Эпистемология и философия науки
2026. Т. 63. № 3. С. 94–112
УДК 167.7

Epistemology & Philosophy of Science
2026, vol. 63, no. 3, pp. 94–112
DOI: https://doi.org/10.5840/eps202663343

Де-антропоцентризм как концептуальная альтернатива антропоморфизации в эпистемологии искусственного интеллекта*

Фролов Константин Геннадьевич – кандидат философских наук, исследователь. Межрегиональная общественная организация «Русское общество истории и философии науки». Российская Федерация, 105062, г. Москва, Лялин пер., д. 1/36, стр. 2; e-mail: konstantin-frolov@yandex.ru



Статья посвящена вопросу о том, могут ли эпистемические понятия – такие как убеждение, знание, обоснование, владение понятиями, сомнение и другие – быть обоснованно приписаны моделям искусственного интеллекта. Рассматривается широкий спектр позиций по данному вопросу – от скептических до радикально оптимистических. При этом отстаиваемая в статье позиция относится скорее к оптимистической части спектра. Ключевой выдвигаемый тезис заключается в том, что уместность приписывания эпистемических понятий моделям искусственного интеллекта определяется не внутренними свойствами системы, но ее способностью функционировать в качестве нормативного эпистемического агента, а также ее способностью корректно взаимодействовать с другими эпистемическими агентами. Такая позиция среди прочего ориентирована на то, чтобы отклонить выдвигаемые в адрес оптимистических подходов возражения о неоправданной антропоморфизации представлений об устройстве и функционировании систем искусственного интеллекта. Эта цель достигается путем отстаивания тезиса де-антропоцентризма. Данный тезис гласит, что ничто ни в природе человека, ни в природе эпистемических понятий не дает нам оснований полагать, что лишь человек в силу его особой природы способен к «подлинному» знанию, пониманию, рассуждению и владению понятиями. При этом обоснование тезиса де-антропоцентризма не требует ни пересмотра человеческой природы, ни трансформации взглядов о природе и статусе существующих технологий искусственного интеллекта, но требует лишь принятия определенной позиции в отношении семантики эпистемических понятий. Тем самым обоснование тезиса де-антропоцентризма относится преимущественно к сфере эпистемологии и философии языка, а не к области когнитивных наук и не к области специальных исследований искусственного интеллекта.

Ключевые слова: эпистемические понятия, искусственный интеллект, знание, семантика, нормативность, агентность, антропоморфизация

* Исследование подготовлено при поддержке Российского научного фонда, проект № 26-18-00851 «Цифровые трансформации семантики понятия науки в социально-сетевую эпоху: лингво-аналитические, теоретико-эволюционные и социально-коммуникативные исследования».

© Фролов К.Г., 2026

DE-ANTHROPOCENTRISM AS A CONCEPTUAL ALTERNATIVE TO ANTHROPOMORPHIZATION IN THE EPISTEMOLOGY OF ARTIFICIAL INTELLIGENCE

Konstantin G. Frolov –
PhD in Philosophy, researcher.
Inter-Regional
Non-Government Organization
"Russian Society of History
and Philosophy of Science".
1/36 Lyalin Lane, bd. 2,
Moscow 105062,
Russian Federation;
e-mail:konstantin-frolov@
yandex.ru

The paper explores whether epistemic concepts – such as belief, knowledge, justification, possession of concepts, and doubt – can be justifiably ascribed to AI models. Various positions on this issue are considered, ranging from skepticism to radical optimism. The position defended in this paper, however, leans more towards the optimistic side of the spectrum. The main claim is that justified ascription of epistemic concepts to AI models is determined not by the system's intrinsic properties, but by its ability to function as a normative epistemic agent as well as its capability to interact properly with other epistemic agents. This optimistic view aims to counter criticisms of unjustified anthropomorphism in our understanding of the design and functioning of AI systems. This is accomplished by advocating for the thesis of de-anthropocentrism. This thesis argues that there is nothing in human nature or in the nature of epistemic concepts that justifies the belief that only humans, due to their unique nature, are capable of possessing 'real' knowledge, understanding, reasoning, and possession of concepts. Remarkably, justifying the de-anthropocentrism thesis does not require a reconsideration of human nature or a transformation of views on AI. Instead, it merely requires a certain stance regarding the semantics of epistemic concepts. Thus, the justification for the de-anthropocentrism thesis primarily pertains to the realm of epistemology and the philosophy of language, rather than the study of cognitive science, or specialized research into artificial intelligence.

Keywords: epistemic concepts, artificial intelligence, knowledge, semantics, normativity, agency, anthropomorphization

Введение

Современный этап развития технологий искусственного интеллекта характеризуется той особенностью, что различные исследователи, обладающие сопоставимой квалификацией и имеющие доступ к одному и тому же набору наблюдаемых фактов и свидетельств, релевантных для описания поведения такого рода моделей, склонны интерпретировать эти свидетельства принципиально различным образом. В результате в исследовательской среде сформировался широкий спектр мнений по поводу возможности и целесообразности приписывания таким моделям тех или иных состояний и способностей, ранее приписывавшихся лишь человеку [Simbolon et al., 2025].

Список этих категорий довольно широк: убеждения, знание, понимание, обоснование, умозаключение, распознавание, владение понятиями, обладание свидетельствами, выдвижение гипотез, поиск решений, планирование, обман, сомнение, уверенность и другие.

Заметим, что все эти понятия изначально принадлежали так называемому словарю «народной эпистемологии» [Hardy-Vallee, Dubreuil, 2010] и в этой связи имеют две особенности.

