Отечественная философия
2024. Т. 2. № 2. С. 5–17
УДК 1(091)+130.3+141

National Philosophy
2024, Vol. 2, No. 2, pp. 5–17
DOI: 10.21146/2949-3102-2024-2-2-5-17

ИССЛЕДОВАНИЯ

А.А. Танюшина

Общий искусственный интеллект, информация и сознание: об интерпретации Д.И. Дубровского

Танюшина Александра Александровна – кандидат философских наук, ассистент. Московский государственный университет имени М.В. Ломоносова. Российская Федерация, 119991, г. Москва, Ленинские горы, 1; e-mail: a.tanyushina@gmail.com

Статья посвящена анализу концепции общего искусственного интеллекта (Artificial General Intelligence; AGI) и её интерпретации, предложенной российским философом Давидом Израилевичем Дубровским в его недавних исследовательских работах. В первой части статьи кратко обозначены существующие сегодня подходы к определению понятия «общий искусственный интеллект», в том числе трактующие его как искусственную интеллектуальную систему, способную достигать общих целей в разнообразных средах. Ссылаясь на тексты наиболее влиятельных зарубежных исследователей и разработчиков, автор демонстрирует параллели между предложенными ими подходами к осмыслению общего искусственного интеллекта и теми трактовками, которые предлагаются Дубровским и его соавторами. В частности, показана общность в интерпретациях концепции «модель мира» (Ян ЛеКун) и концепции «техно-умвельт», а также параллели между гипотезой «универсально-воплощённого ИИ» (Бен Герцель) и рассуждениями российского философа о возможной реализации ИИ посредством его вовлечения в различные типы интеракций с разнообразными мирами, виртуальными и физическим. Во второй части статьи обозначен потенциал применения разработанного Дубровским информационного подхода к решению проблемы «сознание-тело» в качестве основания для объяснения феномена общего искусственного интеллекта. Показано, что несмотря на необходимость доработки предложенной философом концепции информационной причинности, его теория может способствовать лучшему пониманию связи возможных компетенций AGI с явлениями субъективной реальности. В заключении обозначаются ключевые проблемы, которые на данный момент затрудняют поиск ответа на вопрос о зависимости качеств общего искусственного интеллекта от наличия феноменального сознания. Делается акцент на необходимости продолжения междисциплинарного сотрудничества между представителями когнитивных наук, разработчиками и философами, чьё взаимодействие призвано способствовать решению характерных трудностей, связанных как с проблемой интерпретации понятия «общий искусственный интеллект», так и с проблемой выявления сознания у искусственных интеллектуальных систем.

Ключевые слова: искусственный интеллект, общий искусственный интеллект, Д.И. Дубровский, информационная теория сознания, универсально-воплощённый искусственный интеллект, «модели мира», субъективная реальность, феноменальное сознание

Для цитирования: *Танюшина А.А.* Общий искусственный интеллект, информация и сознание: об интерпретации Д.И. Дубровского // Отечественная философия. 2024. Т. 2. № 2. С. 5–17.

Российский философ Давид Израилевич Дубровский снискал широкую известность в среде как отечественных, так и зарубежных учёных во многом благодаря развитию последовательного информационного подхода к объяснению феномена сознания. В рамках разработанной им целостной концепции, призванной, в частности, дать решение фундаментальной философской проблемы соотношения сознания и тела, философ не только предлагает информационную интерпретацию различных аспектов мыслительных процессов, но и углубляется в анализ сопутствующих вопросов из области онтологии, теории познания, нейробиологии, теории искусственного интеллекта и других смежных дисциплин.

В последние годы Давид Израилевич активно публикует работы, посвящённые проблематике общего искусственного интеллекта (англ. *artificial general intelligence*; далее – AGI)[1]. Общий искусственный интеллект представляет собой гипотетическую искусственную вычислительную систему, которая обладает способностью понимать и решать любые интеллектуальные задачи на таком же уровне, как и человек. В отличие от «слабого», или «узкого» ИИ, нацеленного на решение конкретных и узкоспециализированных задач, AGI должен демонстрировать такие навыки, как способность к обобщению и переносу знаний из одной области в другую, умение самостоятельно ставить цели, обучаться, рассуждать (*to reason*) и принимать решения в условиях неопределённости. Кроме того, некоторые исследователи также упоминают о таких возможных качествах общего ИИ, как креативность, эмоциональный интеллект и наличие здравого смысла[2].

Разработка полноценного AGI остаётся одной из самых амбициозных и сложных задач в области искусственного интеллекта, требующей значительных теоретических и технологических разработок. Многие эксперты считают, что для создания AGI необходимы фундаментальные прорывы в понимании природы интеллекта, сознания и процессов мышления, причём не только у человека, но и у нечеловеческих животных[3]. По этой причине возможность скорой разработки AGI долгое время ставилась исследователями под сомнение.

Тем не менее появление ChatGPT в ноябре 2022 г. стало настоящим прорывом в области искусственного интеллекта, заставившим экспертов пересмотреть свои прогнозы относительно сроков создания AGI. Эта революционная модель генеративного искусственного интеллекта, основанная на архитектуре нейронной сети-трансформера, продемонстрировала настолько впечатляющие возможности в области обработки естественного языка и генерации осмысленных ответов на запросы пользователя, что многие специалисты существенно приблизили ожидаемые даты появления полноценного искусственного интеллекта, ни в чём не уступающего человеческому разуму. В 2023–2024 гг. стали активно разрабатываться альтернативные генеративные модели и чат-боты (в том числе и общедоступные модели с открытым исходным кодом), что также способствовало значительному ускорению прогресса в области нейросетевого моделирования и глубокого обучения. Таким

---

1 См., напр.: *Дубровский Д.И.* Эпистемологический анализ социогуманитарной значимости новаций искусственного интеллекта в контексте общего искусственного интеллекта // Философские науки. 2022. Т. 65. № 1. С. 10–26; *Дубровский Д.И.* Значение нейронаучных исследований сознания для разработки Общего искусственного интеллекта // Вопросы философии. 2022. № 2. С. 83–93; *Дубровский Д.И.* Сознание, мозг, общий искусственный интеллект: новые стратегические задачи и перспективы // Человек в системе искусственного интеллекта / Под ред. В.А. Лекторского. СПб., 2022. С. 128–159; *Дубровский Д.И.*, *Ефимов А.Р.*, *Матвеев Ф.М.* Что мешает нам создать Общий искусственный интеллект? Одна старая стена и один старый спор // Вопросы философии. 2023. № 5. С. 39–49.

