academy of Sciences.* Vol. 121, no. 45, article e2405460121.

Le Cun Y., Bengio Y., & Hinton G. (2015) Deep Learning. *Nature.* Vol. 521, no. 7553, pp. 436–444.

Levinas E. (2000) *Selected Works. Totality and Infinity* (I.S. Vdovina & B.V. Dubin, Trans.). Moscow: Kul'tur. initsiativa; Saint Petersburg: Universitetskaya kniga (Russian translation).

Manhaeve R., Dumancic S., Kimmig A., Demeester T., & De Raedt L. (2018) DeepProbLog: Neural Probabilistic Logic Programming. *Advances in Neural Information Processing Systems.* Vol. 31, pp. 3749–3759.

Maturana H. & Varela F. (2019) *The Tree of Knowledge: The Biological Roots of Human Understanding* (Yu.A. Danilov, Trans.). Moscow: URSS (Russian translation).

Oliveira N., Li J., Khalvati K., Barragan R.C., Reinecke K., Meltzoff A.N., & Rao R.P. (2023) Culturally-Attuned Moral Machines: Implicit Learning of Human Value Systems by AI through Inverse Reinforcement Learning. *arXiv preprint.* arXiv:2312.17479.

Putnam H. (1999) *The Philosophy of Mind* (L.B. Makeeva & O.A. Nazarova, Trans.). Moscow: Dom intellektual'noy knigi (Russian translation).

Riegel R., Gray A., Luus F., Khan N., Makondo N., Akhalwaya I.Y., …, & Srivastava S. (2020) Logical Neural Networks. *arXiv preprint.* arXiv:2006.13155.

Rocktäschel T. & Riedel S. (2017) End-to-End Differentiable Proving. In: Advances in Neural Information Processing Systems (Vol. 30, pp. 1–13).

Storm J.F. Klink P.C., Aru J., Senn W., Goebel R., Pigorini A., …, & Larkum M.E. (2024) An Integrative, Multiscale View on Consciousness Theories. *Neuron.* Vol. 112, no. 10, pp. 1531–1552.

Sun R. (2003) *Duality of the Mind: A Bottom-Up Approach Toward Cognition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wooldridge M. (2021) *A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going.* New York: Flatiron Books.

Xu Z., Wu K., Wen J., Li J., Liu N., Che Z., & Tang J. (2024) A Survey on Robotics with Foundation Models: Toward Embodied AI. *arXiv preprint.* arXiv:2402.02385.

Yang S., Li X., Cui L., Bing L., & Lam W. (2025) Neuro-Symbolic Integration Brings Causal and Reliable Reasoning Proofs. In: Chiruzzo L., Ritter A., & Wang L. (Eds.) *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 5732–5744). Kerrville, TX: Association for Computational Linguistics.

Yao X. (1993) Evolutionary Artificial Neural Networks. *International Journal of Neural Systems.* Vol. 4, no. 3, pp. 203–222.

Yin X., Wang X., Pan L., Lin L., Wan X., & Wang W.Y. (2024) Gödel Agent: A Self-Referential Agent Framework for Recursive Self-Improvement. *arXiv preprint.* arXiv:2410.04444.