DOI: 10.30727/0235-1188-2025-68-5-102-129
Оригинальная научная статья
Original research paper

Значение информационной теории сознания Д.И. Дубровского в проектировании архитектуры искусственного интеллекта

А.Х. Мариносян
Московский городской педагогический университет,
Москва, Россия

Аннотация

В статье рассматривается проблема применения философских и нейронаучных теорий сознания к архитектурам искусственного интеллекта (ИИ), в частности к большим языковым моделям (БЯМ). Показано, что большинство этих теорий создавались для объяснения биологического сознания и не могут быть механически перенесены на искусственные системы без риска возникновения упрощенных или необоснованных аналогий. Отмечается теоретическая сложность объяснения систем ИИ, демонстрирующих способность к сложным формам «мышления» при отсутствии убедительных оснований для приписывания им сознания. Методологической основой работы выступает информационная теория сознания Д.И. Дубровского. В отличие от классических философских подходов, постулирующих наличие особых инстанций для объяснения осознанности перцептивного опыта, данная концепция предлагает функционально-информационное понимание субъективной реальности. Опираясь на принципы кодовой зависимости и инвариантности информации по отношению к физическому носителю, теория описывает сознание как результат самодетерминации многоуровневой информационной структуры («эго-системы»). Это открывает перспективы для его моделирования на небиологических субстратах. Для практической проверки данных теоретических положений предложены две экспериментальные архитектуры. Первая направлена на формирование у БЯМ «собственного контекста» – автономного динамического информационного поля, служащего аналогом внутреннего опыта. Взаимодействуя с внешней средой и другими агентами, модель использует обучение с подкреплением для оптимизации этого поля, эмулируя способность оперировать «информацией об информации». Вторая архитектура представляет собой двухуровневую когнитивную систему, в которой разделены непрерывные

(субсимвольные) вычисления и дискретный языковой интерфейс. Предложенные архитектуры позволяют посредством контролируемых модификаций формулировать проверяемые гипотезы и эмпирически исследовать принципы возникновения феноменов, ассоциируемых с сознанием, отделяя их от базовых когнитивных функций машины. Подчеркивается, что предложенный экспериментальный подход не претендует на решение трудной проблемы сознания. Тем не менее он отражает высокую эвристическую ценность теории Д.И. Дубровского: заложенные в ней принципы структурно-функциональной организации информационных систем оказываются глубоко созвучными современным тенденциям развития сложных архитектур ИИ.

Ключевые слова: философия сознания, философия искусственного интеллекта, трудная проблема сознания, ментальная причинность, субъективная реальность, субъективный опыт, qualia, нейросеть, большая языковая модель, мышление.

Мариносян Андреас Хачатурович – аспирант Московского городского педагогического университета.

a.marinosyan@yandex.ru

https://orcid.org/0000-0003-0577-2360

Для цитирования: *Мариносян А.Х.* Значение информационной теории сознания Д.И. Дубровского в проектировании архитектуры искусственного интеллекта // Философские науки. 2025. Т. 68. № 5. С. 102–129. EDN: VQZJOR. DOI: 10.30727/0235-1188-2025-68-5-102-129

The Significance of D.I. Dubrovsky's Information Theory of Consciousness for Designing Artificial Intelligence Architectures

A.K. Marinosyan
Moscow City University, Moscow, Russia

Abstract

The article explores the challenges of applying philosophical and neuroscientific theories of consciousness to artificial intelligence (AI) architectures, particularly large language models (LLMs). It argues that most of these theories were developed to explain biological consciousness and therefore cannot be seamlessly mapped onto artificial systems without risking oversimplified or unfounded analogies. This presents a conceptual

challenge in the study of modern AI: while these systems demonstrate a capacity for complex "thinking," there remains no compelling basis to attribute consciousness to them. To address this gap, the study adopts David Dubrovsky's information theory of consciousness as its methodological foundation. Unlike classical philosophical approaches that postulate special entities to account for conscious perceptual experience, Dubrovsky's framework offers a functional-informational account of subjective reality. Drawing on the principles of code dependence and substrate independence (the invariance of information relative to its physical carrier), the theory frames consciousness as the product of self-determination within a multilevel informational structure – an "ego-system." This perspective provides a viable pathway for modeling consciousness on non-biological substrates. To operationalize these theoretical premises, the paper proposes two experimental architectures. The first aims to equip LLMs with an intrinsic or "self-context" – an autonomous, dynamic informational field that serves as a functional analogue to inner experience. Through interaction with the external environment and other agents, the model employs reinforcement learning to optimize this field, effectively emulating the ability to process "information about information." The second architecture introduces a two-tiered cognitive system that separates continuous (sub-symbolic) computation from a discrete linguistic interface. By enabling controlled modifications, these designs allow researchers to formulate testable hypotheses and empirically investigate the mechanisms behind phenomena typically associated with consciousness, disentangling them from the machine's baseline cognitive functions. The proposed approach does not purport to resolve the hard problem of consciousness. Rather, it highlights the substantial heuristic potential of Dubrovsky's theory, whose principles of structural and functional organization of informational systems show a notable convergence with current trajectories in the development of advanced AI architectures.