2 *Бостром Н.* Искусственный интеллект. Этапы, угрозы стратегии. М., 2016. С. 13.

3 *Butlin P. et al.* Consciousness in Artificial Intelligence: Insights from the Science of Consciousness [Электронный ресурс] // Arxiv.org. 2023. URL: https://arxiv.org/pdf/2308.08708.pdf (дата обращения: 25.04.2024).

образом, произошедшая за последние два года «GPT-революция» существенно всколыхнула сферу разработки искусственных интеллектуальных систем и повысила планку достижений ИИ-моделей, а также заставила научное сообщество переосмыслить временные горизонты, отделяющие нас от создания AGI, потенциально способного изменить траекторию развития человечества. По этой причине представляется, что последовательный концептуальный, методологический и философско-теоретический анализ концепции AGI является сегодня как никогда актуальным.

Далее в настоящей статье будут рассмотрены основные философско-теоретические основания концепции AGI, в том числе анализируемые Дубровским в его статьях. Также будут разобраны методологические вопросы, связанные с созданием общего искусственного интеллекта, и выявлена роль междисциплинарных подходов к разработке и оценке эффективности интеллектуальных систем, претендующих на статус AGI. Наконец, учитывая теоретическую совместимость интерпретации AGI, предложенной Дубровским, с толкованиями ведущих зарубежных исследователей, будет продемонстрирован потенциал применения информационного подхода к объяснению природы феноменального сознания в качестве философского-теоретического фундамента, на котором может базироваться наше дальнейшее концептуальное и онтологическое описание систем общего искусственного интеллекта.

AGI: проблема определения и оценки

Во введении мы уже предварительно обозначили некоторое самое общее понимание выражения «общий искусственный интеллект». Однако следует отметить, что на данный момент в мире не существует ни одного общепринятого способа интерпретации концепции AGI. Набор качеств и компетенций, которым должна обладать подобная интеллектуальная система, остаётся крайне размытым, ведь такие критерии, как «адаптивность», «универсальность» и др., очевидно, не являются удовлетворительными в силу своей концептуальной неоднозначности.

Как уже было отмечено выше, один из наиболее распространённых подходов к трактовке AGI связан с предложением считать искусственную модель обладающей «общим» интеллектом, если она способна воспроизводить любые функции, которые может выполнять носитель естественного интеллекта, т.е. человек. Однако такого рода интерпретация является крайне проблематичной в силу отсутствия единого стандарта человеческого интеллекта, а также подходящих измерительных систем и средств, способных дать объективную оценку естественных когнитивных способностей. Исследователи подчёркивают, что большинство существующих сегодня психометрических тестов являются неадекватными для измерения интеллектуальных способностей человека, и по этой причине оформление аналогичных тестов для оценки интеллектуального уровня искусственных моделей также представляется затруднительным[4]. Так, стандартный тест Тьюринга, а также все его возможные вариации («минимальный тест Тьюринга», «социальный тест Тьюринга» и др.), нацеленные на выявление имитационных и функциональных способностей искусственных вычислительных систем, уже признаются ведущими исследователями неактуальными для оценки эффективности существующих сегодня больших языковых моделей[5]. Дубровский в своих работах также отмечает непригодность теста

---

[4] *Карелов С.В.* «Ловушка Гудхарта» для AGI: проблема сравнительного анализа искусственного интеллекта и интеллекта человека // Учёные записки Института психологии Российской академии наук. 2023. Т. 3. № 3. С. 5–22.

[5] *Bubeck S., Chandrasekaran V., Eldan R., Gehrke J., Horvitz E., Kamar E. et al.* Sparks of Artificial General Intelligence: Early Experiments with GPT-4 [Электронный ресурс] // Cornell University. 2023. URL: https://arxiv.org/pdf/2303.12712.pdf (дата обращения: 25.04.2024).

Тьюринга для оценки деятельности как существующих сегодня моделей, так и будущих систем, потенциально способных стать основанием для разработки AGI:

> Для создания AGI в первую очередь следует преодолеть ограниченности традиционной методологии разработки ИИ, основы которой заложены А. Тьюрингом. Согласно этой методологии, понятие «интеллект» трактуется в сугубо функционально-операциональном смысле, исключающем роль сознания, вся сложная проблематика которого выносится за скобки. Исследование сознания представляется излишним в данном случае: существует когнитивная или практическая задача, которая формулируется и решается посредством сугубо операциональных методов с помощью компьютерных программ… Однако сегодня мы переходим к новому этапу, нуждающемуся в существенных теоретико-методологических обновлениях. Парадигма функционализма, в формирование которой А. Тьюрингом внесён первостепенный вклад, естественно, остаётся действующей, но она требует более широкой интерпретации с учётом её роли в объяснении сознания и способов использования новейших результатов исследований сознания для развития ИИ[6].

Другой подход к интерпретации AGI, предложенный ещё в 2000-е гг., основан на представлении об универсальной целенаправленности искусственного агента, предполагающем способность AGI достигать разнообразных целей в разнообразных условиях и средах[7]. Так, футуролог и специалист по компьютерным наукам Бен Герцель, в своё время популяризировавший выражение “artificial general intelligence”, определил ключевую компетенцию общего искусственного интеллекта как «способность достигать общих целей в сложных средах»[8]. Подобное определение представляется многим исследователям уже более уместным для дальнейшего конструирования интеллектуальных систем и оценки их деятельности, ведь разработчики, занимающиеся машинным обучением, уже достаточно давно предлагают развивать посттьюринговую оценочную методологию, предполагающую, помимо прочего, критический анализ перспектив развития ИИ-систем в контексте их способностей к генерации осмысленного текста.

