УДК: 1 (091)
DOI: 10.15372/PS20250412
EDN: YXULMR

К.А. Родин

ВИТГЕНШТЕЙН И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Статья содержит краткий обзор первого тома сборника статей «Wittgenstein and Artificial Intelligence». Освещается проблема метонимической ловушки. Кратко рассматривается критика возможности сильного ИИ со стороны Н. Хомского и Р. Брэндома. Обсуждается природа «черного ящика» нейросетевых моделей. Отмечается значение радикального антирепрезентализма позднего Витгенштейна в контексте философских проблем ИИ.

Ключевые слова: Витгенштейн; искусственный интеллект; философия сознания; языковые игры; метонимическая ловушка; сильный ИИ; черный ящик; значение как употребление

K.A. Rodin

WITTGENSTEIN AND ARTIFICIAL INTELLIGENCE

This review analyzes key themes from the first volume of “Wittgenstein and Artificial Intelligence.” It addresses the metonymic trap in AI discourse, contrasts Chomsky’s and Brandom’s objections to strong AI, examines neural network opacity, and demonstrates how Wittgenstein’s anti-representationalism offers unique solutions to AI’s philosophical dilemmas.

Keywords: Wittgenstein; Artificial Intelligence; philosophy of mind; language games; metonymic trap; strong AI; black box; meaning as use

В 2024 г. вышел в двух томах сборник статей «Wittgenstein and Artificial Intelligence» под редакцией Брайана Болла, Элис К. Хэллиуэлл и Алессандро Росси [7; 8]. Болл и Росси были организаторами XI конференции Британского общества Витгенштейна (British Wittgenstein Society) «Витгенштейн и ИИ», состоявшейся в 2022 г. Сборник представляет во многом результаты конференции. Второй том посвящен философско-практическим темам (и имеет подзаголовок «Values and Governance»). Первый том концентрируется вокруг проблем сознания и языка. В настоящем кратком обзоре мы ограничимся только первым томом.

В предшествующие годы работы по этой теме были немногочисленными. Единственная посвященная ей монография вышла в 1998 г. (Stuart Shanker. «Wittgenstein's Remarks on the Foundations of AI» [6]). Заслуживает отдельного внимания статья 1983 г. «Wittgenstein on Language and Artificial Intelligence: The Chinese-room Thought Experiment Revisited» за авторством Клауса Обермайера [5]. Иных значимых публикаций не было. Недавние успехи в сфере ИИ (благодаря использованию глубоких нейронных сетей) и появление генеративных систем ИИ поставили новые или модифицировали старые философские проблемы в области машинного мышления. Стало интересным новое обращение к отдельным текстам и их перепрочтение. Витгенштейна обойти стороной не смогли.

Витгенштейн специально не писал и не мог писать на тему ИИ (не существовало такой отдельной области исследований). Однако нам хорошо известны заметки «Голубой книги» относительно способности машины мыслить. Известны также дискуссии Витгенштейна и Алана Тьюринга. В 1939 г. Тьюринг посещал кембриджские лекции Витгенштейна по основаниям математики и был единственным серьезным оппонентом (споры Тьюринга и Витгенштейна законспектированы и включены в издание лекций Витгенштейна). Относительно характера взаимного влияния Витгенштейна и Тьюринга продолжаются споры. Но даже и помимо историко-философских проблем авторы уверены в продуктивности обращения к Витгенштейну, надеясь в перспективе пролить свет на отдельные вопросы в теории и практике применения ИИ.

Первые три главы (Диана Праудфут, Томи Кокконен и Илмари Хирвонен, Артуро Васкес) предлагают критическое переосмысление устоявшихся интерпретаций Витгенштейна в связи с проблемой мышления машин. В первой главе опровергается бихевиористский миф относительно позиций Витгенштейна и Тьюринга. Обосновывается возможность принципиального согласия Витгенштейна и Тьюринга по вопросам атрибуции мышления машинам и значимости аффектов и эмоций. По мнению Кокконена и Хирвонена, энактивизм («воплощенное» взаимодействие со средой в сочетании со способностью к нормативному поведению) должен служить переходным звеном между скептицизмом Витгенштейна и оптимизмом Тьюринга. Васкес предлагает уйти от дихотомии сильного и слабого ИИ через понятие вторичного смысла. Однако больший интерес (не сугубо историко-философский или концептуальный) представляют следующие статьи.