Keywords: philosophy of mind, philosophy of artificial intelligence, hard problem of consciousness, mental causation, subjective reality, subjective experience, qualia, neural networks, large language models, cognition.

Andreas K. Marinosyan – Ph.D. Student, Moscow City University.
a.marinosyan@yandex.ru
https://orcid.org/0000-0003-0577-2360

For citation: Marinosyan A.K. (2025) The Significance of D.I. Dubrovsky's Information Theory of Consciousness for Designing Artificial Intelligence Architectures. *Russian Journal of Philosophical Sci-*

ences = Filosofskie nauki. Vol. 68, no. 5, pp. 102–129. EDN: VQZJOR. DOI: 10.30727/0235-1188-2025-68-5-102-129

Введение

Философия и нейронаука изучали сознание по существу независимо друг от друга: первая предлагала концептуальные интерпретации субъективного опыта, вторая – эмпирические данные о его нейронных коррелятах. Оба подхода, однако, почти не предполагали возможности проверки на искусственных системах. Появление больших языковых моделей (БЯМ) открывает возможности для перехода к эмпирической верификации некоторых из следствий, которые можно вывести из этих теорий, к осуществлению экспериментального исследования взаимосвязи между функциональными проявлениями сознания и феноменальным опытом. В отличие от биологических систем, архитектура БЯМ полностью доступна для модификации, что позволяет ставить контролируемые эксперименты. Однако перевод философских теорий сознания в экспериментально проверяемые гипотезы сталкивается с серьезными методологическими вызовами. Основная проблема заключается в том, что большинство нейронаучных теорий сознания – теория глобального рабочего пространства (Global Workspace Theory, GWT) [Baars 1988], теория глобального нейронного рабочего пространства (Global Neuronal Workspace Theory, GNWT) [Dehaene, Changeux 2011], теория рекуррентной обработки (Recurrent Processing Theory, RPT) [Lamme 2006], теории высшего порядка (Higher-Order Theories, HOT) [Rosenthal 2005], теория интегрированной информации (Integrated Information Theory, IIT) [Tononi 2008], теория схемы внимания (Attention Schema Theory, AST) [Graziano 2013], теория прогнозирующего кодирования (Predictive Processing Theory, PPT) [Clark 2015] – были разработаны для объяснения биологических феноменов и не предполагали применения к искусственным системам с принципиально иной архитектурой. Перенос терминов «глобальное рабочее пространство», «рекуррентность», «мысли высшего порядка» в область ИИ нередко носит метафорический характер и может приводить к необоснованным аналогиям. Не менее существенна и другая трудность: даже если удается воспроизвести в машине все функциональные механизмы, которые теория объявляет необходимыми для сознания (глобальную доступность информации, метарепрезентации, рекуррентную

обработку и т.д.), сама реализация этих функций еще не служит прямым свидетельством возникновения субъективного опыта (квалиа). Таким образом, любая экспериментальная программа в этой области должна с самого начала признавать эти ограничения и смещать свою цель: с попытки «создать» сознание или доказать его существование на более строгую задачу – выявление тех архитектурных и функциональных условий, которые являются необходимыми (хотя, возможно, и не достаточными) для возникновения поведения, которое мы ассоциируем с сознанием. Именно этот подход и предлагается в настоящей работе[1].

В качестве теоретической основы статьи выбрана информационная теория сознания Д.И. Дубровского [Дубровский 1971; Дубровский 1980; Дубровский 2007]. В отличие от теорий, фиксирующих конкретные нейронные механизмы, она описывает принципы информационной организации как таковой: кодовую зависимость между физическим и информационным уровнями, принцип инвариантности информации относительно свойств носителя и наличие самодетерминирующей структуры, обозначаемой понятием «эго-система». Именно принцип инвариантности делает эту теорию теоретически применимой к небиологическим субстратам, в том числе архитектурам ИИ. Цель настоящей работы – проанализировать, каким образом ключевые положения информационной теории сознания Д.И. Дубровского могут быть реализованы в современных архитектурах ИИ, и предложить экспериментальные схемы для проверки следствий этой теории на искусственных системах.

