Эпистемология и философия науки
2026. Т. 63. № 3. С. 113–131
УДК 167.7

Epistemology & Philosophy of Science
2026, vol. 63, no. 3, pp. 113–131
DOI: https://doi.org/10.5840/eps202663344

К НОВЫМ ПОДХОДАМ В ИССЛЕДОВАНИИ МАШИННОГО СОЗНАНИЯ*

Гришин Павел Викторович – стажер-исследователь. Национальный исследовательский университет «Высшая школа экономики». Российская Федерация, 105066, г. Москва, ул. Старая Басманная, д. 21/4, стр. 1;
e-mail: pvgrishin@edu.hse.ru



Статья посвящена современным подходам к исследованию сознания ИИ и их методологическим пределам. Анализ выстраивается на основе классификации Джонатана Бёрча, в рамках которой рассматриваются теоретико-нагруженные, теоретико-нейтральные и теоретико-легкие подходы; отдельно обсуждаются многомерные и многоуровневые стратегии. Исходной установкой исследования выступает позитивный агностицизм относительно возможности машинного сознания. Показано, что, несмотря на существенные различия между существующими подходами, большинство из них в недостаточной степени проблематизирует перенос понятий и методов, разработанных для изучения сознания биологических существ, на вычислительные системы. В связи с этим утверждается, что центральная трудность современного исследования заключается в недостаточной обоснованности указанного переноса. Нехватка эмпирических данных или отсутствие надежного теста, при всей их значимости, не являются ключевыми трудностями. Особое внимание уделяется вычислительному функционализму, который либо явно принимается в качестве предпосылки исследования, либо неявно задает его рамку, либо становится предметом критики. Обосновывается, что наиболее продуктивными в этих условиях являются многомерные и многоуровневые подходы, поскольку они позволяют локализовать ключевые методологические трудности и представить их в виде более частных исследовательских задач, не претендуя на окончательное решение проблем. Делается вывод, что пределы современных исследований машинного сознания во многом определяются унаследованной понятийной рамкой; в этой связи предлагается общая исследовательская установка для дальнейшей работы над проблемой.

Ключевые слова: сознание, искусственное сознание, машинное сознание, ИИ, философия сознания, проблема других сознаний, вычислительный функционализм, антропоцентризм, агностицизм

* Исследование выполнено за счет гранта Российского научного фонда № 26-18-00253, https://rscf.ru/project/26-18-00253/. Russian Science Foundation, RSF. https://rscf.ru/en/project/26-18-00253

© Гришин П.В., 2026

TOWARDS NEW APPROACHES IN MACHINE CONSCIOUSNESS RESEARCH

Pavel V. Grishin –
Research Assistant.
HSE University.
21/4 Staraya Basmannaya St.,
Moscow 105066,
Russian Federation;
e-mail: pvgrishin@edu.hse.ru

This article examines contemporary approaches to the study of AI consciousness and their methodological limits. The analysis is structured around Jonathan Birch's classification, focusing on theory-heavy, theory-neutral, and theory-light approaches, while also treating multidimensional and multilevel strategies as a distinct line of inquiry. The starting point of the article is a position of positive agnosticism regarding the possibility of machine consciousness. It is argued that, despite their significant differences, most existing approaches do not sufficiently problematize the transfer of concepts and methods developed for the study of consciousness in biological beings to computational systems. On this basis, it is argued that the central difficulty in the contemporary research lies in the insufficient justification of the aforementioned transfer. The lack of empirical data or the absence of a reliable test, while significant, do not constitute the key difficulties. Particular attention is paid to computational functionalism, which is either explicitly adopted as a presupposition, implicitly built into the framework of inquiry, or made the object of criticism. In these circumstances, multidimensional and multilevel approaches prove most productive, as they allow one to localize key methodological difficulties and reframe them as more specific research tasks, without claiming to offer a definitive resolution of the problems. The conclusion is that the limits of current research on machine consciousness are shaped to a large extent by an inherited conceptual framework; accordingly, the article proposes a general research orientation for further work on the problem.

Keywords: consciousness, artificial consciousness, machine consciousness, AI, philosophy of mind, philosophy of consciousness, problem of other minds, problem of other consciousness, computational functionalism, anthropocentrism, agnosticism

Введение

Вопрос о сознании ИИ в последние годы перестал быть маргинальным. Быстрый рост возможностей больших языковых моделей сделал эту тему не только общефилософской, но и методологической [Bayne et al., 2024; Butlin et al., 2023; Butlin et al., 2025; Evers et al., 2024; Schwitzgebel, 2025; Shevlin, 2021]. Однако современная дискуссия развивается в условиях двойной неопределенности. С одной стороны, нет общепринятой теории сознания даже для человека. С другой стороны, именно сознание человека по-прежнему служит главным источником понятий, тестов и теорий, которые затем пытаются переносить на искусственные системы.

Из-за этого почти все исследования сознания ИИ строятся вокруг одного и того же хода – переноса. Исследователь либо ищет



выделяемые современной наукой о сознании функциональные и архитектурные пересечения [Butlin et al., 2023; Goldstein, Kirk-Giannini, 2024; Juliani et al., 2022], либо, напротив, подчеркивает, что такой перенос пока не оправдан из-за биологической специфики сознания [Aru et al., 2023; Block, 2025; Findlay et al., 2024; Seth, 2025]. Даже более осторожные программы – от теоретико-нейтральных тестовых стратегий, вроде ACT-теста [Schneider, 2019; Schneider, 2020], до теоретико-легких эмпирических проверок [Dung, 2023], а также многомерных подходов [Chalmers, 2023; Evers et al., 2024; Shevlin, 2021] – все еще работают внутри словаря, выработанного в исследованиях человека и животных.