Четвертая глава посвящена проблеме метонимической ловушки. При приписывании машинам интеллектуальных действий мы впадаем в метонимическую ловушку. Пример метонимического переноса (когда способности или свойства пользователя/оператора переносятся на инструмент) – выражение «калькулятор вычисляет». Калькулятор не вычисляет и не обладает способностью вычислять. Интеллектуальные действия нормативны по природе и осуществляются только внутри разделяемых людьми институтов и практик (калькулятор не разделяет институты и практики). Физические же действия, напротив, понимаются без переноса. «Самолет летает» подразумевает физическую способность летать. Автор главы различает способность летать самолета и способность летать птицы (различает соответствующие высказывания). Птица спонтанно и самопроизвольно актуализирует собственную способность к полету и может не полететь. Самолет летит как предназначенный и используемый инструмент (автор различает односторонние и двусторонние способности: самолет при соответствующих условиях не может не полететь и поэтому в сравнении с птицей обладает односторонней способностью). Однако различие между односторонней способностью самолета и двусторонней способностью птицы не мешает понимать выражение «самолет летает» в физическом и неметонимическом смыс-

лах. Машина может быть функционально эквивалентна человеку лишь при выполнении физических действий (посудомоечная машина при мытье посуды). Восприятие машины как субъекта интеллектуального действия – логическая ошибка (неразличение инструмента и агента действия). «Приписывание когнитивных… действий машинам по необходимости метонимично» [2, p. 99]. Аргументы автора релевантны § 103 «Философских исследований» Витгенштейна [9].

Пятая глава (Лейт Абдель-Рахман) [1] представляет отдельный интерес ввиду развернутой автором критики некоторых философских недоумений по поводу (возможности существования) сильного ИИ. Под атаку попадают позиция Хомского (критика ИИ исходя из интернализма) и позиция Брэндома (критика исходя из экстернализма). Современные большие языковые модели (LLM) основаны на архитектуре Transformer и работают, очевидно, из расчета статистических закономерностей. Но мы оцениваем уровень интеллектуальности LLM (в рамках бихевиористского критерия) сугубо по успешности или неуспешности – в контексте внешнего поведения – имитации человеческой языковой практики. Отсутствуют внутренние и внешние-не-бихевиористские критерии. Хомский, исходя из исследовательской программы универсальной грамматики расценивает язык как врожденную, биологическую способность. Мы создаем бесконечные синтаксические структуры из конечного набора биологически предзаданных элементов: язык одновременно и минималистичен, и эффективен. Мы не прибегаем к вероятностным вычислениям (и, соответственно, к огромному массиву данных). Поэтому путь статистического ИИ не приведет нас к сильному ИИ: статический ИИ не схватывает креативность нашего мышления. Аналитический прагматизм Брэндома объясняет возникновение мышления и значения (семантики) на основе вовлеченности в автономные языковые практики (Autonomous Discursive Practices, ADP). Статистический ИИ лишен возможности участвовать в ADP и не может быть сознательным и понимающим. Равно и интерналистская позиция (и соответствующая критика ИИ) Хомского, и экстерналистская позиция (и соответствующая критика ИИ) Брэндома представляются автору несостоятельными. Витгенштейн будто бы предлагает выход из общей

для Хомского и Брэндома антропологической ловушки: для сильного ИИ не обязательно ни по внутренней биологической структуре, ни по институциональному поведению быть похожим на человека. По Витгенштейну (и согласно автору), приписывание машине ментальных состояний осуществляется на уровне нормативно-функционального описания – не на основе какого-либо научного описания внутреннего мира машины. Поэтому критика Хомского и Брэндома – лишь антропоцентрические проекции. Нам следует допустить возможность нечеловеческих форм разума.