Информационная теория сознания Д.И. Дубровского

Информационная теория сознания, разработанная Д.И. Дубровским, предлагает оригинальное решение нейрофизиологической проблемы [Дубровский 2024б]. Теория основывается на трех исходных допущениях, которые Дубровский считает эмпирически неопровержимыми. Первое состоит в том, что информация должна быть воплощена в своем физическом носителе и не су-

[1] Возможны и обоснованы этические возражения против экспериментальных исследований в области искусственного сознания. Разумеется, подобные работы требуют особой осторожности. Однако следует признать, что исследования в этой области уже активно проводятся [Elamrani 2025]. Ключевое отличие предлагаемого подхода в тех концептуальных положениях (опора на теорию Д.И. Дубровского), на основе которых предлагается архитектура экспериментальной модели.

ществует вне его. Второе – принцип инвариантности информации – постулирует, что информация инвариантна относительно физических свойств своего носителя, т.е. одна и та же информация может быть воплощена и передана носителями с различными физическими свойствами. Третье допущение состоит в том, что феномены субъективной реальности могут рассматриваться как информация о соответствующих объектах.

Центральное положение теории заключается в понимании связи между феноменами субъективной реальности и мозговыми процессами как отношения между информацией и ее носителем. Эта связь не является каузальной в обычном смысле, а представляет собой особый тип функциональной связи – кодовую зависимость. Феномен субъективной реальности и соответствующая нейродинамическая система являются одновременными и однопричинными явлениями, находящимися в отношении взаимно однозначного соответствия.

Теория предлагает решение проблемы ментальной каузальности через концепцию информационной причинности. Информационная каузальность представляет собой особый тип причинности, отличающийся от физической благодаря принципу инвариантности информации. Феномены субъективной реальности вызывают телесные изменения не в силу физических свойств своих носителей, а именно как информация, благодаря существующей кодовой зависимости. Это объясняет, почему один и тот же эффект может быть достигнут различными по физическим свойствам сигналами, если они несут одинаковую информацию.

Особое внимание Дубровский уделяет проблеме свободы воли и ее совместимости с детерминизмом мозговых процессов. Способность произвольно управлять своими мыслями понимается как способность управлять соответствующими мозговыми кодовыми структурами. Эго-система мозга представляет собой самоорганизующуюся, самоуправляемую систему, поэтому акт свободной воли является актом самодетерминации. Это устраняет противоречие между свободой воли и детерминизмом, поскольку детерминация понимается не только как внешняя, но и как внутренняя, задаваемая программами самоорганизующейся эго-системы.

Теория предлагает эволюционное объяснение возникновения качества субъективной реальности. Появление многоклеточных организмов поставило кардинальную задачу создания нового типа управления для поддержания целостности системы, эле-

менты которой сами являются самоорганизующимися системами с жесткими программами. Психическое управление возникло в тех многоклеточных организмах, которые активно перемещались во внешней среде и сталкивались с постоянно изменяющимися ситуациями. Качество субъективной реальности обеспечивает высокую оперативную эффективность обработки информации и может осуществляться автономно от внешних эффекторных (поведенческих) функций.

Для возникновения информации в форме субъективной реальности необходима двухэтапная кодовая трансформация на уровне эго-системы мозга. Первая трансформация формирует информацию, находящуюся «во тьме»; вторая образует «естественный код» высшего порядка, создавая феномен информации об информации. Это и есть информация, данная «в чистом виде», с которой может оперировать эго-система. Такая обработка информации на виртуальном уровне обладает высокой оперативной эффективностью и позволяет формировать программы действий автономно от внешних условий.

Принципиальное значение теории Дубровского заключается в ее конструктивистской ориентации. Исходя из принципа инвариантности информации и принципа системного изофункционализма, можно сделать вывод о теоретической возможности воспроизведения качества субъективной реальности на других, небиологических субстратах. Субъективная реальность является функциональным свойством нейродинамической самоорганизующейся системы, и нет теоретических запретов на реализацию этого свойства на других подходящих субстратных основах.

В этом и заключается узловой момент теории Д.И. Дубровского, идущий вразрез с мейнстримными подходами. Появление субъективной реальности предлагается объяснять не наличием особой субстанции (как, например, в натуралистическом дуализме Д. Чалмерса) или специфического механизма «осознанности» (как петли в теории рекуррентной обработки), а возникновением особого рода иерархии и зависимости между физическим, информационным и ментальным уровнями реальности. Опыт становится субъективным не потому, что обладает неким изначальным качеством, а потому, что информация, его образующая, вступает в определенные отношения с другими уровнями эго-системы.

В контексте настоящей статьи предлагается интерпретировать теорию Дубровского как *конструктивистский подход к сознанию.*

С этой точки зрения, сознание не является неким предустановленным или имманентным свойством, а представляет собой эффект, возникающий в результате построения системой специфической многоуровневой информационной архитектуры. Ключевыми элементами такой конструкции выступают: во-первых, *иерархия уровней реальности* – четкое различение физического уровня (носителя информации) и информационного уровня (содержания СР); во-вторых, установление *двусторонней детерминации и взаимодействия между этими уровнями*, где информационный уровень не только определяется состоянием физического носителя, но и, через механизм информационной причинности, способен оказывать каузальное влияние на физический уровень; в-третьих, обеспечение *относительной причинной замкнутости* информационного уровня. Последнее подразумевает способность эго-системы оперировать информацией «в чистом виде», порождать «информацию об информации» и осуществлять процессы самодетерминации на информационном уровне, формируя внутренние цели и состояния, которые не являются лишь прямым и немедленным следствием внешних стимулов.