Центральная трудность находится глубже чем недостаточность оснований для уверенного ответа на вопрос, сознательны ли современные ИИ, и на вопрос, способна ли машина иметь сознание в принципе. Она состоит в легитимности использования базовых понятий этой дискуссии применительно к искусственным системам, так как наше вопрошание по-прежнему остается в рамках переноса понятий, выработанных для сознания биологического типа. Особую уязвимость здесь демонстрирует понятие феноменального опыта и отчета о нем - к этому мы еще вернемся.

Если сознание ИИ вообще возможно, то весьма вероятно, что оно потребует не частичной модификации готового аппарата, а его радикального пересмотра. Такая ревизия в статье не производится, но все же выделяются общие установки исследования, следуя которым содержание и состав существующих понятий должны измениться.

Статья построена следующим образом. Сначала задаются предпосылки анализа. Затем поочередно разбираются подходы к проблеме других сознаний по Бёрчу (в контексте ИИ). Далее отдельный раздел кратко обозревает дискуссию вокруг вычислительного функционализма (*computational functionalism*). После этого обсуждаются перспективные агностические и многомерные ответы на нынешние затруднения. В заключении формулируется общая проблема переноса понятий и намечается возможный вектор дальнейшего исследования.

Контекст и предпосылки

Прежде всего необходимо уточнить, о каком именно круге вопросов идет речь. Философия сознания (*philosophy of consciousness*) наследует словарь, методы и проблемы философии разума (*philosophy of mind*). Однако фактически философов сознания интересует другой класс вопросов. В то время как философия разума изучает, например,

репрезентации сами по себе, философию сознания они интересуют лишь в той мере, в какой имеют отношение к феноменальному опыту, вопросу «каково это быть», глобальному рабочему пространству и другим концептам сознания. Поэтому перевод mind как сознания является неточностью, преодоление которой входит в задачу строгого исследования[1]. По этой причине далее речь пойдет исключительно о проблеме других сознаний, а не о проблеме других разумов.

Проблема других сознаний гораздо более детально проработана в контексте животных, чем в контексте ИИ. Поэтому смежная литература по сознанию животных полезна здесь не как готовое решение, а как источник методологических предупреждений. Так, Джонатан Бёрч подчеркивает, что изучение нечеловеческого сознания требует специальных стратегий, поскольку имеющиеся модели сознания невозможно механически расширять на другие системы [Birch, 2022]. Генри Шевлин называет эту трудность *проблемой специфичности* [Shevlin, 2021]. Если теория формулируется слишком близко к когнитивному устройству человека, она дает ложные отрицания: *справедливо ли заключить, что осьминоги не имеют сознания, только потому, что морфология их мозга кардинально отличается от морфологии мозга человека?* Если формулируется слишком абстрактно, она начинает порождать ложные положительные результаты: *можно ли назвать ИИ сознательным при условии, что его архитектура в некоей степени функционально повторяет мозг человека?* Для ИИ проблема специфичности стоит особенно остро, поскольку разрыв между человеком и машиной больше, чем между человеком и любым животным.

Эти трудности получают дальнейшее развитие у Хизер Браунинг и Уолтера Вейта, предлагающих разделять *проблему валидности индикаторов* и *проблему экстраполяции* [Browning, Veit, 2020]. Смысл первой состоит в том, что маркер сознания может оказаться ненадежным: трудно гарантировать, что он фиксирует именно сознательный опыт, а не сопутствующий когнитивный процесс. Иными словами, еще до попытки переноса необходимо доказать, что выбранный маркер действительно имеет устойчивую связь с сознанием, а не просто коррелирует с ним в случае человека. Проблема экстраполяции возникает на следующем шаге: даже если маркер хорошо работает на примере людей, это не гарантирует его релевантность для животных и тем более для искусственных систем. Это

[1] Автор может рассматривать вопрос о существовании врожденных языковых механизмов – а это дискуссия из области философии разума. В русской литературе его текст с большой вероятностью окажется под рубрикой философии сознания, хотя речь идет исключительно о бессознательных процессах. Это абсурдная ситуация.

разграничение помогает локализовать методологические препятствия: оно четко разводит сомнения в надежности самих индикаторов и сомнения в правомерности их переноса на принципиально иные системы.

Нужно учитывать и материальный разрыв между человеком и машиной: в одном случае речь идет о живой биологической ткани, состоящей из клеток, белков, воды и других органических структур, в другом – об устройствах, построенных из кремния, металлов и иных неорганических компонентов. Уже по одному этому основанию любое животное ближе к человеку, чем нынешние системы ИИ. Машина может демонстрировать языковые компетенции – способность, являющуюся одной из главных отличительных черт человека. При этом она остается радикально чуждой по своему устройству. И напротив, существо, вовсе не владеющее языком, может быть намного ближе к человеку как биологический родственник и носитель сходной формы организации нервной системы и тела вообще. *Отсюда возникает парадоксальная асимметрия, которую важно иметь в виду.*

Из-за этой асимметрии совпадение по поведению или по части функций еще не дает оснований предполагать более глубокое сходство в релевантных для сознания вещах. Вместе с тем этот кардинальный разрыв между человеком и машиной частично сужается. Переход от символического ИИ к коннекционизму сделал его в некотором смысле ближе к биологическим системам. Однако это сходство остается грубым и частичным: глубокие нейросети заимствуют лишь общий принцип иерархичности. Основным методом обучения служит обратное распространение ошибки – алгоритм, биологическая достоверность которого остается спорной [Lillicrap et al., 2020; Whittington, Bogacz, 2019]. Рекуррентность лишь частично приближается в отдельных разработках вроде спайковых нейронных сетей, однако они пока сохраняют лабораторный статус [Pfeiffer, Pfeil, 2018]. Поэтому методологическая задача состоит не в простой фиксации несходства ИИ и человека, а в более точном анализе того, какие различия здесь действительно принципиальны и какие сходства вообще можно считать значимыми.