Наиболее влиятельной и цитируемой работой, посвящённой данной проблематике, является статья «Искусственный интеллект и пределы языка», написанная в 2022 г. философом Якобом Браунингом и специалистом по машинному обучению Яном ЛеКуном[9]. В данной работе учёные утверждают, что, несмотря на впечатляющие достижения больших языковых моделей в области обработки естественного языка, они по-прежнему страдают от фундаментальных ограничений, обусловленных самой природой их архитектуры и принципами обучения: будучи натренированными на огромных текстовых корпусах, подобные системы способны улавливать статистические закономерности и корреляции в данных, но тем не менее не обладают истинным пониманием смысла и контекста. Иными словами, они всё ещё оперируют данными на уровне синтаксиса, а не содержания, что приводит к генерации правдоподобных, но зачастую бессмысленных или некогерентных высказываний. С точки зрения авторов, естественный язык – это не просто система знаков и правил, но неотъемлемая часть человеческого познания, укоренённая в нашем сенсомоторном опыте, социальном взаимодействии и культурном контексте.

---

6 *Дубровский Д.И.* Задача создания Общего искусственного интеллекта и проблема сознания // Философские науки. 2021. Т. 64. № 1. С. 18.

7 *Legg S., Hutter M.* Universal Intelligence: A Definition of Machine Intelligence // Minds and Machines. 2007. No. 17 (4). P. 394.

8 *Goertzel B.* Contemporary Approaches to Artificial General Intelligence / Artificial General Intelligence (Goertzel B., Pennachin C. eds.). Springer. 2005. P. 22.

9 *Browning J., LeCun Y.* AI And The Limits Of Language [Электронный ресурс] // Noemamag.com. 2022. URL: https://www.noemamag.com/ai-and-the-limits-of-language/ (дата обращения: 25.04.2024).

Большие языковые модели принципиально ограничены в таких аспектах, как понимание, рассуждение и абстрактное мышление, поскольку они обучаются исключительно на текстовых данных, без возможности активного взаимодействия с окружающей средой. Для создания по-настоящему интеллектуальных машин, способных к глубокому пониманию языка и мышлению, потребуются принципиально новые подходы, выходящие за рамки простого обучения моделей на больших данных с целью имитации языковых паттернов. Разработка подобных подходов потребует тесного междисциплинарного сотрудничества специалистов в области ИИ, лингвистики, когнитивной науки и философии.

Ян ЛеКун предложил развивать генеративные интеллектуальные системы в рамках целеориентированного подхода к разработке ИИ (*objective-driven AI*), где ключевой акцент делается на идее развития у искусственных нейросетевых структур «моделей мира», т.е. внутренних представлений системы об окружающей среде, в которой она функционирует[10]. Наличие «модели мира» подразумевает интеграцию знаний, убеждений и допущений системы о структуре, свойствах и закономерностях мира: сюда относятся понимание системой типов объектов, понятий, сущностей и отношений между ними, представление о причинно-следственных связях, необходимых для предиктивной работы (а также учёт вероятностной природы мира), знания о физических свойствах среды и пространственных отношениях, понимание контекста и, наконец, понимание действий других когнитивных агентов. Для построения подобных богатых и согласованных внутренних репрезентативных структур необходимы методы обучения ИИ-систем, основанные на иерархическом планировании, а также принципах универсально-воплощённой агентности.

Разработка универсально-воплощённого искусственного интеллекта (*universal embodied AI*) предполагает создание систем обучения, в рамках которых интеллектуальные агенты обладают как физическим, так и виртуальным воплощением, что обеспечивает их способность взаимодействовать с реальной или смоделированной средой[11]. В отличие от систем «узкого ИИ», функционирующих исключительно в абстрактном символическом пространстве, универсально-воплощённые ИИ-агенты могут быть обучены в виртуальных и игровых средах, имитирующих реальные физические взаимодействия, а также могут быть интегрированы в робототехнические устройства, обладающие сенсорными и моторными системами и позволяющие ИИ воспринимать окружающий мир, манипулировать объектами и перемещаться в пространстве. Похожие соображения высказывает и Дубровский в статье «Что мешает нам создать Общий искусственный интеллект? Одна старая стена и один старый спор», написанной в соавторстве с А.Р. Ефимовым и Ф.М. Матвеевым:

> Общий искусственный интеллект, то есть машина, которая способна мыслить и действовать подобно человеку, неизбежно будет следствием, продуктом, зафиксированным в опыте разносторонней интеракции человека и окружающего мира. Такая интеракция может быть вербальной или невербальной. Она может проходить в обоих видах реальности – в виртуальной и реальной[12].

Создание агентов, способных автономно и целенаправленно действовать в различных физических и мультимодальных виртуальных средах, безусловно, вынуждает

---

[10] *LeCun Y.* A Path Towards Autonomous Machine Intelligence [Электронный ресурс] // OpenReview. 2022. URL: https://openreview.net/pdf?id=BZ5a1r-kVsf (дата обращения: 25.04.2024).

[11] *Luo J., Longfei M., Chang Z., Fei W., Hongxia Y.* Universal Embodied Intelligence: Learning from Crowd, Recognizing the World, and Reinforced with Experience. [Электронный ресурс] // OpenReview. 2023. URL: https://openreview.net/pdf?id=3e5nHhhRK93 (дата обращения: 25.04.2024).

[12] *Дубровский Д.И., Ефимов А.Р., Матвеев Ф.М.* Что мешает нам создать Общий искусственный интеллект? Одна старая стена и один старый спор // Вопросы философии. 2023. № 5. С. 42–45.

исследователей не только искать новые методологические подходы к конструированию условий для обучения подобных ИИ-агентов, но и формулировать новые теоретические и философско-концептуальные рамки для анализа концепции «модели мира» (или аналогичной концепции техно-умвельта, предложенной одним из соавторов Дубровского Альбертом Рувимовичем Ефимовым, определившим «техно-умвельт» как «домен мировосприятия – то, что и как воспринимает машина, отображая окружающий мир»[13]). Давид Израилевич и соавторы также отмечают, что для реализации искусственных интеллектуальных агентов нужно ввести слой специальных интерпретаторов, которые смогут переводить онтологии и опыт одной области в понятия и опыт другой[14].