Первая особенность заключается в том, что все эти категории в той или иной мере сформировались в языке еще до того этапа, когда человек значительно продвинулся в теоретическом постижении областей теоретической эпистемологии, психологии и когнитивных наук. Слова, выражающие эти понятия, изначально не были терминами каких-либо теорий, но приобрели статус терминов значительно позже формирования практики их повседневного употребления.

Такая особенность происхождения влечет умножение смыслов. С одной стороны, смысл этих слов в первом приближении понятен любому носителю языка, и, более того, с лингвистической точки зрения именно носитель языка является последней инстанцией, к кому следует обращаться за их толкованием. Эти слова как термины повседневного словаря выражают то, что имеют в виду люди – широкий класс людей, носителей языка, – когда они используют их в своей речи. С другой стороны, эти слова, будучи терминами различных теорий, приобретают в рамках этих теорий различный смысл и условия их корректного употребления. Причем, что важно, речь идет именно о *различных* теориях и, как следствие, различных смыслах этих терминов. Тем самым конкуренция между теориями за статус наиболее адекватного способа описания и объяснения некоторого фрагмента действительности сопровождается конкуренцией между теориями за наиболее оправданное и обоснованное толкование соответствующих терминов и выражаемых ими понятий. При этом никакого универсального способа сопоставления различных теорий по степени их признанности и обоснованности не существует. Если две теории расходятся в своем толковании тех или иных терминов, то помимо такого расхождения между теориями по поводу трактовки терминов также возможно расхождение во мнениях среди сторонних квалифицированных наблюдателей по поводу того, какая из этих двух теорий лучше обоснована и является более предпочтительной.

В итоге мы имеем довольно парадоксальную картину семантики подобных терминов. С одной стороны, их смысл понятен любому носителю языка в той мере, в какой он не испытывает затруднений с их повседневным использованием [Вострикова, 2025, с. 227]. В самом деле, как бы он это делал, если бы не понимал их смысла? С другой стороны, по поводу их смысла и условий корректного применения идут оживленные дискуссии между исследователями,

которые далеки от достижения какого-либо консенсуса и преодоления разногласий между собой[1].

Вторая особенность понятий из приведенного выше списка заключается в том, что все они так или иначе встроены в определенную нормативную семантическую рамку. Их смысл раскрывается в контексте категорий рациональности, уместности, правильности, целесообразности и ряда иных нормативных категорий [Peacocke, 1999, p. 18]. Так, например, еще У. Селларс в своей работе «Эмпиризм и философия сознания» утверждал, что «существенный момент заключается в том, что, характеризуя некоторое состояние как состояние познания, мы не даем эмпирического описания этого состояния; мы помещаем его в логическое пространство доводов, обоснования и возможности обосновать сказанное» [Sellars, 1997, p. 76]. То есть помещаем в нормативную семантическую рамку, которая характеризует уместные коммуникативные реакции, но ничего не говорит о специфике внутренних состояний агентов.

При этом ясно, что хронологически человек усвоил эпистемические понятия задолго до того, как в его мышлении сформировалась и выделилась самостоятельная философская категория рациональности. И все же эта категория была изначально имплицитно встроена, «вшита» в смысл этих понятий народной эпистемологии, призванных объяснять наблюдаемое поведение людей и оценивать его на предмет правильности. То есть оценивать на предмет того, насколько в том или ином случае уместно или даже необходимо внести изменение в характер наблюдаемых действий кого-то из окружающих путем внесения изменений в его ненаблюдаемые эпистемические установки, которые лежат в основании всякого разумного и целенаправленного поведения.

Современный этап развития коммуникативных практик характеризуется трансформацией и расширением класса коммуникативных взаимодействий за счет включения в число их участников множества разнообразных искусственных систем. Соответственно, целью данной статьи является исследование того, в какой мере понятия народной эпистемологии могут быть распространены на модели искусственного интеллекта.

При этом искать ответ на этот вопрос можно двумя различными способами. Можно руководствоваться имеющимся у каждого

[1] Так, например, среди различных трактовок понятия желания в аналитической философии можно встретить как подходы, исходящие из того, что желания представляют собой подкласс убеждений (а именно, убеждений, что некое положение дел *p* является благом [Price, 1989]), так и подходы, согласно которым желания и убеждения относятся к двум различным, непересекающимся классам установок [Anscombe, 2000], первый из которых называется классом когнитивных установок, а второй – классом некогнитивных установок.

носителя языка интуитивным схватыванием эпистемических понятий и основной задачей исследования избрать вопрос о том, реализуют ли те или иные искусственные системы с учетом специфики их архитектуры, способа обучения и т.д. такие состояния, которые можно было бы охарактеризовать посредством данных эпистемических понятий. Например, можно ли сказать про шахматную нейросеть AlphaZero, что она помнит все сыгранные ею партии? Можно ли сказать, что управляемый специализированной нейросетью автономный автомобиль видит пакет на дороге? Для ответа на эти вопросы обращение к принципам устройства и функционирования той или иной модели может быть необходимым в том случае, если мы ясно и отчетливо понимаем, что значит «помнит» и «видит», и при этом нас интересует вопрос о том, обладают ли соответствующими способностями и состояниями данные конкретные системы.