На этом фоне наиболее оправданной исходной позицией выглядит *агностицизм*. Он фиксирует, что имеющиеся данные пока не позволяют обоснованно перейти ни к уверенному признанию, ни к уверенному отрицанию этой возможности [McClelland, 2024; Schwitzgebel, 2025]. Вместе с тем агностицизм не обязательно должен быть нейтральным. В рамках обзора он понимается скорее как позитивный агностицизм – признание нынешней неопределенности при одновременной открытости к тому, что искусственное сознание принципиально возможно, даже если условия его возникновения и критерии верификации еще не прояснены. *Если искусственное*

сознание возможно, то ключевой вопрос не в том, будет ли оно «настоящим», т.е. как у человека. Но в том, какова специфика вычислительных аналогов сознательных процессов и какими средствами эту специфику вообще можно исследовать.

В обзор включены наиболее показательные линии современной дискуссии о сознании ИИ. В них особенно ясно проявляется вопрос о правомерности переноса готовых философских и научных понятий сознания с биологических систем на компьютерные. Новая классификация подходов здесь не предлагается. В качестве рабочей используется классификация Бёрча [Birch et al., 2022]. *Теоретико-нагруженный* подход ищет строгие соответствия между взятой теорией сознания и архитектурой рассматриваемой системы. *Теоретико-нейтральный* подход, напротив, отказывается от теории и, как правило, оценивает сознание на основе поведения. *Теоретико-легкий* вместо структурных соответствий полагается на центральную и проверяемую эмпирически гипотезу.

Эта классификация ставит в центр отношение того или иного подхода к теории сознания, что делает ее особенно уместной: научное исследование сознания неизбежно строится вокруг теорий, тогда как сама наука о сознании еще находится в младенческом состоянии и не располагает общепризнанными, хорошо подтвержденными теориями. В таких условиях именно отношение подхода к теории предопределяет обоснованность получаемых выводов. Критерием оценки служит то, насколько обоснованно тот или иной подход работает с разрывом между человеческим и искусственным сознанием и насколько в нем отрефлексированы ограничения, связанные с переносом понятий.

Теоретико-нагруженный подход и его цена

Наиболее влиятельное и систематическое воплощение теоретико-нагруженный подход получил в недавних работах группы исследователей под руководством Патрика Батлина [Butlin et al., 2023]. Выводя из нескольких научных теорий сознания ряд функциональных и архитектурных требований, они проверяют, действительно ли современный ИИ выполняет эти требования. В этом же состоит и более поздняя формулировка метода [Butlin et al., 2025].

Главное преимущество такого подхода – перевод абстрактных споров о машинном сознании в плоскость анализа конкретных архитектур и функций. С точки зрения теории глобального рабочего пространства (GWT) содержание сознания должно быть глобально доступно всем функциональным модулям системы. Например, попавшая через модуль зрения в рабочее пространство зрительная

информация доступна модулю производства речи – люди могут рассказать, что они в данный момент видят[2]. Аналогичным образом из других научных теорий выводятся собственные индикаторные свойства. Благодаря этому исследование приобретает вид систематичной научной программы: проверка того, какие именно теоретически значимые механизмы сознания реализованы в конкретной модели ИИ.

Внутри этого направления важно различать два разных исследовательских хода. Первый – это поиск индикаторов сознания в уже существующих системах. Второй – попытка изначально разработать систему на основе механизмов, которые ведущие теории считают необходимыми для возникновения сознания[3]. Несмотря на очевидные различия между анализом готовых моделей и созданием новых систем, оба случая относятся именно к теоретико-нагруженному подходу. В обоих случаях определяющую роль играет предварительно принятая теория сознания и строгость соответствия ее требованиям.

Вместе с тем именно эта теоретическая определенность образует и главное ограничение. С самого начала подход опирается на истинность вычислительного функционализма: предполагается, что реализованная при помощи вычислений правильная функциональная организация необходима и достаточна для сознания [Butlin et al., 2023; Butlin et al., 2025]. Этот тезис не выводится из эмпирических данных, а постулируется заранее – как мост, открывающий путь

[2] В случае GWT речь идет о наличии множества специализированных систем, способных работать параллельно [GWT-1], рабочего пространства с ограничением пропускной способности и механизмом селективного внимания [GWT-2], глобальной трансляции, т.е. доступности содержаний рабочего пространства для всех модулей [GWT-3], и зависящего от состояния внимания, позволяющего последовательно опрашивать разные модули при выполнении сложных задач [GWT-4]. Аналогичным образом из других теорий выводятся собственные индикаторные свойства, включая рекуррентную обработку, метакогнитивный мониторинг, предиктивное кодирование, схему внимания, агентность и элементы воплощенности [Butlin et al., 2023].

[3] Эта линия представлена в [Juliani et al., 2022], в тесной связи с [Juliani et al., 2022], где архитектура Perceiver прямо интерпретируется как функциональное глобальное рабочее пространство: вычислительный аналог области, интегрирующей информацию из разных модальностей и делающей ее доступной для дальнейшей обработки. К этой же линии примыкает исследование, где предлагается архитектура Global Latent Workspace, т.е. глобальное латентное пространство [Rullen, Kanai, 2020]. В [Graziano, 2017] схема внимания трактуется как основа для конструирования искусственного сознания, а в [Wilterson, Graziano, 2021] и [Dossa et al., 2024] уже на уровне моделей и агентов показано, что подобные архитектурные элементы могут давать функциональные преимущества в контроле внимания и мультимодальной координации.

к переносу концепций человеческого сознания на ИИ[4]. Это означает, что философски спорное допущение предшествует исследованию, а не следует из него. Подробный критический разбор вычислительного функционализма приведен далее.

В результате теоретико-нагруженный подход покупает эмпирическую операциональность ценой сильного метафизического допущения. Он позволяет начать исследование, но не доказывает, что его понятийная рамка адекватна предмету. В лучшем случае он показывает, что некоторые архитектурные свойства, ассоциированные с человеческим сознанием, могут быть вычислительно реализованы, инженерно-воспроизведены и проверены в ИИ. Однако из этого следует лишь то, что такие механизмы могут быть когнитивно полезны. То, что мы научились воспроизводить отдельные положения научных теорий сознания в виде вычислений, еще не делает законным перенос на ИИ понятийного аппарата, выработанного для феноменального мира человека.