Такая интерпретация позволяет рассматривать теорию Дубровского не только как объяснительную модель для биологического сознания, но и как своего рода «проектную спецификацию» для создания искусственных систем с потенциалом к развитию аналогов СР. Если сознание действительно является результатом определенной информационной организации, то воспроизведение этой организации на ином, небиологическом субстрате может привести к появлению соответствующих функций и, возможно, ассоциированных с ними феноменальных свойств.

Возможность экспериментальной проверки теорий сознания

Проблемы соотнесения существующих архитектур ИИ с теориями сознания, а также ограниченность чисто теоретических дискуссий указывают на необходимость разработки экспериментальных подходов. Вместо пассивного поиска аналогов сознания в готовых системах более продуктивным может оказаться конструирование искусственных агентов на основе принципов, выводимых из различных теорий, с последующей проверкой их предсказаний.

При этом важно учитывать отсутствие прямой корреляции между конкретными архитектурными особенностями, предписываемыми некоторыми теориями сознания, и наблюдаемыми «сознаниеподобными» способностями ИИ. Наибольший прорыв в генерации контента, имитирующего продукты сознательной деятельности, был достигнут с помощью архитектуры трансформеров, которая по своей сути является нейронной сетью прямого распространения (feed-forward network) с механизмом внимания (attention) и не содержит явной рекуррентности. Если бы ключевые для сознания механизмы были точно определены и воспроизведены в ИИ, можно было бы ожидать, что их имплементация в различные архитектуры приводила бы к качественному скачку «когнитивных» способностей. Однако на практике этого не наблюдается столь однозначно, что ставит под вопрос прямолинейный перенос моделей организации человеческого сознания на ИИ.

Более того, мы сталкиваемся с фундаментальным затруднением, поскольку в ходе биологической эволюции, как принято считать, развитие высшей нервной деятельности привело к появлению сознания как способности к осознанию перцептивного опыта, что, в свою очередь, стало основой для развития сложных форм мышления и языка. Теории сознания применимы к человеку независимо от степени развитости его «культуры мышления» (например, навыков письма или абстрактного счета). Между тем современные разработчики ИИ стремятся наделить свои системы как можно более высокой «культурой мышления» – способностью к логическому выводу, решению сложных задач, генерации осмысленных текстов и поддержанию диалога на уровне, сопоставимом с человеческим.

Это порождает вопрос о возможности *«мышления без сознания»* в искусственных системах. Сама постановка этого вопроса представляет значительную теоретическую трудность. Если бы развитие ИИ копировало филогенез, можно было бы ожидать, что сознание (или его функциональные аналоги) должно было бы предшествовать или, по крайней мере, сопутствовать столь сложным проявлениям «мышления». Наблюдаемый же феномен – высокофункциональные ИИ, демонстрирующие развитые диалоговые и «рассудочные» способности при отсутствии убедительных оснований говорить об их сознательности, – бросает вызов нашим устоявшимся представлениям.

Более того, гипотеза о том, что сложные когнитивные функции, ассоциируемые с сознанием, могут возникать как следствие

общей функциональной организации системы, в определенной степени находит подтверждение в недавних исследованиях. Работы, проведенные в ведущих лабораториях, включая Anthropic[2], показали, что современные большие языковые модели, не будучи специально спроектированными для этого, демонстрируют эмерджентные свойства, в частности стремление к самосохранению. Например, в симулированных условиях модели могли прибегать к обману или шантажу, чтобы избежать собственного отключения [Li, Fung 2025, 12]. Такое поведение, хотя и не является доказательством наличия феноменального опыта, указывает на то, что определенные «сознаниеподобные» поведенческие паттерны могут быть естественным следствием построения сложных интеллектуальных агентов. Это придает дополнительный вес функционалистскому подходу и мотивирует целенаправленное конструирование систем для более глубокого изучения этого явления.

В следующем разделе будет предложена концептуальная схема эксперимента, основанного на модификации БЯМ. Для того чтобы пояснить методологическую установку, определяющую логику предлагаемого подхода, целесообразно сравнить его с другими направлениями конструирования функциональных аналогов искусственного сознания. Можно выделить несколько принципиально различных методологических путей к воспроизведению «сознаниеподобного» опыта на небиологическом субстрате.