Однако к достижению обозначенной выше цели можно подойти и несколько иным путем. В фокусе исследовательского внимания может быть не устройство той или иной технической системы, но вопросы о том, что значит «помнит» и что значит «видит». Эти вопросы не имеют непосредственного отношения к специфике философии искусственного интеллекта как таковой, но относятся к ведению двух других областей – эпистемологии и семантики[2]. При этом принятие той или иной точки зрения по поводу этих вопросов может значительно упростить исследование множества разнообразных частных случаев. Более того, мы можем сделать еще один шаг на пути абстрагирования и отвлечься от специфики не только тех или иных технических систем, но и от специфики отдельных эпистемических понятий, таких как «помнит» и «видит», и занять определенную теоретическую позицию в отношении семантики всех эпистемических понятий как таковых. Это даст нам наиболее обобщенный критерий приписываемости эпистемических понятий *любым* искусственным системам.

Этот критерий таков: эпистемическое понятие С может быть обоснованно приписано искусственной системе М в том случае, если эта система может быть обоснованно включена в систему нормативных эпистемических отношений, задаваемых понятием С.

Структура и характер этих нормативных отношений при этом определяются содержанием и спецификой того или иного конкретного эпистемического понятия. При этом в дальнейшем мы раскроем, при выполнении каких условий, на наш взгляд, искусственные системы оказываются способны удовлетворять данному обобщенному

[2] Также эти вопросы некоторым образом относятся к области философии сознания, однако этот ракурс проблематики мы в данном исследовании вынужденно обойдем вниманием.



критерию. Однако прежде обратимся к иллюстративному обзору позиций, встречающихся в литературе, по рассматриваемой нами проблематике.

Разнообразие позиций

Как уже было отмечено выше, спектр мнений по вопросу уместности приписывания системам искусственного интеллекта таких состояний, как убеждения, знания, понимание, владение понятиями и прочих, довольно широк [Simbolon et al., 2025]. Наш обзор этих мнений не претендует на полноту, но ориентирован лишь на то, чтобы показать, что различия в наблюдаемых позициях среди исследователей действительно существенны, а встречающиеся в литературе оценки порой диаметрально противоположны.

На одном конце спектра, который мы будем называть скептическим, можно найти множество исследователей[3], среди которых мы в качестве примера рассмотрим позицию Стефана Буйсмана, полагающего, что «мы не можем приписывать системам искусственного интеллекта владение понятиями» [Buijsman, 2025, p. 3]. Он обосновывает это тем, что такое владение должно предполагать возможность оперативного учета экспертного мнения: «Когда мы говорим об обращении к экспертам в случае понятий, мы имеем в виду механизм, в рамках которого применение понятий быстро корректируется, когда эксперт указывает на допущенную концептуальную ошибку. И хотя мы можем легко исправить такие ошибки на уровне выходных данных модели, нет аналогичного способа напрямую скорректировать соотношения весов, используемые системой. Мы не можем, например, изменить нейронную сеть так, чтобы карты важности признаков более точно соответствовали тому, что мы считаем релевантными признаками» [Ibid., p. 11]. Тем самым система, откликаясь на экспертное мнение, ориентирована в первую очередь на то, чтобы корректировать непосредственно сам ответ, а не распределение весов признаков. Со временем эти веса постепенно корректируются, но не так оперативно, как это происходит в случае реакции человека.

Это довольно серьезное возражение. Модели такого рода тем самым не просто «не понимают», но и *не демонстрируют поведения*,

[3] Наиболее типичное изложение скептической позиции можно найти в статье Мюррэя Шанахана, который утверждает, что «чат-боты на основе LLM, строго говоря, не обладают убеждениями. Чат-бот не обладает *подлинным* знанием (does not *really* know), что Бурунди находится к югу от Руанды» [Shanahan, 2024, p. 72].

характерного для понимания человеком. Однако на это можно ответить в духе описываемого ниже подхода, что отождествление всякого понимания с поведением, характерным именно для человеческого понимания, является проявлением антропоцентризма. Искусственные модели демонстрируют в таких случаях несколько иное поведение, однако это может лишь означать, что они *понимают иначе* и иначе овладевают понятиями. Они понимают медленнее. Однако какая-либо определенная скорость осуществления ревизий не является сущностной характеристикой понимания как такового. Понимание может протекать с различной скоростью.

При этом на противоположном, оптимистическом, краю рассматриваемого спектра мнений можно обнаружить таких исследователей, как Герман Каппелен и Джош Девер[4]. Они выдвигают следующий тезис, название которого на русский язык можно перевести как «во все тяжкие» (Whole Hog Thesis): «ChatGPT – это полноценный лингвистический и когнитивный агент наравне с человеком. Он способен осмысленно использовать язык, делать утверждения, задавать и отвечать на вопросы, предлагать гипотезы и отдавать команды. <…> Он может познавать мир (получая информацию из рассказов или рассуждая на основе уже имеющихся знаний). Он может строить планы и рассуждать о том, как их достичь. Он может предпринимать действия для реализации этих планов» [Cappelen, Dever, 2025a, p. 13].

В основе этого чрезвычайно амбициозного исследовательского проекта лежит так называемая гиперэкстерналистская[5] установка, согласно которой вопрос о том, является ли некоторое состояние, например, убеждением, определяется не внутренними факторами и характеристиками того объекта, которому это состояние приписывается, но характером его взаимодействия с миром [Shin, 2025, p. 1555] и, что наиболее важно для нашего изложения, «практикой предоставления и запроса оснований, в которую он встроен» [Cappelen, Dever, 2025b, p. 11].

По мнению Каппелена и Девера, в действительности проблема не в том, что мы склонны к антропоморфизации при описании поведения искусственных систем, но в том, что наши теории и фигурирующие в них термины слишком антропоцентричны [Cappelen, Dever, 2021, p. 69].

[4] К этой же части спектра можно, например, отнести Уинли Ма и Винсента Валтона [Ma, Valton, 2024, p. 7].