Теоретико-нейтральный подход

Ввиду отсутствия общепризнанной теории сознания некоторые исследователи отказываются от теоретической опоры вообще. Именно так устроены различные теоретико-нейтральные предложения, включая ACT-тест Сьюзан Шнайдер [Schneider, 2019; Schneider, 2020] и тест Тьюринга. ACT-тест является попыткой модифицировать тест Тьюринга для оценки именно сознания, а не только способности имитировать человеческое языковое поведение. Он включает серию вопросов, затрагивающих такие понятия, как субъективный опыт, восприятие, самосознание, эмоции и отношение системы к собственному существованию: «Есть ли что-то, что значит быть вами прямо сейчас?», «Считаете ли вы свое существование чем-то большим, чем просто выполнение программы?», «Как бы вы отнеслись к перспективе необратимого удаления вашей программы?». Основная идея состоит в том, что осмысленные, глубокие и последовательные ответы на подобные вопросы (при условии, что система не была специально обучена на философских и научных текстах о сознании) могут свидетельствовать о наличии некоторой формы саморефлексии и сознания [Schneider, 2019; Schneider, 2020]. Иначе говоря, концепция сознания должна самостоятельно возникать из предполагаемой ментальной

[4] Показательно, что теория интегрированной информации (IIT) прямо исключается из основного набора теорий именно потому, что в стандартной форме несовместима с вычислительным функционализмом [Butlin et al., 2023]. Будет пояснено в разделе про вычислительный функционализм.



жизни системы, а не быть простым воспроизведением заученных паттернов. Достоинство таких тестов в том, что они дают исследователю хотя бы некоторую процедуру проверки в ситуации, когда наука о сознании распадается на конкурирующие теории.

Критика этих тестов, однако, сегодня приобрела особую убедительность. Даже весьма изощренные ответы на вопросы о сознании могут порождаться продвинутыми языковыми моделями за счет статистических закономерностей в обучающих данных и сложной лингвистической комбинаторики, без какого-либо реального внутреннего опыта и без специального обучения [Udell, Schwitzgebel, 2021]. Дополнительно показано, что некоторые высказывания ИИ о собственном устройстве можно понимать лишь как поверхностные или минимальные формы самоотнесения, которые не дают оснований говорить о феноменальном опыте [Comsa, Shanahan, 2025]. Тем самым развитая способность системы говорить о себе перестает быть убедительным аргументом в пользу сознания и, говоря откровенно, никогда им не была.

Убедительнее выглядит позиция Бёрча, согласно которой полностью теоретико-нейтральный подход к сознанию внутренне противоречив, поскольку в процессе развития в конечном счете приходит к необходимости хотя бы минимальных теоретических установок [Birch et al., 2022]. Отказ от явной теоретической рамки не устраняет теоретические допущения, а лишь делает их неявными и потому менее контролируемыми. Поэтому отсутствие теории здесь не преимущество, а скорее слабость подхода. Кроме того, именно теоретико-нейтральный подход наименее рефлексивно относится к разрыву между человеком и машиной, поскольку опирается исключительно на внетеоретические бихевиоральные данные, с помощью которых этот разрыв в принципе невозможно ухватить: поведенческое сходство само по себе не показывает, какие именно внутренние процессы стоят за наблюдаемыми ответами ИИ.

Научное изучение сознания вообще невозможно без теоретической основы: выводы о наличии или отсутствии сознания не могут основываться исключительно на критериях фолк-психологии, интуициях или впечатлении от правдоподобного поведения. Они должны опираться на проверяемую гипотезу, из которой ясно, почему именно тот или иной признак считается значимым, и допускать дальнейшую критику. В этом смысле теоретико-нейтральный подход не может служить инструментом научного изучения сознания, поскольку противоречит самой логике научной программы.

Теоретико-легкий подход

В версии Бёрча теоретико-легкий подход означает не полный отказ от теории, а отказ от привязки к одной научной теории сознания в пользу более ограниченной, но эмпирически проверяемой центральной гипотезы. В случае Бёрча такой гипотезой выступает *гипотеза усиления*: феноменально осознанное восприятие стимула, по сравнению с бессознательным восприятием, повышает успешность выполнения определенных когнитивных задач, связанных с этим стимулом [Birch, 2022].

Для пояснения обратимся к классическим психологическим опытам на человеке. Представим, что испытуемому на долю секунды демонстрируют слово, мгновенно перекрывая его другим изображением (маскирование). Увиденное слово не доходит до содержания сознания человека, однако на бессознательном уровне обработка все же происходит. Если убрать маскирование и дать человеку время осознать стимул, его когнитивные возможности драматически возрастают: теперь он может удерживать это слово в памяти, использовать его для построения сложных ассоциаций, гибко менять свое поведение через несколько минут или вообще отказаться от выполнения задачи. Именно эта разница в успешности действий – с осознанием стимула и без него – образует *профиль усиления*. Бёрч полагает, что, если мы обнаружим схожий профиль у животного, это может служить весомым свидетельством в пользу наличия у него сознания.

Есть исследователи, выступающие за применение этого же подхода к ИИ. В работе Леонарда Данга утверждается, что тесты на сознание для животных можно перенести и на машины, если исключать ad hoc системы, специально подстроенные под прохождение теста, и искать у неподогнанного ИИ такие кластеры способностей, которые в человеческом случае усиливаются сознательным восприятием [Dung, 2023]. Сюда же относятся маскирование и двойная диссоциация. В этом смысле данное предложение развивает бёрчевскую теоретико-легкую линию и гипотезу усиления [Birch, 2022] применительно к искусственным системам.