Однако для дальнейших исследований AGI необходимо более конкретное понимание когнитивных и ментальных факторов, влияющих на формирование «моделей мира» у носителей естественного интеллекта. Именно на этом этапе среди исследователей и возникают ключевые разногласия: подразумевая, что данные «модели мира» являются внутренними представлениями процессов, происходящих в окружающей агента среде, учёные и разработчики, как правило, говорят исключительно об их репрезентативно-функциональной роли: «модели мира» необходимы для проектирования последствий действия агента в среде и предсказания исхода будущих вероятностных событий, поэтому само по себе обладание «моделью мира» не подразумевает, к примеру, наличия высокоуровневых ментальных свойств вроде феноменального сознания. Так, в книге 2021 г. «Сильный искусственный интеллект. На подступах к сверхразуму», на которую также в своих статьях ссылается и Дубровский, авторы отмечают:

> Делая акцент на широком диапазоне сред, область AGI позволяет нам избавиться от антропоцентричных предпочтений и предлагает сфокусироваться на общих решениях, пригодных для разных агентов (человека, животных, роботов, ботов и т.д.), действующих в разных условиях… Как повышение уровня решения узких задач не потребовало «сильных» качеств, так и расширение общности методов решения задач вовсе не обязательно подразумевает преднамеренное движение в сторону сильного ИИ[15, 16].

С другой стороны, при анализе возможных качеств и компетенций, присущих потенциальному AGI, некоторые разработчики также упоминают «способность к интроспекции», «конструирование агентом образа себя», «эмпатичное поведение», «рефлексию» и т.д. Именно на этот факт и обращает внимание Д.И. Дубровский в своих работах: «Современный специалист в области ИИ как бы забывает о том, что является носителем ЕИ [естественного интеллекта] и обладает искомыми свойствами AGI, постоянно применяет их в своей профессиональной работе, несмотря на то, что

---

[13] *Дубровский Д.И., Ефимов А.Р., Матвеев Ф.М.* Что мешает нам создать Общий искусственный интеллект? Одна старая стена и один старый спор // Вопросы философии. 2023. № 5. С. 46.

[14] Стоит отметить, что подобные системы уже активно разрабатываются: можно упомянуть ИИ-модель Newton от компании Archetype AI, которая изучает физический мир непосредственно на основе данных любых датчиков (камер, микрофонов, термометров, радаров, лидаров, инфракрасных датчиков и др.), становясь единой базовой моделью, поддерживаемой универсальным семантическим метаязыком. Различные исследовательские лаборатории также разрабатывают агентов, способных к автономному обучению в виртуальных игровых средах (например, ИИ-агент Sima от Google DeepMind), а крупнейшие робототехнические компании интегрируют большие языковые модели в своих антропоморфных роботов, способных самостоятельно взаимодействовать с окружающей средой и человеком (например, робот-гуманоид от компании Figure, работающий на ChatGPT-4).

[15] Сильный искусственный интеллект. На подступах к сверхразуму / Науч. ред. А.С. Потапов. М., 2021. С. 29–30.

[16] Отметим, что в цитируемом исследовании под «сильным» ИИ подразумевается человекоподобная интеллектуальная система, обладающая сознанием.

они не выделяются им, остаются скрытыми для него в процессах его сознательной деятельности, направленной на разработку AGI»[17].

По этой причине следующим важным вопросом, тесно связанным с проблемой разработки последовательного теоретического и методологического аппарата, пригодного для анализа свойств AGI, является вопрос о том, как связаны компетенции общего искусственного интеллекта с наличием феноменального ментального опыта. Именно данной проблеме Давид Израилевич и посвящает многие из своих недавних статей и исследований.

Субъективная реальность как необходимое качество AGI: информационный подход

Выражение «субъективная реальность», предназначенное для описания уникальных приватных качеств ментального опыта, Дубровский впервые использовал ещё в своей знаменитой статье 1968 г. «Мозг и психика»[18]. Данное понятие определённым образом соотносится с понятием «объективной реальности», применяемой для характеристики физических структур, которые, в свою очередь, могут выступать в качестве носителей субъективной реальности. Согласно философу, некоторые сложные физические системы, обладающие достаточно высокой степенью структурно-функциональной организации, являются носителями информационных свойств, которые при должной кодовой обработке могут быть представлены как внутренние приватные состояния данных систем: будучи зафиксированными в нейродинамической структуре мозга, эти информационные качества представляют собой феноменальное содержание психики человека, обозначаемое философом термином «субъективная реальность»[19]. Однако поскольку подобная информационная структура, ассоциированная с появлением ментальных переживаний, может быть реализована на любых функционально-изоморфных носителях, отличающихся по своим физическим характеристикам от естественных носителей сознания, субъективная реальность может появляться и у искусственных интеллектуальных систем, сконструированных из небиологических субстратов.

Таким образом, в рамках своего информационного подхода Д.И. Дубровский предпринимает попытку решения фундаментальной психофизической проблемы путём указания на несводимость сознания к простому набору функционально-организационных свойств и когнитивных компетенций, что обыкновенно препятствует последовательному объяснению онтологической реальности феноменального опыта. Субъективные ментальные переживания, по утверждению философа, являются необходимым условием реализации интеллектуальным агентом его интегральных функций, в числе которых сенсомоторные и когнитивные функции, отвечающие за поддержание организмом своей целостности[20].

Таким образом, рассуждая о проблемах конструирования AGI, Давид Израилевич пишет:

---

[17] *Дубровский Д.И.* Сознание, мозг, общий искусственный интеллект: новые стратегические задачи и перспективы // Человек в системе искусственного интеллекта / Под ред. В.А. Лекторского. СПб., 2022. С. 135.

[18] *Дубровский Д.И.* Мозг и психика // Вопросы философии. 1968. № 8. С. 125–135.