[5] Приставка «гипер-» в этом термине указывает на то, что речь идет о крайней, последовательно экстерналистской позиции в рамках дискуссий в таких областях, как эпистемология и семантика.

Антропоцентризм на примере доверия

Антропоцентризм являет собой представление о том, что определенные черты и аспекты тех или иных способов взаимодействия между людьми являются сущностными чертами этих способов взаимодействия, не допускающими их распространения на более широкий круг участников, включающий искусственные системы.

В качестве весьма характерной иллюстрации рассмотрим вопрос о возможности приписывания моделям искусственного интеллекта благонадежности. Как и по другим аналогичным вопросам, здесь среди исследователей наблюдается широкий спектр мнений[6]. На скептическом краю этого спектра можно встретить такого исследователя, как Кристиан Будник, который утверждает, что «вопрос о том, когда мы можем доверять цифровым технологиям, представляет собой своего рода категориальную ошибку: доверять цифровым технологиям вообще невозможно, поскольку цифровые артефакты, в первую очередь, не относятся к числу подходящих объектов доверия» [Budnik, 2025, p. 2]. Он обосновывает свое мнение тем, что доверие предполагает наличие нормативных ожиданий, что в свою очередь требует допущения о том, что сторона, на которую распространяются эти нормативные ожидания, способна к намеренным и ответственным действиям [Ibid., p. 7]. Обосновать же наличие у моделей искусственного интеллекта способности к намеренным и ответственным действиям, по его мнению, было бы крайне трудной задачей.

На противоположном краю спектра мнений по данному вопросу располагаются такие исследователи, как Мона Симион и Кристоф Келп. Они признают, что в литературе распространена точка зрения, согласно которой благонадежность представляет собой добродетель в морально релевантном смысле: «В то время как надежный (reliable) человек просто добросовестно выполняет свои обязательства, заслуживающий доверия (trustworthy) человек выполняет свои обязательства добросовестно в силу своей доброй воли по отношению к доверителю» [Simion, Kelp, 2023, p. 3]. Однако, как отмечают эти же авторы, трактовка подобного рода является чрезмерно антропоцентричной: «Правильно ли считать, что способность ИИ к тому, чтобы заслуживать доверия, зависит от его способности проявлять волю или демонстрировать добродетели характера? И в более общем смысле, следует ли нам, обнаружив, что конкретный артефакт,

[6] Стоит оговорить, что речь здесь идет не о конкретных существующих на сегодняшний день моделях LLM, которые в силу принципов своей архитектуры систематически склонны галлюцинировать, но о более отвлеченном вопросе критериев приписываемости благонадежности искусственным интеллектуальным системам как таковым.

который мы считали заслуживающим доверия, на самом деле не способен проявлять эти психологические черты, пересмотреть наше отношение к этому артефакту?» [Simion, Kelp, 2023, p. 3]. В качестве ответа на данный вопрос они приводят цитату из монографии Саймона Блэкберна: «Мы часто соглашаемся доверять, не зная многого о психологии того, кому доверяем, просто предполагая, что у него есть… способности, достаточные для выполнения соответствующей работы» [Blackburn, 1998]. В конечном итоге они утверждают, что заслуживающий доверия искусственный интеллект можно определить как «искусственный интеллект, который выполняет свои функционально обусловленные обязательства, причем последние могут быть либо заданы на этапе проектирования, либо обусловлены этиологически (etiologically determined) через преимущества, которые приносит соответствующий ИИ» [Simion, Kelp, 2023, p. 11].

Как видно, в рамках данного подхода способность заслуживать доверие объясняется через встраиваемость в структуру нормативных отношений определенного рода. В таком случае исследование наблюдаемых практик взаимодействия объектов с окружающей средой, а также их встраиваемости в нормативные отношения должно быть стартовой точкой в исследовании уместности приписывания искусственным объектам и системам эпистемических состояний [Alvarado, 2023].

Объекты и нормативные отношения

Рассмотрим, например, автономно управляющийся автомобиль. Способен ли такой автомобиль нарушать правила дорожного движения? С экстерналистской точки зрения это несложный эмпирический вопрос. Для поиска ответа на него достаточно в течение некоторого времени пронаблюдать за тем, что происходит с таким автомобилем на дорогах. Вполне представимо, что в определенный момент траектория его движения может оказать таковой, что ее вполне можно охарактеризовать как нарушение правил дорожного движения. Например, представимо, что автономно управляемый автомобиль выезжает на встречную полосу движения, пересекая при этом двойную сплошную разделительную черту.

Ясно, что столь непосредственная и прямая трактовка такого события может вызвать закономерные возражения, указывающие на то, что нарушением правил движения может быть лишь контролируемое и намеренное действие агента. Если же автомобиль выехал на встречную полосу вследствие механической поломки, то такое событие не будет нарушением правил просто потому, что оно не будет ничьим действием. Нарушением в данном случае может быть лишь



бездействие владельца автомобиля, вовремя не прошедшего положенный по регламенту технический осмотр своего транспортного средства.

Тем не менее мы можем дополнить наш сценарий, уточнив, что такая траектория движения была контролируемой и не была результатом механической поломки. Такое обстоятельство вполне можно установить по итогам проведенной экспертизы. В таком случае для оценки данного события нам, по-видимому, потребовалось бы оценить те основания, в силу которых автомобиль двигался по такой траектории. Например, исходя из отчетов такого автомобиля, нам может быть известно, что выбор в пользу указанной траектории был совершен на том основании, что любая иная физически реализуемая траектория привела бы к аварии с потенциально летальным исходом, вероятность которого была бы крайне высокой. В таком случае выезд на встречную полосу не является нарушением правил дорожного движения, поскольку правила допускают подобное в случае крайней необходимости.