В шестой главе (Ян Граунд) [4] проводится различие между символьным ИИ и современным нейросетевым ИИ. Состояния символьного ИИ можно отобразить на понятия предметной области (символьный ИИ предметно-однороден). Состояния нейросетевого ИИ (автор говорит про сверточные нейронные сети) не репрезентируют объекты реального мира (ИИ предметно-гетерогенен), он становится «машиной Юма»: обнаруживает корреляции без необходимости причинного объяснения. Непрозрачность (черный ящик) сверточных нейронных сетей должна приниматься как конститутивная особенность (не как недостаток). Работа через нахождение статистических корреляций вне репрезентации и логического вывода подрывает нашу платоническую модель познающего разума. Витгенштейн был радикально против репрезентализма и иллюзии автономного (картезианского) субъекта. Мы часто принимаем решение на уровне фронесиса (практической мудрости Аристотеля) и не можем артикулировать причины принятых решений. Историчность, контекстуальность и неполная прозрачность равно присущи естественным познавательным агентам и ИИ.

Витгенштейн считал теоретизирование над природой языка очень плохой философией. Призыву Витгенштейна смотреть (вместо производства философских концепций) Джованни Галли в седьмой главе сопоставляет «отсутствие теоретической нагруженности в технологиях глубокого обучения». Витгенштейн создает общую почву для разработчиков систем обработки естественного языка (NLP) [3, p. 146]. Коннекционизм современных нейросетевых моделей ИИ якобы можно проследить до текстов позднего Витгенштейна.

В известном мысленном эксперименте против существования индивидуального (приватного) языка «жук в коробке» (см. «Философские исследования», § 243–256) Витгенштейн не допускает возможности установления значения слов через внутреннее (индивидуальное) указание на объекты приватного опыта. Значение определяется публичным использованием слов внутри языковых практик. Согласно Галли, система Word2Vec реализует принцип Витгенштейна «значение как употребление». Слова преобразуются алгоритмом в векторы (последовательности чисел). Слова из похожих контекстов будут помещены рядом и в векторном пространстве. Поэтому слова-векторы лишены какой-либо сущности и не отсылают к какому-либо референту. Значение определяется только контекстом (пусть контекст остается лингвистическим и не учитывает жизненный мир). Векторно-символьные архитектуры (VSA) разрывают связь между словом и объектом. Пусть даже какой-то символ (какое-то слово) будет представлен случайным вектором (и для каждого пользователя вектор будет собственным – по аналогии с жуком в коробке), агенты очевидно смогут оперировать случайными векторами из-за общих разделяемых правил и, следовательно, иметь успешную коммуникацию. Галли обсуждает аналогию между приватным опытом и непрозрачностью (черным ящиком) работы некоторых систем ИИ. Непрозрачность ИИ – лишь техническая недоступность. Техническую недоступность не следует путать с приватным опытом. Сложные (до известной степени не поддающиеся анализу) математические преобразования не могут создавать субъективный опыт. ИИ не могут обладать изолированным от публичных языковых практик (не-приватных по определению) приватным субъективным опытом. Мы описываем наш субъективный (якобы приватный) опыт при помощи известных и всегда публичных языковых игр. Обученную на контексте наших языковых игр машину не следует подозревать в наличии какой-либо приватной субъективности.

Восьмая глава поднимает проблему релевантного упрощения информации. При любом упрощении определение неустранимых (сохраняющихся при упрощении) элементов не может не исходить из соответствующих контекста и целей (даются очевидные ссылки на позд-

него Витгенштейна). Поэтому по-настоящему интеллектуальный ИИ должен иметь настоящие цели и быть способным быть приверженным чему-то. Девятая глава посвящена аналогичному рассуждению. Обсуждается возможность построения универсальной модели. Модель должна однозначно определять условия корректности аналогичного умозаключения. Рассмотрение отдельных философских моделей аналогичного умозаключения на фоне ИИ и критических замечаний Витгенштейна завершается оптимистической уверенностью в некоторой перспективности движения в сторону универсализации.