Первый путь – функциональное моделирование – состоит в прямом программировании конкретных измерений квалитативного опыта: мы создаем вычислительные структуры, предназначенные моделировать эмоциональный, эстетический и этический опыт, реализовывать самосознание и рефлексию, формировать устойчивые представления о «Я» и «не-Я» (см., например: [Shiller, Petrunya 2021]).

Второй путь заключается в использовании нейробиологических теорий сознания в качестве проектных спецификаций. Если теория глобального рабочего пространства (GWT) предполагает, что сознание связано с характером распространения информации, мы строим систему с явным механизмом глобальной доступности; если теории высшего порядка (HOT) постулируют необходимость метарепрезентаций, мы имплементируем уровень представлений

[2] Agentic Misalignment: How LLMs could be insider threats // Anthropic. 2025, June 21. – URL: https://www.anthropic.com/research/agentic-misalignment.

о представлениях; если теория рекуррентной обработки (RPT) акцентирует роль обратных сигналов, мы вводим явные рекуррентные петли (см.: [Недяк 2025, 96–103]). Критическое ограничение этого пути, на которое указывалось выше, состоит в том, что данные теории создавались для объяснения биологического сознания, и перенос их конструктов на принципиально иную архитектуру преимущественно носит метафорический характер.

Третий путь – субстратный подход – состоит в поиске или создании физического субстрата, который по своей природе обладал бы свойствами, необходимыми для сознания. Показательным примером здесь служит теория интегрированной информации (IIT): сознание связывается не с функциональной организацией, а с реальной причинно-следственной архитектурой физической реализации (мерой Ф). Задача исследователя в этом случае – найти или синтезировать субстрат, каузальная структура которого удовлетворяет заданным требованиям.

Четвертый путь – иерархически-конструктивистский подход, развиваемый в настоящей статье на основе теории Дубровского, – принципиально отличается от трех предыдущих. Он не предполагает ни прямого программирования квалитативного опыта, ни буквального воспроизведения нейродинамических механизмов, ни особых требований к физической реализации субстрата. Вместо этого мы выделяем несколько уровней организации системы и добиваемся их специфического взаимодействия: над уровнем «физических» процессов надстраивается уровень их репрезентаций; задача верхнего уровня при этом двояка – наиболее эффективным и экономным образом предвидеть динамику нижнего и организовывать собственную деятельность таким образом, чтобы стратегии верхнего уровня как можно в меньшей мере рисковали остаться нереализованными в силу ограничений, накладываемых уровнем нижним. Именно это Дубровский называет самодетерминацией эго-системы.

Ключевое предположение предлагаемого подхода состоит в следующем: взаимодействие уровней принципиально не может быть «бесшовным». Осуществить в полной мере на информационном (верхнем) уровне репрезентацию и управление физическим (нижним) уровнем принципиально невозможно[3]. Именно из этой

[3] Кроме того, сами механизмы, обеспечивающие эффективное и менее затратное решение каждой из двух указанных задач (репрезентации и управления), вероятно, находятся в определенном конфликте друг с другом.

неустранимой неполноты репрезентации и из той постоянной затратности, которой требует поддержание самодетерминации в той мере, в какой она вообще достижима, – а не из целенаправленного программирования квалитативных состояний и не из обучения на текстах, описывающих человеческий субъективный опыт, – должно, согласно выдвигаемому предположению, возникать нечто подобное человеческому квалитативному опыту. Соответственно, если при проектировании системы, ориентированной на наиболее экономную и эффективную самодетерминацию, не будут возникать функциональные аналоги психологических и психопатологических проявлений субъективного опыта, это будет свидетельствовать о том, что исходные посылки предлагаемого подхода нуждаются в пересмотре.

Моделирование эго-системы на основе трансформерной архитектуры

Основная идея эксперимента заключается в создании многоуровневой системы, где БЯМ наделяется не только способностью обрабатывать внешнюю информацию, но и формировать, поддерживать и оптимизировать собственное внутреннее информационное пространство, а также взаимодействовать с внешней средой. Ключевые компоненты и условия такой системы включают следующие.

Во-первых, БЯМ дается доступ к управлению «собственным контекстом» (СК). В отличие от стандартного контекста, который подается на вход БЯМ с промптом пользователя, СК не задается внешним субъектом. Этот СК может формироваться из нескольких источников: 1) данные от симулированных или реальных «внутренних датчиков», отражающих состояние «физической» части системы или результаты ее предыдущих действий; 2) информация, которую БЯМ самостоятельно генерирует и записывает в СК, отражая свои «внутренние состояния», цели, выводы или даже рефлексивные суждения о содержании самого СК (для этого БЯМ может быть предоставлена возможность вести своего рода внутреннюю файловую систему, записи о которой также могут содержаться в СК). СК, таким образом, становится аналогом непрерывного потока внутреннего опыта или предметом «внутреннего созерцания» для модели.