Однако пусть мы теперь обнаруживаем, что основанием для выезда на встречную полосу было то, что автономно управляемый автомобиль принял черный полиэтиленовый пакет за лежащего на проезжей части человека и, как и в предыдущем случае, выбрал указанную траекторию во избежание возможных жертв. В этом случае никаких условий крайней необходимости в действительности не было и выбор такой траектории определенно является нарушением действующих правил.

Тем самым мы оказываемся перед непростым выбором терминов для описания происходящего. Мы можем либо согласиться с тем, что такой автомобиль может анализировать поступающие данные и свидетельства, способен оценивать их нормативный вес, а также в состоянии на основании этих свидетельств определять траекторию своего контролируемого движения и тем самым может либо следовать правилам дорожного движения, либо гипотетически способен нарушать их. Либо – и в этом заключается вторая доступная нам опция – мы можем утверждать, что использование подобных терминов для описания данной ситуации является проявлением антропоморфизации [Scorici et al., 2024] и тем самым является некорректным способом описания происходящего.

Основных проблем со второй опцией две.

Первая проблема заключается в том, что у нас может не быть другого, более подходящего языка описания и объяснения подобного рода событий. Объяснение, согласно которому автономный автомобиль выехал на встречную полосу потому, что принял лежащий на земле полиэтиленовый пакет за живого человека, может оказаться наиболее адекватным, информативным и релевантным по сравнению с любым иным способом описания и объяснения этих же событий.

Вторая проблема заключается в том, что если структура описания оснований, учитываемых системой искусственного интеллекта, встроенной в автомобиль, совпадает со структурой оснований, учитываемой человеком, который мог бы находиться на месте водителя, то у нас нет причин для отказа приписывать искусственному интеллекту те состояния, природа которых сводится к определенным уместным реакциям на соответствующие типы фактов.

Гиперэкстернализм

Стоит признать, что гиперэкстерналистский подход к толкованию терминов словаря «народной эпистемологии» исходит из того, что поведенческая имитация эпистемических состояний (при выполнении ряда условий, о которых речь пойдет ниже) тождественна обладанию ими. Это в буквальном смысле одно и то же. Нет никакого довеска, которым «подлинное» знание или понимание могли бы отличаться от демонстрируемой в поведении «имитации». Если вы, например, научились идеально имитировать способность сомневаться в своих ответах, включая правильное воспроизводство влияния такого имитируемого сомнения на характер и содержание ваших ответов, то вы приобрели способность сомневаться в собственных ответах. С гиперэкстерналистской точки зрения сомнение – это не то, что качественно переживается агентами где-то внутри глубин их сознания, но то, что проявляет себя в актах коммуникации между агентами и определяет уместность тех или иных реакций при их взаимодействии друг с другом. Если характер коммуникативного взаимодействия между агентами и его уместные последствия таковы, что всякий носитель языка с готовностью опишет их как проявление сомнения, то, значит, такое взаимодействие может быть обоснованно охарактеризовано таким образом. С гиперэкстерналистской точки зрения понятие сомнения определяется не характером переживаний вовлеченных агентов, но той структурой нормативных отношений, которая определяет, какие коммуникативные реакции являются уместными в ответ на то или иное коммуникативное действие. В свою очередь, противоположной позицией будет та, согласно которой в основе сомнения – а также знания, понимания, желаний и других подобных состояний – находится не их нормативная роль в рамках механизмов порождения поведения и уместных коммуникативных реакций, но качественная специфика обладания ими, переживаемая из перспективы первого лица. Такая позиция с точки зрения экстернализма приводит к неоправданному сужению класса объектов, которым могут быть обоснованно приписаны соответствующие состояния и способности, что и представляет собой антропоцентризм.



В качестве дополнительной иллюстрации, раскрывающей суть рассматриваемой экстерналистской позиции, мы можем рассмотреть вопрос о том, умеет ли калькулятор умножать числа. По этому поводу возможны по меньшей мере три точки зрения. Первая из них предполагает утвердительный ответ. Умножение чисел – это одна из основных функций калькулятора, с которой имеющиеся в мире калькуляторы систематически хорошо справляются [Humphreys, 2004, p. 5]. Таких наблюдаемых свидетельств вполне достаточно для того, чтобы дать утвердительный ответ на этот вопрос. Вторая возможная позиция – скептическая. Она предполагает, что приписывание калькуляторам способности умножать числа является некорректным и необоснованным, поскольку калькуляторы не обладают понятием числа, не имеют соответствующих репрезентаций [Бажанов, 2025, с. 141] и не обладают пониманием арифметических отношений. Вычислительные процессы, происходящие с калькуляторами, являются каузальными, физическими процессами. Приписывание им когнитивных способностей оправданно не более, чем приписывание таких способностей механическим счетам. Такое приписывание представляет собой необоснованную антропоморфизацию [Arsenault, 2026].

Однако по этому вопросу возможна также третья позиция: калькуляторам действительно неуместно приписывать наличие когнитивных способностей, но не потому, что они не владеют понятием числа и не обладают пониманием арифметических отношений, но потому, что они не способны к вариативному контролируемому поведению, не способны к оценке свидетельств и вследствие этого не являются агентами, чьи действия могут быть вписаны в структуру нормативных отношений. Калькуляторы не могут получать и оценивать свидетельства, не могут ошибаться в этих своих оценках иначе, чем в силу каузальных сбоев, и не могут оценивать вероятность собственной ошибки. В свою очередь, модели искусственного интеллекта такими способностями обладают [Cash et al., 2025, p. 397]. Их ошибки – это не результат сбоев и поломок, но результат неверной оценки доступных свидетельств. Рассматривая ответы моделей таким образом, мы тем самым предполагаем, что они могли бы быть точнее в своих ответах. Такая опция проявления немного иного, более обоснованного поведения является для них номологически возможной. Именно поэтому их коммуникативное поведение может быть вписано в структуру нормативных отношений, что позволяет приписывать им эпистемические понятия, содержание которых полностью описывается их ролью в этой структуре.