Однако в случае ИИ трудность оказывается гораздо глубже, чем риск банальной мимикрии системы под требования эксперимента. У Бёрча и затем у Данга маскирование понимается как способ проверить гипотезу усиления [Birch, 2022; Dung, 2023]. Представим, что ученые, прежде изучавшие лошадей, переходят к человеку и спрашивают: «Скачут ли люди?». Такой вопрос уже содержит категориальную ошибку. Любая эмпирическая проверка человеческого движения, построенная на этом вопросе, будет давать вводящие в заблуждение результаты.



То же происходит и с маскированием в отношении ИИ. Нет оснований полагать, что категория феноменального опыта, о котором ИИ должен был бы отчитаться, вообще к нему применима. Это не доказывает невозможность машинного сознания и не исключает, что такая система может обладать некоторой формой опыта. Но это означает, что тесты на сознание, основанные исключительно на отчете о феноменальном опыте, здесь невозможны в строгом смысле, поскольку не показано, что данная категория вообще имеет денотат в случае ИИ. Следовательно, теоретико-легкий подход в версии Данга и гипотеза усиления Бёрча применительно к искусственному интеллекту оказываются крайне ограниченными.

Проблемы вычислительного функционализма

Биологические и каузальные возражения направлены прежде всего против той философской предпосылки, на которой держится значительная часть современных исследований сознания ИИ. Мейнстримный теоретико-нагруженный подход опирается на вычислительный функционализм вполне явно, тогда как другие подходы нередко сохраняют ту же установку неявно. Вычислительный функционализм в данном контексте закрывает разрыв между человеком и машиной: согласно нему для реализации сознания на компьютере необходимо и достаточно реализовать при помощи вычислений лежащую в основе сознания человека функциональную организацию. Если сознание сводится лишь к определенной функциональной организации, это открывает путь к некритическому переносу любых умозрительных построений о сознании, допускающих вычислительную интерпретацию. Первоначально эта позиция была разработана в контексте философии разума, прежде всего в работах Хилари Патнэма [Putnam, 1960; Putnam, 1967]. Впоследствии вычислительный функционализм был перенесен и в философию сознания, где использовался для аргументации в пользу принципиальной возможности сознательного ИИ, в частности у Чалмерса [Chalmers, 1996].

Однако эта позиция сталкивается с серьезной эмпирической критикой, настаивающей на несводимости сознания к абстрактным вычислениям. Так, с позиций нейронауки подчеркивается, что современные ИИ-системы лишены фундаментальных биологических механизмов, определяющих сознание млекопитающих [Aru et al., 2023]. Среди них – глубокая сенсорная укорененность в физическом мире, таламо-кортикальная архитектура мозга, многоуровневая самоподдерживающаяся организация и эволюционно обусловленная агентность. Ценность такого нейробиологического взгляда в том, что он переходит от общих деклараций к демонстрации конкретных

структурных лакун между большими языковыми моделями и живыми организмами.

Еще более радикально эту критику развивает Анил Сет, последовательно отстаивающий принципы *биологического натурализма* [Seth, 2025]. Его аргумент строится вокруг идеи о том, что сознание неразрывно связано с природой живого, самоподдерживающегося организма. Сет связывает вычисления с материальной смертью (*mortal computation*), указывая, что подлинно сознательные процессы неотделимы от конечности, уязвимости когнитивной системы. Напротив, алгоритмы современного ИИ безразличны к физической платформе, на которой они запущены: программный код здесь принципиально отделен от аппаратного обеспечения. Из этого Сет делает вывод, что на пути развития классического цифрового ИИ возникновение сознания крайне маловероятно. Перспектива создания сознательной машины остается открытой разве что для нейроморфных или синтетико-биологических систем, которые будут не просто имитировать мозг, но функционировать по законам биологии.

Философский вызов функционализму бросает теория интегрированной информации (IIT). Ее сторонники доказывают невозможность приравнивания функциональной эквивалентности к феноменальной эквивалентности [Findlay et al., 2024]. Два устройства могут выполнять одну и ту же задачу, обладать идентичными входами и выходами и быть полностью взаимозаменяемыми на уровне вычислительных ролей, но при этом иметь принципиально различную внутреннюю каузальную структуру – то, что в IIT называют проводкой (*wiring*). Сознание определяется не тем, что система *делает*, а тем, как ее физические элементы *интегрированы* друг с другом каузально. Простой калькулятор и человек могут выдать одинаковый результат вычисления «2+2», однако каузальная сторона внутренних процессов в этих случаях несопоставима.

Кристоф Кох заключает, что архитектура современного цифрового компьютера физически неспособна обеспечить ту степень внутренней интеграции, которая необходима для феноменального опыта: его транзисторы слишком независимы друг от друга, чтобы образовать единое каузальное целое [Koch, 2019]. Следовательно, любой ИИ, реализованный на типовом компьютере, с этой точки зрения не будет обладать сознанием, даже если он способен пользоваться языком, как человек, рефлексировать, выражать эмоции или напрямую заявлять о наличии у себя сознания.

Если отвлечься от методологического удобства вычислительного функционализма и спросить именно о силе доводов в его пользу, то они выглядят заметно слабее, чем критика этой позиции. Ее сторонники пребывают в парадигме *компьютационализма*, который сводит или отождествляет сознание, как и мышление, с вычислением.



Кох назвал это «царствующей доктриной англосаксонской философии разума, департаментов компьютерных наук и индустрии технологий» [Koch, 2019][5]. То есть устойчивость этой позиции во многом держится не столько на аргументах, сколько на дисциплинарных конвенциях, исследовательской инерции, коллективных интуициях и фантазиях.

Сторонники IIT правы в своей критике вычислительного функционализма[6], однако упускают один существенный момент. Когда они говорят о сознании, они имеют в виду именно биологическое сознание. Создатели теории берут свойства биологического сознания и объявляют их обязательными для любого сознания (для них сознание – это конкретное и единственное естественное явление). Поэтому их критика по существу сводится к тезису: *биологическое сознание невозможно реализовать через вычисления*. В таком виде позиция кажется почти тривиальной.