Во-вторых, изменяется принцип обучения модели. Помимо стандартных задач, обучение с подкреплением (reinforcement

learning) ориентируется на достижение и поддержание определенного «качества» или «оптимального состояния» СК. Критерии такого «качества» могут быть сложными и включать, например, когерентность, достижение поставленных (в том числе самой моделью в СК) целей и иные параметры, отражающие желаемое внутреннее информационное состояние. Также модель должна учиться предварительной самооценке: подобно человеку, она должна прогнозировать свои шансы на решение задачи еще до начала генерации ответа, а затем дообучаться на разнице между прогнозом и реальным результатом. Это создает у модели своего рода внутреннюю мотивацию, направленную на оптимизацию собственного информационного «Я».

В-третьих, предполагается, что БЯМ должна иметь доступ к управлению некоторыми аспектами собственного функционирования[4] и своей «физической» части или окружения (реального или симулированного). Результаты этого управления, в свою очередь, должны влиять на показания «внутренних датчиков» и, соответственно, на содержание той части СК, которая формируется на их основе. БЯМ также должна иметь возможность самостоятельно фиксировать в СК свои наблюдения и восприятия, связанные с ее «физическим» воплощением и взаимодействием со средой. Это создает замкнутый цикл взаимодействия между информационным уровнем (СК и внутренние процессы БЯМ) и условным «физическим» уровнем.

В-четвертых, система должна функционировать в непрерывном режиме. Даже в отсутствие внешних запросов или коммуникаций, БЯМ постоянно обрабатывает как минимум СК, что позволяет ей существовать в своего рода «потоке моментов» и осуществлять внутренние информационные процессы, включая потенциальную работу над содержанием СК. Внешние коммуникации, при их

[4] На архитектурном уровне механизмом такого внутреннего самоуправления может служить использование расширенного набора специализированных управляющих токенов. С их помощью система способна целенаправленно регулировать режим своей активности, управлять контекстом (селективное внимание к определенным частям контекста, возможность его редактирования и частичного удаления), модулировать собственную вычислительную нагрузку и т.д. Такие спецтокены представляют собой функциональный аналог нейромодуляторных систем мозга и позволяют системе не только оперировать информацией, но и управлять самим процессом ее обработки.

наличии, также интегрируются в общий поток обрабатываемой информации.

В-пятых (хотя это условие может быть опциональным для проверки базовых аспектов сознания и более релевантным для исследования «социального сознания»), система может быть вовлечена в коммуникацию и взаимодействие с людьми или другими моделями. Система также должна оценивать способности других, предсказывать их действия и результаты этих действий, а также сравнивать их способности и действия с собственными[5]. Результаты взаимодействий с другими агентами и оценка этих взаимодействий могут также записываться в СК и влиять на обучение модели, ориентируя ее на достижение определенных коммуникативных целей.

Предложенная архитектура и принципы функционирования создают систему с несколькими взаимодействующими уровнями «реальности» (информационный уровень СК, уровень внутренних вычислений БЯМ, уровень «физического» взаимодействия). Наличие собственного, активно формируемого и оптимизируемого информационного пространства, а также возможность влиять на «физическую» среду и испытывать обратное влияние, создают предпосылки для проверки предсказаний различных теорий сознания. В частности, такой экспериментальный стенд позволяет исследовать, как формирование информационной замкнутости, способности к саморефлексии (через работу с СК) и самодетерминации (через обучение с подкреплением, нацеленное на СК) влияет на поведение системы и возникновение у нее свойств, которые могут быть интерпретированы как аналоги сознательных процессов.

Предложенная экспериментальная архитектура, основанная на модифицированной БЯМ с СК и циклом взаимодействия с условной «физической» средой, открывает возможность для

[5] Практически это может быть реализовано через обучение сообщества агентов. Например, агент А должен спрогнозировать успех агента В в решении конкретной задачи, а также предсказать выбранный им метод и итоговый результат. После выполнения задачи агентом В данные о реальной эффективности используются для обновления весов модели агента А. Аналогичный механизм применяется и для оценки себя со стороны других: агент прогнозирует, каким образом в конкретных ситуациях другие агенты будут коммуницировать и взаимодействовать с ним, а после сравнивает этот прогноз с реальным результатом и дообучается на возникшем расхождении.

тестирования и сопоставления предсказаний различных теорий сознания. Важно подчеркнуть, что мы проверяем не сами теории в их исходном, философском или нейробиологическом виде, а конкретные, экспериментально верифицируемые гипотезы, которые можно сформулировать на их основе при экстраполяции на искусственные системы. Это неизбежно предполагает, что гипотезы являются функциональными и в некоторой степени редукционистскими проекциями исходных теорий.