Вариативность поведения и отзывчивость к основаниям

Ясно, что такое расширение круга участников нормативных отношений следует проводить с осторожностью. При этом требуется уточнить и прояснить критерии такого включения, которые согласовывались бы со здравым смыслом. В противном случае мы можем обнаружить, что практика нашего словоупотребления, в частности, позволяет утверждать, что некоторые здания способны нарушать те или иные действующие нормы – например, нормы высотной застройки. В свою очередь, некоторые законы способны нарушать нормы международного права. Причем они способны их нарушать даже в том случае, если люди, написавшие и принявшие эти нормативные акты, действовали в установленном законом порядке и ничего при этом не нарушали. По-видимому, речь в данном случае идет о побочных эффектах языка как естественного и потому несовершенного феномена. Эти эффекты не следует трактовать как свидетельства в пользу того, что здания, фотографии или видеоизображения сами способны вступать в нормативные отношения и сами способны что-либо нарушать или чему-либо соответствовать.

Способность к нарушению, следованию и соответствию целесообразно сохранить исключительно за так называемыми нормативными агентами, т.е. агентами, способными к обладанию и проявлению собственных нормативных установок, а также способными к проявлению отзывчивости по отношению к нормативным основаниям.

Ключевой характеристикой агентности является способность к вариативности поведения и контролю за реализацией стратегии поведения, выбранной из множества доступных альтернатив [Cochrane, 2018, p. 185].

В свою очередь, вариативность проявлений активности предполагает, что события, случающиеся с объектом, имеют место не в рамках линейной каузальной цепочки, а в рамках ветвящихся каузальных траекторий. Спонтанные отклонения изначальных параметров в сочетании с высочайшей чувствительностью к этим параметрам, характерной для нелинейных процессов, приводят к тому, что, например, две идентичные клетки, находящиеся в одинаковых начальных условиях, могут двигаться по различным траекториям. Причем такое различие является не результатом ограниченности нашего знания, а следствием того, как функционируют физические системы определенного рода вдали от термодинамического равновесия.

Однако наличия такой вариативности и фундаментальной непредсказуемости в проявлении объектом той или иной активности еще недостаточно для полноценной агентности. Вариативность



проявлений активности означает только то, что объект способен двигаться по различным траекториям из одного и того же исходного состояния. Ключевой шаг происходит тогда, когда объект в силу своей природы и структурной организации находится не просто перед лицом ветвящихся траекторий, но способен влиять – а в пределе и полностью определять, – какая из доступных для него траекторий будет в конечном итоге реализована в мире. Достигается это за счет разнообразных механизмов обратной связи, устроенных таким образом, что одни малейшие отклонения усиливаются, тогда как другие случайные отклонения погашаются.

Ключевым обстоятельством здесь является то, что, хотя различные траектории действительно могут быть реализуемы из одного и того же исходного состояния объекта и его среды, ни одна из этих траекторий не выходит за рамки номологически возможного, т.е. за рамки совместимого с имеющимися законами природы.

Наконец, когда живые системы становятся достаточно сложными, они приобретают способность к моделированию окружающей среды. Выбор траектории из числа доступных альтернатив в этом случае становится не просто слепой физической реакцией объекта на внешний стимул, но результатом учета возможных последствий такого выбора в рамках проводимого моделирования. Если говорить в терминах концепции Д. Деннета, то на этом этапе «скиннеровские» существа становятся «попперовскими» [Dennett, 1997]. Такой переход на следующий уровень в структурном развитии позволяет системам, которые уже с полным правом можно назвать агентами [Bratman, 2007, p. 62–63], формировать гипотезы в рамках имеющейся у них модели действительности. Теперь агенты могут позволить гипотезам погибать вместо себя. Здесь вариативность перестает быть слепой. Альтернативные траектории поведения могут быть «проиграны» агентом в его модели действительности до того, как какая-либо из них будет реализована в физическом мире.

Структурными элементами такого рода моделей действительности, которыми обладают агенты, являются установки этих агентов.

Установки – это интенциональные состояния агентов, т.е. состояния, описываемые как имеющие некоторое содержание, «состояния о чем-то».

Ключевая особенность этого определения заключается в том, что речь здесь идет о возможности описания некоторых состояний агента определенным образом (как имеющих некоторое содержание). Такого рода условие позволяет отождествлять приписываемость тех или иных установок некоторому агенту с их наличием у агента. Иметь установку и быть тем, кому эта установка может быть обоснованно приписана, – это в таком случае буквально одно и то же. В чем же причина такого нашего стремления отождествить, казалось бы, различные вещи?

Дело в том, что все разнообразные термины, выражающие установки, выполняют в естественном языке одну и ту же функцию – их приписывание позволяет агентам объяснять и предсказывать наблюдаемые действия друг друга. Такого рода объяснения и предсказания составляют самую суть такого явления, как «народная психология», а сами установки представляют собой основу ее словаря. При этом вопрос об уместности приписывания агентам тех или иных установок – это вопрос о соотношении наблюдаемого нами поведения этих агентов и разделяемых нами установок народной психологии. Сами термины, выражающие в языке такие установки, появляются в нем и осваиваются агентами ради объяснения и предсказания поведения *других* людей, а не самих себя. Это так потому, что для знания собственных психологических состояний и объяснения мотивов собственного поведения агенту, вообще говоря, не требуется язык. Ему незачем было бы описывать самому себе мотивы собственных действий. Однако научившись обоснованно применять термин «хочет» к другим людям, агент в какой-то момент оказывается способен применить его и к себе и тем самым оказывается способен объяснить себе и окружающим собственные действия.