[19] *Дубровский Д.И.* Психические явления и мозг: философский анализ проблемы в связи с некоторыми актуальными задачами нейрофизиологии, психологии и кибернетики. М., 1971. С. 292.

[20] *Дубровский Д.И.* Зачем субъективная реальность, или «Почему информационные процессы не идут в темноте?» (Ответ Д. Чалмерсу) // Сознание, мозг, искусственный интеллект: сб. статей. М., 2007. С. 159.

> Качество [субъективной реальности] нередко выносится за скобки теми, кто занимается задачами сугубо функционалистского, бихевиорального типа (как «сделать» что-то, в том числе создать программу для компьютера или создать робота). В большинстве описанных случаев нет необходимости выделять и анализировать качество субъективной реальности, поскольку внимание сосредоточено на поставленной задаче и наличном процессе деятельности; мы будто не замечаем это качество, и даже когда напрягаем, корректируем собственное мышление; оно подобно воздуху, который мы тоже «не замечаем» во многих интервалах жизни[21].

Однако, по утверждению философа, специфические качества феноменального опыта, связанные прежде всего с его интегральными функциями, должны учитываться при построении AGI:

> Возникновение у животных СР [субъективной реальности] стало исторически первой формой виртуальной реальности, открывающей по мере развития всё более широкий диапазон способностей к абстрагированию, обобщению, прогнозированию, планированию, пробным действиям в виртуальном плане, иным схожим операциям (в форме «мысленных экспериментов»), повышающим приспособляемость к среде. Это качество виртуальности во многом определяет способность формирования новых навыков, приспособления к изменившейся среде, т.е., по существу, главное свойство, необходимое для AGI, – способность определять и самостоятельно решать задачи «в широком диапазоне сред»[22].

Представляется, что именно на этом этапе рассуждений появляются основные онтологические и методологические вопросы, связанные с дальнейшим анализом вопроса о необходимости наличия сознания у AGI-систем. С одной стороны, на сегодняшний день наука и философия ещё не уточнили вопрос о наличии прямых логических и номических связей между качествами феноменального опыта (интенциональность, наличие приватных ментальных переживаний, ощущение «каково это быть» носителем данного состояния и т.д.) и когнитивными способностями, которые современные исследователи склонны приписать будущему AGI (автономное поведение в различных окружающих средах, способность к рассуждению, планированию, пониманию контекстных отношений и др.). С другой стороны, даже при условии допущения, что феноменальный опыт может быть расценён как непременный конститутивный элемент перечисленных выше интеллектуальных компетенций, мы всё ещё сталкиваемся с вопросом о каузальной роли субъективных переживаний, который является одним из центральных внутри философской проблемы «сознание-тело».

Отстаивая тезис о несводимости сознания к простому набору функциональных свойств системы, Д.И. Дубровский развивает концепцию информационной причинности, по которой физические и определяемые своей информационной структурой ментальные процессы являются одновременными и однопричинными и находятся в отношении «взаимооднозначного соответствия» в силу кодовой зависимости, определяющей отношения между информацией и её носителем. Иллюстрируя концепцию информационной причинности, философ использует следующий пример:

> Возьмём простой пример: я говорю вам: «Подайте мне, пожалуйста, эту книгу», и вы даёте её мне. Очевидно, что это действие зависит не от самих по себе физических свойств звукового сигнала, а именно от информации на основе сложившейся у вас кодовой зависимости. То же самое действие я могу вызвать у вас множеством других сигналов с другими физическими свойствами. В этом и состоит специфика информационной причинности, которая не противоречит физической причинности, но представляет по сравнению с ней новый тип вызываемых агентом следствий, характерный

---

[21] *Дубровский Д.И.* Сознание, мозг, общий искусственный интеллект: новые стратегические задачи и перспективы // Человек в системе искусственного интеллекта / Под ред. В.А. Лекторского. СПб., 2022. С. 131.

[22] Там же. С. 141–142.

для причинно-следственных отношений в функционировании биологических, социальных и ряда технических систем. Психическая, ментальная причинность, стоявшая для естествознания под большим вопросом, является видом информационной причинности. Тем самым она получает научное объяснение, позволяет дать обоснованный ответ на классический вопрос о воздействии ментального на физическое[23].

Доказывая таким образом каузальную действенность феноменального сознания, Давид Израилевич показывает, что именно явления субъективной реальности играют решающую роль в процессах управления, интеграции и саморегуляции высокоразвитых когнитивных агентов, которые, в свою очередь, должны стать ключевыми и для интеллектуальных систем вроде AGI[24].

С другой стороны, современные разработчики и философы задаются встречным вопросом: каких качеств не хватает современным ИИ-системам, чтобы обладать сознанием? Рассматривая различные варианты ответа на данный вопрос в своей статье «Могут ли большие языковые модели быть сознательными?», австралийский философ Дэвид Чалмерс, в своё время высоко оценивший информационный подход Дубровского, перечисляет следующие возможные пункты: отсутствие у ИИ органической формы реализации, отсутствие воплощённости и чувств (senses), отсутствие глобального рабочего пространства или способности к рекуррентной обработке информации, отсутствие объединённой агентности и автономности и, наконец, отсутствие у ИИ «модели мира» и «модели себя»[25]. Австралийский философ приходит к выводу, что некоторые из этих пунктов опираются на весьма спорные предпосылки о сознании (например, на утверждение о том, что сознание требует биологии, с которым также не согласен и Д.И. Дубровский), в то время как самые сильные возражения, с точки зрения Чалмерса, связаны как раз с отсутствием у современных ИИ-моделей глобального рабочего пространства, способности к рекуррентной обработке информации или других факторов, с которыми, согласно наиболее актуальным сегодня нейробиологическим теориям сознания, могут быть ассоциированы высокоуровневые субъективные переживания[26]. Обязательное наличие «моделей мира» для появления сознания, полагает философ, является далеко не очевидным

---

[23] *Дубровский Д.И.* Сознание, мозг, общий искусственный интеллект: новые стратегические задачи и перспективы // Человек в системе искусственного интеллекта / Под ред. В.А. Лекторского. СПб., 2022. С. 141–142.