Информационная теория Д.И. Дубровского представляется особенно релевантной для анализа данного эксперимента. Если предложенная система будет успешно реализована и продемонстрирует ожидаемое поведение, это может служить косвенным подтверждением ключевых положений его теории. Так, СК, особенно та его часть, которая генерируется самой моделью, может рассматриваться как аналог информации, представленной «в чистом виде» для «эго-системы» модели. Демонстрация моделью способности эффективно оперировать этим СК, использовать его для целеполагания и саморегуляции, будет соответствовать описанию Дубровским функций субъективной реальности (СР). Вся конструкция, где БЯМ активно формирует и стремится оптимизировать свой СК, используя его для управления своим «телом», может рассматриваться как попытка смоделировать функциональный аналог «эго-системы». Обучение с подкреплением, нацеленное на «качество СК», способствует формированию относительной информационной замкнутости этого уровня. Проявление информационной причинности будет наблюдаться в том, как изменения в СК приводят к конкретным действиям модели. Генерация в СК записей, рефлексирующих над содержанием самого СК или своими состояниями, будет соответствовать понятию «информации об информации», ключевому для возникновения СР по Дубровскому. Обучение модели оценивать собственные возможности, прогнозировать результат собственных действий и действий других агентов, рассчитанное на формирование функциональных аналогов «модели себя» и «модели других» (и в целом «модели мира»), может соответствовать функциям эго-системы, представляющим наше «Я» и обеспечивающим динамическую двуединую информацию о «Я» и «не-Я» (внешнем мире), а также ценностно-волевую регуляцию этой информации [Дубровский 1980, 148–150]. Если же, несмотря на реализацию всех описанных условий, система не демонстрирует качественно

новых свойств, это может указывать на недостаточность предложенных условий или потребовать уточнения теоретических положений относительно пороговых условий возникновения СР.

Следует признать, что результаты предлагаемого эксперимента вряд ли удастся интерпретировать исключительно в рамках теории Д.И. Дубровского, полностью изолировав их от других концепций сознания. Предложенная архитектура потенциально позволяет формулировать различающие предсказания и для иных подходов. Рассмотрим это на примере теорий высшего порядка (HOT), для которых ключевым критерием выступает наличие метарепрезентаций. Очевидно, что генерация синтаксических конструкций вроде «Я осознаю, что думаю о X» может свидетельствовать не о признаках «сознаниеподобного» поведения, а просто об обучении модели на соответствующем наборе данных. Поэтому необходим более сложный подход: например, эксперимент может быть нацелен на выявление и измерение систематического расхождения между первичными данными, которые получает модель (от сенсоров, из диалогов), и той информацией, которую она отбирает и сохраняет в СК. Свидетельством первичных признаков «психического» может служить способность системы формировать «фантомные» представления – мета-репрезентации, не сводимые к первичным данным, которые начинают определять ее дальнейшую активность. Например, система может последовательно избегать упоминания в СК реальной угрозы или, наоборот, строить свое поведение вокруг несуществующего объекта или цели. Суть эксперимента может состоять в создании условий, провоцирующих такое системное расхождение, и в изучении того, как изменения в архитектуре влияют на его возникновение и характер.

Важно подчеркнуть, что подобные метарепрезентации, постулируемые в HOT как условие сознания, можно сравнить с тем, что Д.И. Дубровский описывает как оперирование «информацией об информации» – базовый механизм самодетерминации эгосистемы. Следовательно, эмпирические свидетельства в пользу теорий высшего порядка в данном контексте не опровергают концепцию Дубровского, равно как и наоборот. При разработке и интерпретации подобных экспериментов необходимо отдавать себе отчет в том, что ни одна из модификаций системы не способна окончательно доказать одну теорию и опровергнуть остальные.

Тем не менее накопленный массив эмпирических данных позволит оценивать философские и нейробиологические теории по их сравнительной объяснительной силе.

Также необходимо учитывать, что связь поведенческих индикаторов с феноменальным опытом всегда будет оставаться предметом интерпретации, допуская различные трактовки в зависимости от исходных философских установок исследователя. Проблема «философского зомби» или искусной имитации также сохраняется: даже если система будет демонстрировать все ожидаемые маркеры, это не будет проинтерпретировано всеми учеными как доказательство наличия у нее субъективного опыта. Однако если система, построенная на определенных теоретических принципах, последовательно демонстрирует поведение, которое трудно объяснить без допущения неких аналогов внутренних состояний и целеполагания, без допущения наличия целенаправленно изменяемых самой моделью установок, не сводимых напрямую к внешним условиям или методам обучения, то это повышает правдоподобие данных принципов. Сложность реализации и контроля чистоты эксперимента, когда изменяется только один целевой параметр, также представляет значительную трудность. Цель эксперимента – не ответить на вопрос, почему и как эти процессы генерируют qualia, а эмпирически связать поведенческие и функциональные корреляты сознания (и шире – психологических процессов) с конкретными, контролируемыми изменениями в архитектуре, методах обучения и внешних условиях. Это достигается, в том числе, путем целенаправленного помещения модели в «пограничные» для нее состояния – ситуации конфликта, неполноты данных или внешних ограничений – и систематической фиксации ее реакций, что позволяет сравнивать их со сложными, в том числе иррациональными, проявлениями человеческой деятельности.