Шаг, который позволяет уйти и от жесткого вывода IIT о невозможности сознания ИИ, и от слабых попыток навести мосты между человеческим и искусственным сознанием через вычислительный функционализм, состоит в другом допущении: *возможное сознание ИИ – это не биологическое сознание, а иной род сознания*. Это соображение опирается на общеизвестный факт: материал и архитектура ИИ кардинально отличаются от биологических систем, и он не разделяет с нами конституирующего фактора – эволюционной истории развития. *Следовательно, если машинное сознание возможно, оно практически с необходимостью будет сознанием другого рода*. Тогда критика IIT бьет мимо цели, а вычислительный функционализм просто перестает быть нужен, потому что вместе с ним отпадает и нужда в самих его мостах.

Из сказанного не следует, что сознание ИИ невозможно вообще. Однако если искусственное сознание возможно, то его нельзя заранее мыслить ни как функциональный аналог человеческого сознания, ни как его биологически точное воспроизведение на другом носителе. Исследование должно быть открыто тому, что возможное сознание ИИ окажется иной формой сознания, не совпадающей с человеческим по своей внутренней организации и потому плохо описываемой готовыми понятиями. Его изучение

[5] См. раздел “Consciousness and Computationalism”.

[6] Интересно, что IIT напрямую отвергает функционализм, но предполагает нейтральность к субстрату, т.е. множественную реализуемость. Требования к интеграции, выдвигаемые теорией, могут быть реализованы не только на биологическом субстрате, но и на любом другом. Это яркий пример, подтверждающий, что функционализм и множественная реализуемость суть две разные позиции, а их логические отношения неоднозначны. В литературе же обычно предполагается, что они необходимо идут вместе.

требует иных философских оснований и не может быть продуктивным, пока вычислительный функционализм определяет фон дискуссии по умолчанию.

От агностицизма к многомерным моделям сознания ИИ

По мере того как обнаруживаются жесткие ограничения сильного функционализма и сильного биологизма, вектор дискуссии смещается к более взвешенным позициям. Одной из них становится агностицизм, сторонники которого утверждают: существующие теории и эмпирические данные пока не позволяют вынести обоснованное решение в пользу признания или отрицания искусственного сознания [McClelland, 2024]. Дело в том, что наука о сознании построена на рассмотрении живых существ, из-за чего ее трансляция на ИИ упирается в непреодолимую недоопределенность. Биологический натуралист настаивает на необходимости органического субстрата, другие же принимают спорный метафизический тезис вычислительного функционализма. Доказательная база слаба у обеих сторон.

Этот вывод подтверждается скептическим анализом Эрика Швитцгебеля [Schwitzgebel, 2025]: весьма вероятно, что ИИ-системы ближайшего будущего будут признаны сознательными по одним влиятельным теориям и объявлены несознательными по другим. Поле исследования движется не к скорому консенсусу, а к ситуации, когда теоретические вердикты начнут непримиримо расходиться.

Именно поэтому сегодня наиболее перспективны не те подходы, что обещают окончательный ответ на вопрос о сознательности ИИ, а те, что принципиально избегают поспешной бинарности. Их общая стратегия – переход от грубого вердикта к многомерному исследованию: вместо вопроса о наличии сознания «вообще» – вопросы о том, какие свойства, уровни и типы внутренней организации присущи системе, какие из них могут быть релевантны для сознания и насколько правомерен сам перенос человеческих категорий на этот случай. Такие исследования тяготеют к теоретико-легким подходам, но настолько далеко отходят от исходной рамки Бёрча, что я выделяю их в отдельную группу.

Осторожный Дэвид Чалмерс, рассуждая о возможности сознания у больших языковых моделей, уходит от строгого «да или нет» [Chalmers, 2023]. Он предлагает *теоретико-сбалансированный* подход, где оценка строится на сопоставлении целого спектра условий: от наличия моделей мира и самого себя до рекуррентной обработки, глобального рабочего пространства, сенсорной укорененности



и единой агентности. Кроме того, Чалмерс фактически разделяет вопрос о сознании по нескольким измерениям, различая сенсорный, аффективный, когнитивный и агентивный аспекты феноменального опыта. Благодаря этому обсуждение перестает быть полностью бинарным: система может обладать одними предпосылками и не обладать другими. Современные языковые модели предстают здесь не просто бессознательными автоматами, а скорее архитектурно неполными кандидатами. Отсюда и осторожный вывод философа: хотя нынешние языковые модели, скорее всего, лишены сознания, будущие интегрированные системы (так называемые *LLM+*), объединяющие мультимодальность, воплощенность и агентность, уже нельзя будет так легко списать со счетов [Chalmers, 2023].

Еще более важный шаг делает Шевлин. В качестве альтернативы он выдвигает скромный теоретический подход (*modest theoretical approach*). Его суть – в удержании баланса между поведенческими маркерами и анализом внутренней архитектуры. Все начинается с выделения сильных кандидатов на наличие сознания на основе относительно очевидных индикаторов, а затем постепенно как бы пересобирается взятая теория сознания: отсеиваются те ее элементы, которые оказались лишь специфической особенностью человеческой биологии. Для исследований ИИ эта стратегия незаменима: она избавляет нас как от догматического навязывания машине человеческих лекал, так и от растворения проблемы в чрезмерно абстрактных функциональных формулах [Shevlin, 2021].

Наиболее разработанную форму многомерный и многоуровневый поворот получает у Катинки Эверс [Evers et al., 2024]. Здесь сознание предлагается понимать не как единое свойство, которое либо есть, либо нет, а как сложный профиль, состоящий из форм, компонентов, измерений и уровней. Авторы различают, с одной стороны, формы сознания, например феноменальное и доступное, с другой – компоненты вроде бодрствования и осознавания (*awareness*), а далее раскладывают их на отдельные измерения, функциональные и содержательные.