Литература

1. *Abdel-Rahman L*. The forms of artificially intelligent life: Brandom, Chomsky and Wittgenstein on the possibility of strong-AI // Wittgenstein and Artificial Intelligence. Vol. I: Mind and Language. Anthem Press, 2024. P. 105–122.

2. *Boisseau É*. The metonymical trap // Wittgenstein and Artificial Intelligence. Vol. I: Mind and Language. Anthem Press, 2024. P. 85–104.

3. *Galli G*. Language models and the private language argument: A Wittgensteinian guide to machine learning // Wittgenstein and Artificial Intelligence. Vol. I: Mind and Language. Anthem Press, 2024. P. 145–164.

4. *Ground I*. Black boxes, beetles and beasts // Wittgenstein and Artificial Intelligence. Vol. I: Mind and Language. Anthem Press, 2024. P. 123–144.

5. *Obermeier K.K*. Wittgenstein on language and artificial intelligence: The Chinese-room thought experiment revisited // Synthese, 1983. Vol. 56, No. 3. P. 339–349.

6. *Shanker S*. Wittgenstein's Remarks on the Foundations of AI. Routledge, 1998.

7. *Wittgenstein* and Artificial Intelligence. Vol. I: Mind and Language / Ed. by B. Ball, A.C. Helliwell and A. Rossi. Anthem Press, 2024.

8. *Wittgenstein* and Artificial Intelligence. Vol. II: Values and Governance / Ed. by B. Ball, A.C. Helliwell and A. Rossi. Anthem Press, 2024.

9. *Wittgenstein L*. Philosophical Investigations. Wiley-Blackwell, 1953.

References

1. *Abdel-Rahman, L*. (2024). The forms of artificially intelligent life: Brandom, Chomsky and Wittgenstein on the possibility of strong-AI. In: Wittgenstein and Artificial Intelligence. Vol. I: Mind and Language. Anthem Press, 105–122.

2. *Boisseau, É*. (2024). The metonymical trap In: Wittgenstein and Artificial Intelligence. Vol. I: Mind and Language. Anthem Press, 85–104.

3. *Galli, G*. (2024). Language models and the private language argument: A Wittgensteinian guide to machine learning. In: Wittgenstein and Artificial Intelligence. Vol. I: Mind and Language. Anthem Press, 145–164.

4. *Ground, I*. (2024). Black boxes, beetles and beasts. In: Wittgenstein and Artificial Intelligence. Vol. I: Mind and Language. Anthem Press, 123–144.

5. *Obermeier, K.K*. (1983). Wittgenstein on language and artificial intelligence: The Chinese-room thought experiment revisited. Synthese, Vol. 56. No. 3, 339–349.

6. *Shanker, S*. (1998). Wittgenstein's Remarks on the Foundations of AI. Routledge.

7. *Ball, B., A.C. Helliwell & A. Rossi* (Eds.). (2024). Wittgenstein and Artificial Intelligence. Vol. I: Mind and Language. Anthem Press. 2024.

8. *Ball, B., A.C. Helliwell & A. Rossi* (Eds.). (2024). Wittgenstein and Artificial Intelligence. Vol. II: Values and Governance. Anthem Press.

9. *Wittgenstein, L*. (1953). Philosophical Investigations. Wiley-Blackwell.

Информация об авторе

Родин Кирилл Александрович – Институт философии и права СО РАН (630090, Новосибирск, ул. Николаева, 8).

rodin.kir@gmail.com

Information about the author

Rodin, Kirill Aleksandrovich – Institute of Philosophy and Law, Siberian Branch of the Russian Academy of Sciences (8, Nikolaev St., Novosibirsk, 630090, Russia).

rodin.kir@gmail.com

Дата поступления 16.09.2025.
Принята к публикации 10.11.2025.