[24] Отметим, что именно концепция информационной причинности представляется самым уязвимым местом в теории Давида Израилевича: наделение информационных структур особенными причинными качествами, несводимыми к физическим каузальным свойствам, является, на наш взгляд, решением, нуждающимся как минимум в более последовательном обосновании: если феномены субъективной реальности являются результатом «кодового преобразования» некоторой информации и при этом, по утверждению философа, обладают особыми каузальными силами, то возникает вопрос о том, почему именно «преобразованная» когнитивной системой информация является каузально релевантной, в то время как «непреобразованные» информационные структуры остаются без должной каузальной силы. С другой стороны, если допустить, что информация, не подвергнутая кодовому преобразованию, также обладает должной причинной силой, необходимой для управления и самоорганизации организма, встаёт вопрос о необходимости более высокоуровневых информационных процессов, ассоциированных с субъективными феноменальными переживаниями. Иначе говоря, информационный подход к объяснению сознания пока не даёт последовательного решения проблемы ментальной каузальности и, как следствие, не способен дать точный ответ на вопрос, является ли феноменальный субъективный опыт определяющим фактором реализации высокоуровневых когнитивных свойств. По этой причине данный аспект теории нуждается, на наш взгляд, в некоторой концептуальной доработке.

[25] *Chalmers D.J.* Could a Large Language Model Be Conscious? [Электронный ресурс] // Boston Review. 2023. URL: https://philpapers.org/archive/CHACAL-3.pdf (дата обращения: 25.04.2024).

[26] Подробнее см.: *Butlin P. et al.* Consciousness in Artificial Intelligence: Insights from the Science of Consciousness [Электронный ресурс] // Arxiv.org. 2023. URL: https://arxiv.org/pdf/2308.08708.pdf (дата обращения: 25.04.2024).

предположением; более того, на сегодняшний день мы уже не можем с уверенностью утверждать, что у некоторых уже существующих ИИ-систем полностью отсутствуют простейшие «модели мира» (в то время как в наличии у них феноменальных переживаний на данный момент сомневается большинство исследователей).

Таким образом, можно заключить, что на данный момент у исследователей нет надёжных оснований полагать, что наличие у ИИ «модели мира» может быть обусловлено наличием у подобной системы высокуровневных ментальных свойств, ассоциированных с ментальными переживаниями и явлениями субъективной реальности; в то же время ведущие философы и учёные также не берутся утверждать, что разработка систем AGI с внутренними репрезентациями процессов, происходящих в окружающей среде, автоматически подразумевает появление у подобных систем приватного сознания.

Заключение

Сегодня эксперты, специализирующиеся на философии ИИ, акцентируют внимание на двух ключевых проблемах, значительно затрудняющих дальнейшее развитие концепции AGI и её осмысление в контексте вопросов о связи «общих» интеллектуальных способностей с феноменальным сознанием. Первая проблема сводится к вопросу о природе сознания у человека и нечеловеческих животных, отсутствие последовательного ответа на который препятствует дальнейшему ходу рассуждений о возможности появления сознания у искусственных интеллектуальных систем. Вторая проблема связана с трудностями интерпретации работы больших языковых моделей, до сих пор функционирующих по принципу «чёрного ящика»: в силу того, что разработчики подобных систем всё ещё не до конца понимают, что именно происходит «внутри» данных моделей, затруднено и толкование факторов, при соблюдении которых подобные модели в какой-то момент смогут достичь уровня AGI или стать феноменально сознательными.

Для решения этих проблем исследователи предлагают прицельнее рассмотреть существующие сегодня нейробиологические подходы к объяснению сознания у людей и животных, чтобы затем изучить возможность их адаптации под потребности, связанные с изучением систем AGI[27] (о чём также регулярно упоминает и Д.И. Дубровский, отмечая необходимость детального исследования нейронных коррелятов сознания для дальнейшей разработки общего ИИ[28]).

Представляется, что ключевым критерием успешности подобных исследований является обеспечение продуктивного междисциплинарного взаимодействия между учёными-когнитивистами, специалистами по машинному обучению и философами, способными критически осмыслить теоретические и методологические границы исследования, а также подвергнуть тщательному концептуальному анализу существующие сегодня понятия и определения, связанные с искусственными интеллектуальными системами. Зачастую невозможность прийти к консенсусу по поводу таких важных категорий, как, например, «общий искусственный интеллект», является закономерным следствием отсутствия устойчивого сотрудничества между представителями различных научных дисциплин и крупными технологическими корпорациями, занимающимися самостоятельным изучением и разработкой больших языковых моделей. Необходимо подчеркнуть, что именно специалисты в области

---

[27] См., напр.: *Le Doux J., Birch J., Andrews K., Clayton N.S.* Consciousness Beyond the Human Case // Current Biology. 2023. No. 33 (16). P. 832–840.

[28] *Дубровский Д.И.* Значение нейронаучных исследований сознания для разработки Общего искусственного интеллекта // Вопросы философии. 2022. № 2. С. 90–93.

философии ИИ, философии сознания и философии науки и техники потенциально способны успешно осуществлять взаимный перевод терминологических аппаратов различных дисциплин и обеспечивать концептуализацию и интеграцию существующих сегодня способов решения как «проблемы сознания у ИИ», так и «проблемы AGI». По этой причине системный междисциплинарный подход, предложенный Давидом Израилевичем Дубровским, представляется сегодня весьма актуальным, а разрабатываемая им информационная теория сознания при должной доработке способна стать подходящим онтологическим основанием для решения проблемы сознания у искусственного интеллекта.

Список литературы

Бостром Н. Искусственный интеллект. Этапы, угрозы стратегии. М.: Манн, Иванов и Фербер, 2016. 496 с.

Дубровский Д.И. Психические явления и мозг: философский анализ проблемы в связи с некоторыми актуальными задачами нейрофизиологии, психологии и кибернетики. М.: Наука, 1971. 400 с.

Дубровский Д.И. Сознание, мозг, искусственный интеллект. М.: Стратегия-Центр, 2007. 272 с.