Интерпретативистские подходы обладают одним весьма значимым преимуществом по сравнению со своими конкурентами. Они позволяют сравнительно простым и понятным образом ответить на вопрос о том, каким образом какие-либо состояния, события или процессы, случающиеся с агентами, могут «быть о чем-то», отличном от этих процессов, о чем-то, находящемся вне тел агентов, а, быть может, находящемся на значительном от них удалении как в пространстве, так и во времени. Ответ интерпретативизма заключается в том, что мы можем описывать эти состояния агентов так, *как будто* они представляют собой состояния «о чем-то». Причем практики такого нашего приписывания некоторым состояниям агентов того или иного содержания будут обоснованными в той мере, в какой такие приписывания позволяют нам корректно объяснять и предсказывать поведение данных агентов.

Приписывая агентам пропозициональные установки, мы приписываем им уместность реагирования определенным образом при выполнении тех или иных условий, характеризуемых в терминах интенционального содержания. Тем самым мы вписываем агентов в нормативный контекст.

Условия уместности включения в нормативные отношения

Таким образом, уместность приписывания системам искусственного интеллекта понятий из словаря «народной эпистемологии» определяется выполнением условий для того, чтобы их можно было характеризовать в качестве нормативных эпистемических агентов [Sosa, 2013]. При этом наша склонность приписывать кому-либо ту или иную установку не является безошибочной. Соответственно, «уместность приписывания» представляет собой нормативное понятие, содержание которого выражается в том, что идеально рациональный и информированный агент в рассматриваемой ситуации был бы склонен приписать некоторому другому агенту – в нашем случае искусственному интеллекту – определенную пропозициональную установку.

Это, в свою очередь, требует выполнения следующих условий: (а) наличие у систем искусственного интеллекта способности к вариативному поведению; (б) наличие у них контроля за выбором реализуемой траектории поведения; (в) наличие у них способности к взвешиванию оснований при выборе реализуемой траектории из числа номологически доступных альтернатив.

Эти условия позволяют отличать осмысленное и целенаправленное поведение от детерминированных последовательностей физических событий, в отношении которых неуместно вынесение нормативной оценки, как это имеет место, например, в случае физических процессов, реализующих вычисления в калькуляторе.

При этом выполнение данных условий является принципиально доступным для проверки.

Во-первых, существующие модели искусственного интеллекта демонстрируют способность к реализации различных форм поведения (ответов) при одинаковых внешних стимулах (запросах) [Wang et al., 2023]. В основе такой вариативности, так же как и в случае вариативности поведения живых организмов, лежат принципы нелинейной динамики.

Во-вторых, модели искусственного интеллекта демонстрируют способность придерживаться определенной стратегии поведения в течение некоторого времени. Стоит, конечно, признать, что параметр так называемой агентной стабильности (agent stability) остается пока одним из наиболее уязвимых мест для существующих в настоящий момент искусственных систем [Milliere, Buckner, 2024, p. 32]. И все же при всех подобного рода оговорках мы имеем все основания говорить об имеющей у них минимальной агентной стабильности, поскольку в противном случае мы наблюдали бы перманентную непоследовательность в их поведении, чего в действительности не происходит.

В-третьих, эти модели демонстрируют отзывчивость к предъявляемым им основаниям [Steyvers et al., 2025, p. 222]. Если при взаимодействии с ними им предоставить ссылку на авторитетные данные, вступающие в прямое противоречие с выбранной ими стратегией дискурсивного поведения, то они склонны релевантным образом реагировать на такие свидетельства и менять свою стратегию поведения.

Выполнение этих трех условий свидетельствует о том, что существующие на сегодняшний день модели искусственного интеллекта могут быть охарактеризованы в качестве нормативных эпистемических агентов. Они способны вступать в полноценные нормативные эпистемические отношения с другими эпистемическими агентами, включая людей. Это позволяет обоснованно приписывать им широкий спектр эпистемических понятий с тем, чтобы наиболее информативным и релевантным образом объяснять их дискурсивное поведение для других эпистемических агентов.

Список литературы

Бажанов, 2025 – *Бажанов В.А.* Феномен числа с позиций психологизма: можно ли утверждать, что числа «сочиняются»? // Эпистемология и философия науки. 2025. Т. 62. № 1. С. 134–147.

Вострикова, 2025 – *Вострикова Е.В.* Знание и контекст. Размышления о книге Майкла Бломе-Тилльманна «Семантика атрибуции знания» // Эпистемология и философия науки. 2025. Т. 62. № 1. С. 215–233.

References

Alvarado, 2023 – Alvarado, R. “AI as an Epistemic Technology”, *Science and Engineering Ethics*, 2023, vol. 29, pp. 1–18.

Anscombe, 2000 – Anscombe, E. *Intention*. Cambridge, MA: Harvard University Press, 2000, 94 p.

Arsenault, 2026 – Arsenault, L.M. “Humanizing: On the Anthropomorphization of AI as a Tactic of Cultural Integration”, *AI & Society*, 2026, vol. 41, pp. 3789–3799.