В результате появляется множество локальных уровней сознания по разным осям. Это смещает фокус с вопроса «сознательна ли система?» к многомерной оценке ее профиля по параметрам агентности, исполнительного контроля, управления вниманием, метакогниции, семантического понимания, сенсорного богатства. Такой подход особенно продуктивен в случае ИИ, поскольку он заранее допускает профиль, радикально расходящийся с человеческим, – например, высокое семантическое понимание при низкой сенсорной насыщенности или развитый контроль в отсутствие феноменальности в привычном смысле [Ibid.].

Именно в этом направлении и смещается наиболее зрелая часть современной дискуссии. Агностицизм здесь не является финальной

остановкой. Он играет роль дисциплинирующего перехода: сначала мы отказываемся от слишком жестких метафизических и теоретических вердиктов, а затем переходим к более осторожным многомерным и многоуровневым схемам. Такие схемы не решают проблему искусственного сознания окончательно и не устраняют трудность переноса понятий из науки о сознании.

Ценность многомерных подходов состоит в том, что они позволяют локализовать ее и разложить на отдельные вопросы. Они не обещают решающий тест, зато позволяют точнее работать с теми конкретными трудностями, которые обнаружились в предыдущих подходах. В отличие от теоретико-нейтральных стратегий, здесь не воспроизводится проблема отсутствия явной теории и слепого полагания на бихевиоральные данные. В отличие от теоретико-нагруженного подхода, удается избежать некритического переноса человеческих когнитивных механизмов на машину и связанного с этим принятия вычислительного функционализма. В отличие от теоретико-легких версий, такие модели не сводят исследование к одной центральной гипотезе, эмпирическая проверка которой опирается на нерефлексивное применение понятия феноменального опыта.

Выводы и перспективы

Современное исследование сознания ИИ зиждется на негласной предпосылке: терминологический аппарат, созданный для человека и животных, можно переносить на машины – обычно через вычислительную интерпретацию. Но сама обоснованность этого переноса систематически выносится за скобки.

Именно здесь проходит настоящая граница дискуссии. Спор идет не просто между оптимистами и скептиками. Он идет между теми, кто считает, что у нас уже есть достаточно специфический понятийный аппарат для постановки вопроса о сознании ИИ, и теми, кто полагает, что наш аппарат пока слишком сильно укоренен в специфике биологических систем. Литература последних лет все чаще подтверждает вторую линию. Даже авторы, наиболее открытые идее сознательного ИИ, вынуждены формулировать свои тезисы в условной, вероятностной форме – с обязательной экспликацией теоретических допущений.

Из рассмотренной литературы не следует, что сознание ИИ невозможно. Но из нее следует, что его нельзя надежно изучать так, будто речь идет о еще одном варианте человеческого или животного сознания. Поэтому наиболее сильный вывод состоит в смене исследовательской установки.



Такую установку можно описать в нескольких пунктах.

- Исходным должен быть не вопрос, имеет ли ИИ феноменальное сознание как у человека, а вопрос, в чем заключается специфика вычислительных аналогов сознательных процессов и какими средствами эту специфику возможно исследовать.
- Ни концепты науки о сознании (скажем, феноменальный опыт), ни ее теоретические модели не могут автоматически транслироваться на искусственные системы – каждый такой перенос требует самостоятельного обоснования применимости.
- Наличие теории всегда лучше ее отсутствия. Анализ не может игнорировать внутренние свойства системы и строиться исключительно на самоотчете ИИ.
- Доказывать истинность или ложность вычислительного функционализма не обязательно, как и не обязательно его принимать или отвергать.
- Необходимо допускать, что возможное сознание ИИ есть не сознание наподобие человеческого или животного, а иная форма сознания, плохо улавливаемая сегодняшним аппаратом изучения.

В предельной формулировке это означает, что исследование сознания ИИ должно начинаться почти с нулевой точки. Не в том смысле, что нужно отбросить всю науку о сознании человека, а в том смысле, что ни один ее ключевой концепт не должен автоматически сохранять силу при переходе к искусственным системам. *Наиболее продуктивная стратегия на данном этапе состоит в том, чтобы использовать существующие теории как временные эвристики, постоянно проверяя границы их применимости.* Это особенно выражено в многомерных и многоуровневых подходах.

Сознание – это конкретное естественное явление, которое мы обнаруживаем у человека и имеем веские основания искать у животных. Однако его ли мы ищем, когда исследуем ИИ? Мейнстримная позиция отвечает на этот вопрос утвердительно, используя для этого вычислительный функционализм. *Я же полагаю, что в случае ИИ мы имеем дело с поиском иного естественного явления*. Если учитывать богатство природы и фундаментальные различия между биологическими организмами и машинами, насколько вероятно, что сознание, возникающее у биологических существ, окажется эквивалентным возможному машинному сознанию? Очевидно, что такая вероятность крайне мала.

Благодарности

Благодарю Владислава Эриковича Тереховича за обсуждение и помощь в подготовке моей первой публикации.

Список литературы / References

Aru et al., 2023 – Aru, J., Larkum, M., and Shine, J. "The Feasibility of Artificial Consciousness Through the Lens of Neuroscience", *Trends in Neurosciences*, 2023. https://doi.org/10.48550/arXiv.2306.00915

Bayne et al., 2024 – Bayne, T., Seth, A., Massimini, M., et al. "Tests for Consciousness in Humans and Beyond", *Trends in Cognitive Sciences*, 2024, vol. 28 (5), pp. 454–466.

Birch, 2022 – Birch, J. "The Search for Invertebrate Consciousness", *Nous*, 2022, vol. 56 (1), pp. 133–153.

Block, 2026 – Block, N. "Can Only Meat Machines Be Conscious?", *Trends in Cognitive Sciences*, 2026, vol. 30 (4), pp. 298–308.

Browning, Veit, 2020 – Browning, H., and Veit, W. "The Measurement Problem of Consciousness", *Philosophical Topics*, 2020, vol. 48 (1), pp. 85–108.