Дубровский Д.И. Задача создания Общего искусственного интеллекта и проблема сознания // Философские науки. 2021. Т. 64. № 1. С. 13–44.

Дубровский Д.И. Значение нейронаучных исследований сознания для разработки Общего искусственного интеллекта // Вопросы философии. 2022. № 2. С. 83–93.

Дубровский Д.И. Мозг и психика // Вопросы философии. 1968. № 8. С. 125–135.

Дубровский Д.И. Проблема «Сознание и мозг»: теоретическое решение. М.: Канон+, 2015. 208 с.

Дубровский Д.И. Сознание, мозг, общий искусственный интеллект: новые стратегические задачи и перспективы // Человек в системе искусственного интеллекта / Под ред. В.А. Лекторского. СПб.: Юридический центр, 2022. С. 128–159.

Дубровский Д.И. Эпистемологический анализ социогуманитарной значимости новаций искусственного интеллекта в контексте общего искусственного интеллекта // Философские науки. 2022. Т. 65. № 1. С. 10–26.

Дубровский Д.И., Ефимов А.Р., Матвеев Ф.М. Что мешает нам создать Общий искусственный интеллект? Одна старая стена и один старый спор // Вопросы философии. 2023. № 5. С. 39–49.

Карелов С.В. «Ловушка Гудхарта» для AGI: проблема сравнительного анализа искусственного интеллекта и интеллекта человека // Учёные записки Института психологии Российской академии наук. 2023. Т. 3. № 3. С. 5–22.

Сильный искусственный интеллект. На подступах к сверхразуму / Науч. ред. А.С. Потапов. М.: Альпина паблишер, Интеллектуальная Литература, 2021. 236 с.

Browning J., LeCun Y. AI And The Limits Of Language [Электронный ресурс] // Noemamag.com. 2022. URL: https://www.noemamag.com/ai-and-the-limits-of-language/ (дата обращения: 25.04.2024).

Bubeck S. et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4 [Электронный ресурс] // Cornell University. 2023. URL: https://arxiv.org/pdf/2303.12712.pdf (дата обращения: 25.04.2024).

Butlin P. et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness [Электронный ресурс] // Arxiv.org. 2023. URL: hhttps://arxiv.org/pdf/2308.08708.pdf (дата обращения: 25.04.2024).

Chalmers D.J. Could a Large Language Model Be Conscious? [Электронный ресурс] // Boston Review. 2023. URL: https://philpapers.org/archive/CHACAL-3.pdf (дата обращения: 25.04.2024).

Goertzel B. Contemporary Approaches to Artificial General Intelligence / Artificial General Intelligence (Goertzel B., Pennachin C. eds.). Springer. 2005. P. 1–28.

Le Doux J., Birch J., Andrews K., Clayton N.S. Consciousness Beyond the Human Case // Current Biology. 2023. No. 33 (16). P. 832–840.

LeCun Y. A Path Towards Autonomous Machine Intelligence [Электронный ресурс] // OpenReview. 2022. URL: https://openreview.net/pdf?id=BZ5a1r-kVsf (дата обращения: 25.04.2024).

Legg S., Hutter M. Universal Intelligence: A Definition of Machine Intelligence // Minds and machines. 2007. No. 17 (4). P. 391–444.

Luo J., Longfei M., Chang Z., Fei W., Hongxia Y. Universal embodied intelligence: learning from crowd, recognizing the world, and reinforced with experience. [Электронный ресурс] // OpenReview. 2023. URL: https://openreview.net/pdf?id=3e5HhhRK93 (дата обращения: 25.04.2024).

General Artificial Intelligence, Information and Consciousness: on D.I. Dubrovsky's Interpretation

Alexandra A. Tanyushina – Candidate of Sciences in Philosophy, associate researcher. Lomonosov Moscow State University. 1 Leninskie Gory, Moscow, 119991, Russian Federation; email: a.tanyushina@gmail.com

The article is devoted to the analysis of the concept of artificial general intelligence (AGI) and its interpretation proposed by the Russian philosopher David Dubrovsky in his recent research papers. The first part of the article briefly outlines the current approaches to defining the concept of "artificial general intelligence", including interpreting it as an artificial intelligent system capable of achieving common goals in a variety of environments. Referring to the texts of the most influential foreign researchers and developers, the author demonstrates the parallels between their proposed approaches to understanding general artificial intelligence and those interpretations proposed by David Dubrovsky and his co-authors. In particular, the commonality in the interpretations of the concept of the "world model" (Yan LeCun) and the concept of "techno-umwelt" is shown, as well as parallels between the hypothesis of "universal embodied AI" (Ben Herzel) and the arguments of the Russian philosopher about the possible implementation of AI through its involvement in various types of interactions with various worlds, virtual and physical. In the second part of the article, the potential of using the information approach developed by David Dubrovsky to solve the mind-body problem as a basis for explaining the phenomenon of general artificial intelligence is outlined. It is shown that despite the need to refine the concept of information causality proposed by the philosopher, his theory can contribute to a better understanding of the connection of possible AGI competencies with the phenomena of subjective reality. In conclusion, the key problems that currently make it difficult to find an answer to the question of the dependence of the qualities of general artificial intelligence on the presence of phenomenal consciousness are outlined. The emphasis is placed on the need to continue interdisciplinary cooperation between representatives of cognitive sciences, developers and philosophers, whose interaction is designed to help solve the characteristic difficulties associated with both the problem of conceptualizing the concept of "artificial general intelligence" and the problem of identifying consciousness in artificial intelligent systems.

Keywords: artificial intelligence, general artificial intelligence, D.I. Dubrovsky, information theory of consciousness, universal embodied artificial intelligence, "world models", subjective reality, phenomenal consciousness

For citation: Tanyushina, A.A. Obshchij iskusstvennyj intellekt, informatsiya i soznanie: ob interpretatsii D.I. Dubrovskogo [General Artificial Intelligence, Information and Consciousness: on D.I. Dubrovsky's Interpretation], *Otechestvennaya filosofiya* [National Philosophy], 2024, Vol. 2, No. 2, pp. 5–17. (In Russian)

References

Bostrom, N. *Iskusstvennyj intellekt. Etapy, ugrozy strategii* [Superintelligence: Paths, Dangers, Strategies]. Moscow: Mann, Ivanov i Ferber, 2016. 496 c. (In Russian)

Browning, J., LeCun, Y. AI And The Limits Of Language, *Noemamag.com*, 2022. URL: https://www.noemamag.com/ai-and-the-limits-of-language/ (25.04.2024).