Bazhanov, 2025 – Bazhanov, V.A. “Fenomen chisla s pozitsii psikhologizma: mozhno li utverzhdat’, chto chisla ‘sochinyayutsya’?” [The Phenomenon of Number From The Lens of Psychologism: Can We Claim that Numbers Are Being “Made Up”?], *Epistemology & Philosophy of Science*, 2025, vol. 62, no. 1, pp. 134–147. (In Russian)

Blackburn, 1998 – Blackburn, S. *Ruling Passion: A Theory of Practical Reasoning*. Oxford: Clarendon Press, 1998.

Bratman, 2007 – Bratman, M. *Structures of Agency: Essays*. Oxford: Oxford University Press, 2007.

Budnik, 2025 – Budnik, C. "Can We Trust Artificial Intelligence?", *Philosophy & Technology*, 2025, vol. 38, pp. 1–23.

Buijsman, 2025 – Buijsman, S. "Conceptual Disconnect: The Prospect of Concept-Based AI Explanations", in: Pantsar, M., Stjernfelt, F., Gramelsberger, G., Olteanu A. (eds.) *Philosophy of Artificial Intelligence: Optimistic and Pessimistic Views*. Cham: Springer, 2025, pp. 1–16.

Cappelen, Dever, 2021 – Cappelen, H., Dever, J. *Making AI Intelligible: Philosophical Foundations*. Oxford: Oxford University Press, 2021, 175 pp.

Cappelen, Dever, 2025a – Cappelen, H., Dever, J. *Going Whole Hog: A Philosophical Defense of AI Cognition*, 2025. DOI: 10.48550/arXiv.2504.13988.

Cappelen, Dever, 2025b – Cappelen, H., Dever, J. "A Hyper-Externalist Manifesto for LLMs", in: Cappelen, H., Dever, J. (eds.) *Communicating with AI: Philosophical Perspectives*. Oxford: Oxford University Press, 2025, pp. 1–21.

Cash et al., 2025 – Cash, T., Oppenheimer, D., Christie, S., Devgan, M. "Quantifying Uncert-AI-nty: Testing the Accuracy of LLMs' Confidence Judgments", *Memory & Cognition*, 2025, vol. 54, pp. 375–400.

Cochrane, 2018 – Cochrane, T. *The Emotional Mind: A Control Theory of Affective States*. Cambridge: Cambridge University Press, 2018.

Dennett, 1997 – Dennett, D. *Kinds of Minds: Toward an Understanding of Consciousness*. New York: Basic Books, 1997.

Hardy-Vallee, Dubreuil, 2010 – Hardy-Vallee, B., Dubreuil, B. "Folk Epistemology as Normative Social Cognition", *Review of Philosophy and Psychology*, 2010, vol. 1, pp. 483–498.

Humphreys, 2004 – Humphreys, P. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. New York: Oxford University Press, 2004.

Ma, Valton, 2024 – Ma, W., Valton, V. "Toward an Ethics of AI Belief", *Philosophy & Technology*, 2024, vol. 37, pp. 1–28.

Milliere, Buckner, 2024 – Milliere, R., Buckner, C. "A Philosophical Introduction to Language Models – Part II: The Way Forward", 2024, DOI: 10.48550/arXiv.2405.03207.

Shanahan, 2024 – Shanahan, M. "Talking about Large Language Models", *Communications of the ACM*, 2024, vol. 67, no. 2, pp. 68–79.

Peacocke, 1999 – Peacocke, C. *Being Known*. Oxford: Oxford University Press, 1999.

Price, 1989 – Price, H. "Defending Desire-as-Belief", *Mind*, 1989, vol. 98, pp. 119–127.

Scorici et al., 2024 – Scorici, G., Schultz, M.D., Seele, P. "Anthropomorphization and Beyond: Conceptualizing Humanwashing of AI-enabled Machines", *AI & Society*, 2024, vol. 39, pp. 789–795.

Sellars, 1997 – Sellars, W. *Empiricism And The Philosophy Of Mind*. Cambridge, MA: Harvard University Press, 1997.

Shin, 2025 – Shin, D. Automating Epistemology: How AI Reconfigures Truth, Authority, and Verification", *AI & Society*, 2025, vol. 41, pp. 1553–1559.

Simbolon et al., 2025 – Simbolon, L., Manugeren, M., Barus, E. "Does AI Know Things? An Epistemological Perspective on Artificial Intelligence", *Journal of English Language and Education*, 2025, vol. 10, pp. 1022–1028.

Simion, Kelp, 2023 – Simion, M., Kelp, C. “Trustworthy Artificial Intelligence”, *Asian Journal of Philosophy*, 2023, vol. 2, pp. 1–12.

Sosa, 2013 – Sosa, E. “Epistemic Agency”, *The Journal of Philosophy*, 2013, vol. 110, pp. 585–605.

Steyvers et al., 2025 – Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Smyth, P. “What Large Language Models Know and What People Think They Know”, *Nature Machine Intelligence*, 2025, vol. 7, pp. 221–231.

Vostrikova, 2025 – Vostrikova, E. “Znanie i kontekst. Razmyshleniya o knige Maikla Blome-Till’manna “Semantika atributsii znaniya” [Knowledge and Context. Reflections on Michael Blome-Tillmann’s “Semantics of Knowledge Attribution”], *Epistemology & Philosophy of Science*, 2025, vol. 62, no. 1, pp. 215–233. (In Russian)

Wang et al., 2023 – Wang, C., Liu, X, Awadallah, A. “Cost-Effective Hyperparameter Optimization for Large Language Model Generation Inference”, *Proceedings of Machine Learning Research*, 2023, vol. 224, pp. 1–17.