Butlin, Long, Elmoznino et al., 2023 – Butlin, P., Long, R., Elmoznino, E., et al. "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness", 2023. https://doi.org/10.48550/arXiv.2308.08708

Butlin, Long, Bayne et al., 2025 – Butlin, P., Long, R., Bayne, T., et al. "Identifying Indicators of Consciousness in AI Systems", *Trends in Cognitive Sciences*, 2025. https://doi.org/10.1016/j.tics.2025.10.011

Chalmers, 1996 – Chalmers, D.J. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press, 1996.

Chalmers, 2023 – Chalmers, D.J. "Could a Large Language Model Be Conscious?", 2023. [https://doi.org/10.48550/arXiv.2303.07103, accessed on 10.04.2026]

Comsa, Shanahan, 2025 – Comsa, I.M., and Shanahan, M. "Does It Make Sense to Speak of Introspection in Large Language Models?", 2025. [https://doi.org/10.48550/arXiv.2506.05068, accessed on 10.04.2026]

Dossa et al., 2024 – Dossa, R.F.J., Arulkumaran, K., Juliani, A., et al. "Design and Evaluation of a Global Workspace Agent Embodied in a Realistic Multimodal Environment", *Frontiers in Computational Neuroscience*, 2024, vol. 18, 1352685.

Dung, 2025 – Dung, L. "Tests of Animal Consciousness Are Tests of Machine Consciousness", *Erkenntnis*, 2025, vol. 90 (4), pp. 1323–1342.

Evers et al., 2025 – Evers, K., Farisco, M., Chatila, R., et al. "Preliminaries to Artificial Consciousness: A Multidimensional Heuristic Approach", *Physics of Life Reviews*, 2025, vol. 52, pp. 180–193.

Findlay et al., 2024 – Findlay, G., Marshall, W., Albantakis, L., et al. "Dissociating Artificial Intelligence from Artificial Consciousness", 2024. [https://doi.org/10.48550/arXiv.2412.04571, accessed on 10.04.2026]

Goldstein, Kirk-Giannini, 2024 – Goldstein, S., and Kirk-Giannini, C.D. "A Case for AI Consciousness: Language Agents and Global Workspace Theory", 2024. [https://doi.org/10.48550/arXiv.2410.11407, accessed on 10.04.2026]

Graziano, 2017 – Graziano, M.S.A. "The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness", *Frontiers in Robotics and AI*, 2017, vol. 4, 60.

Juliani et al., 2022 – Juliani, A., Arulkumaran, K., Sasai, S., and Kanai, R. "On the Link Between Conscious Function and General Intelligence in Humans and

Machines", *Transactions on Machine Learning Research*, 2022.[https://doi.org/10.48550/arXiv.2204.05133, accessed on 10.04.2026]

Juliani et al., 2021 – Juliani, A., Kanai, R., and Sasai, S. "The Perceiver Architecture Is a Functional Global Workspace", *Annual Meeting of the Cognitive Science Society*, 2021.

Koch, 2019 – Koch, C. *The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed*. Cambridge, MA: MIT Press, 2019.

Lillicrap et al., 2020 – Lillicrap, T.P., Santoro, A., Marris, L., et al. "Backpropagation and the Brain", *Nature Reviews Neuroscience*, 2020, vol. 21 (6), pp. 335–346.

McClelland, 2024 – McClelland, T. "Agnosticism About Artificial Consciousness", 2024. [https://doi.org/10.48550/arXiv.2412.13145, accessed on 10.04.2026]

Pfeiffer, Pfeil, 2018 – Pfeiffer, M., and Pfeil, T. "Deep Learning with Spiking Neurons: Opportunities and Challenges", *Frontiers in Neuroscience*, 2018, vol. 12, 774.

Putnam, 1960 – Putnam, H. "Minds and Machines", in: Hook, S. (ed.) *Dimensions of Mind: A Symposium*. New York: New York University Press, 1960, pp. 138–164.

Putnam, 1967 – Putnam, H. "Psychological Predicates", in: Capitan, W.H., and Merrill, D.D. (eds.) *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press, 1967, pp. 37–48.

Schneider, 2019 – Schneider, S. *Artificial You: AI and the Future of Your Mind*. Princeton, NJ: Princeton University Press, 2019.

Schneider, 2020 – Schneider, S. "How to Catch an AI Zombie: Testing for Consciousness in Machines", in: Liao, S.M. (ed.) *Ethics of Artificial Intelligence*. New York: Oxford University Press, 2020, pp. 439–458.

Schwitzgebel, 2025 – Schwitzgebel, E. "AI and Consciousness", 2025. [https://doi.org/10.48550/arXiv.2510.09858, accessed on 10.04.2026]

Seth, 2025 – Seth, A.K. "Conscious Artificial Intelligence and Biological Naturalism", *Behavioral and Brain Sciences*, 2025, pp. 1–42.

Shevlin, 2021 – Shevlin, H. "Non-Human Consciousness and the Specificity Problem: A Modest Theoretical Proposal", *Mind & Language*, 2021, vol. 36 (2), pp. 297–314.

Udell, Schwitzgebel, 2021 – Udell, D.B., and Schwitzgebel, E. "Susan Schneider's Proposed Tests for AI Consciousness: Promising but Flawed", *Journal of Consciousness Studies*, 2021, vol. 28 (5–6), pp. 121–144.

VanRullen, Kanai, 2021 – VanRullen, R., and Kanai, R. "Deep Learning and the Global Workspace Theory", *Trends in Neurosciences*, 2021, vol. 44 (9), pp. 692–704.

Whittington, Bogacz, 2019 – Whittington, J.C.R., and Bogacz, R. "Theories of Error Back-Propagation in the Brain", *Trends in Cognitive Sciences*, 2019, vol. 23 (3), pp. 235–250.

Wilterson, Graziano, 2021 – Wilterson, A.I., and Graziano, M.S.A. "The Attention Schema Theory in a Neural Network Agent: Controlling Visuospatial Attention Using a Descriptive Model of Attention", *Proceedings of the National Academy of Sciences of the United States of America*, 2021, vol. 118 (33), e2102421118.