Bubeck, S., Chandrasekaran V., Eldan R, Gehrke J., Horvitz E., Kamar E. et al. Sparks of Artificial General Intelligence: Early Experiments with GPT-4, *Cornell University*, 2023. URL: https://arxiv.org/pdf/2303.12712.pdf (25.04.2024).

Butlin, P. et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness, *Arxiv.org*, 2023. URL: https://arxiv.org/pdf/2308.08708.pdf (25.04.2024).

Chalmers, D.J. Could a Large Language Model Be Conscious? *Boston Review*, 2023. URL: https://philpapers.org/archive/CHACAL-3.pdf (25.04.2024).

Goertzel, B. Contemporary Approaches to Artificial General Intelligence, *Artificial General Intelligence* (Goertzel B., Pennachin C. eds.). Springer, 2005, pp. 1–28.

Dubrovskij, D.I. *Psikhicheskie yavleniya i mozg: filosofskij analiz problemy v svyazi s nekotorymi aktualnymi zadachami nejrofiziologii, psikhologii i kibernetiki* [Mental Phenomena and the Brain: a Philosophical Analysis of the Problem in Connection with Some Urgent Tasks of Neurophysiology, Psychology and Cybernetics]. Moscow: Nauka, 1971. 400 c. (In Russian)

Dubrovskij, D.I. *Soznanie, mozg, iskusstvennyj intellect* [Consciousness, Brain, Artificial Intelligence]. Moscow: 2007. 272 c. (In Russian)

Dubrovskij, D.I. Zadacha sozdaniya Obshchego iskusstvennogo intellekta i problema soznaniya [The Task of Creating a Common Artificial Intelligence and the Problem of Consciousness], *Filosofskie nauki* [Russian Journal of Philosophical Sciences], 2021, Vol. 64, No. 1, pp. 13–44. (In Russian)

Dubrovskij, D.I. Znachenie nejronauchnykh issledovanij soznaniya dlya razrabotki Obshchego iskusstvennogo intellekta [The Importance of Neuroscientific Consciousness Research for the Development of General Artificial Intelligence], *Voprosy filosofii* [The Problems of Philosophy], 2022, No. 2, pp. 83–93. (In Russian)

Dubrovskij, D.I. Mozg i psikhika [Brain and Psyche], *Voprosy filosofii* [The Problems of Philosophy], 1968, No. 8, pp. 125–135. (In Russian)

Dubrovskij, D.I. *Problema "Soznanie i mozg": teoreticheskoe reshenie* [The Problem of "Consciousness and the Brain": a Theoretical Solution]. Moscow: Kanon+, 2015. 208 c. (In Russian)

Dubrovskij, D.I. Soznanie, mozg, obshchhij iskusstvennyj intellekt: novye strategicheskie zadachi i perspektivy [Consciousness, Brain, General Artificial Intelligence: New Strategic Tasks and Prospects], *Chelovek v sisteme iskusstvennogo intellekta* [The Human Being in the System of Artificial Intelligence], ed. by V.A. Lektorsky. Saint-Petersburg: Yuridicheskij tsentr, 2022, pp. 128–159. (In Russian)

Dubrovskij, D.I. Epistemologicheskij analiz sociogumanitarnoj znachimosti novatsij iskusstvennogo intellekta v kontekste obshchego iskusstvennogo intellekta [Epistemological Analysis of the Socio-humanitarian Significance of Artificial Intelligence Innovations in the Context of General Artificial Intelligence], *Filosofskie nauki* [Russian Journal of Philosophical Sciences], 2022, Vol. 65, No. 1, pp. 10–26. (In Russian)

Dubrovskij, D.I., Efimov, A.R., Matveev, F.M. Chto meshaet nam sozdat' Obshchij iskusstvennyj intellekt? Odna staraya stena i odin staryj spor [What Prevents Us from Creating a Common Artificial Intelligence? One Old Wall and One Old Dispute], *Voprosy filosofii* [The Problems of Philosophy], 2023, No. 5, pp. 39–49. (In Russian)

Karelov, S.V. "Lovushka Gudkharta" dlya AGI: problema sravnitel'nogo analiza iskusstvennogo intellekta i intellekta cheloveka [The "Goodhart Trap" for AGI: the Problem of Comparative Analysis of Artificial Intelligence And Human Intelligence], *Uchenye zapiski Instituta psikhologii Rossijskoj akademii nauk* [Scientific Letters of the Institute of psychology of Russian Academy of Sciences], 2023, Vol. 3, No. 3, pp. 5–22. (In Russian)

Le Doux, J., Birch, J., Andrews, K., Clayton, N.S. Consciousness Beyond the Human Case, *Current Biology*, 2023, No. 33 (16), pp. 832–840.

LeCun, Y. A Path Towards Autonomous Machine Intelligence, *OpenReview*, 2022. URL: https://openreview.net/pdf?id=BZ5a1r-kVsf (25.04.2024).

Legg, S., Hutter, M. Universal Intelligence: A Definition of Machine Intelligence, *Minds and machines*, 2007, No. 17 (4), pp. 391–444.

Luo, J., Longfei, M., Chang, Z., Fei, W., Hongxia, Y. Universal embodied intelligence: learning from crowd, recognizing the world, and reinforced with experience, *OpenReview*, 2023. URL: https://openreview.net/pdf?id=3e5nHhhRK93 (25.04.2024).

Sil'nyj iskusstvennyj intellekt. Na podstupakh k sverkhrazumu [Strong Artificial Intelligence. On the Outskirts of the Superintelligence]. Moscow: Intellektual'naya Literatura, 2021. 236 c. (In Russian)