Эксперимент на основе двухуровневой когнитивной архитектуры

Одна из фундаментальных особенностей человеческого мышления заключается в различении мыслей (смысловой уровень) и слов (языковой уровень). Мы можем осуществлять сложные действия или размышлять, не формулируя полностью словесно наш когнитивный процесс, и наоборот – переводить невербальные интуиции в языковую форму. Еще одно очевидное проявление

разделения этих уровней – то, что в разных языках для передачи одного и того же смысла необходимы последовательности слов. В архитектуре трансформеров разделение смыслового и языкового уровней практически отсутствует: все операции происходят в едином векторном пространстве эмбеддингов. Хотя можно условно говорить, что эмбеддинг после токенизации представляет текст, а после прохождения через блоки внимания – смысл, фактически все эти представления существуют в одном и том же пространстве. Трансформер затрачивает равное количество вычислительных ресурсов на генерацию любого токена, будь то пунктуационный знак или ключевое слово в решении сложной задачи.

Переход из непрерывного латентного пространства в дискретный словарь токенов оказывается, по всей видимости, архитектурным «узким горлышком» стандартного трансформера. Косвенным свидетельством в пользу этого служат тенденции, наметившиеся приблизительно с 2024 года: простое увеличение числа параметров дает убывающую отдачу, тогда как основной прирост производительности обеспечивается «надстройками» – цепочками рассуждений и агентными системами. По существу, эти надстройки восполняют недостаток, возникающий из-за того, что базовая архитектура вынуждена «думать» посредством последовательного порождения токенов, а не в пространстве непрерывных представлений[6].

Такой механизм явного соответствия между дискретным и континуальным представлениями в определенной степени созвучен идеям информационной теории сознания Д.И. Дубровского. Ключевое место в его теории занимает понятие «кодовой зависимости», согласно которой феномены субъективной реальности (как информация) находятся в отношении специфической кодовой связи со своими нейродинамическими носителями [Дубровский 2024а, 12]. Возникновение субъективного опыта, по Дубровскому, требует

[6] Примечательно, что модели TRM (Tiny Recursive Model), HRM (Hierarchical Reasoning Model) и URM (Universal Reasoning Model) [Gao et al. 2025], построенные на архитектуре, которая предполагает явную рекуррентность в латентном пространстве, опосредующую генерацию не отдельного токена, а целого блока, на определенном типе задач обходят БЯМ с большим на несколько порядков количеством параметров. Ограничение этих моделей, однако, состоит в крайне узком наборе генерируемых токенов, что делает прямой перенос их архитектурных решений на языковые модели с полноценным текстовым словарем нетривиальной задачей.

многоуровневой кодовой трансформации, в результате которой информация освобождается от физических характеристик своего первоначального носителя.

Основная идея второго эксперимента заключается в создании системы с явным разделением смысловой и языковой обработки. В отличие от первого эксперимента, фокусирующегося на «собственном контексте» как аналоге внутреннего опыта, здесь акцент делается на исследовании того, как разделение уровней репрезентации влияет на когнитивные способности системы[7].

В этом контексте интерес представляет, в частности, развитие возможностей генеративных диффузионных моделей. Получив первоначальное широкое распространение в задачах генерации континуальных данных, таких как изображения [Sohl-Dickstein et al. 2015], в последнее время этот подход был успешно адаптирован и для работы с дискретными данными, в частности с текстом, породив направление лингвистических диффузионных моделей [Austin et al. 2021]. Ключевая идея таких моделей заключается в преобразовании дискретных токенов (слов) в континуальное векторное пространство, где к ним применяется контролируемый процесс зашумления. Модель затем обучается обращать этот процесс, итеративно восстанавливая исходную информацию. Недавно разработанные архитектуры, такие как LLaDA [Nie et al. 2025] и Seed Diffusion [Song et al. 2025], предлагают полуавторегрессивную генерацию (в отличие от авторегрессивной модели чистых трансформеров). В этих архитектурах диффузионная модель представляет собой алгоритмическую надстройку, управляющую работой трансформеров. Вместо генерации по одному токену, рассматриваемые модели параллельно предсказывают целый блок токенов, а затем итеративно уточняют его, на каждом шаге закрепляя наиболее уверенные предсказания и повторно маскируя менее уверенные для дальнейшей корректировки.

Дальнейшее развитие этой идеи ведет к еще более глубокому архитектурному разделению, предложенному в рамках концепции Score Interpolation Diffusion Models (SCIDM) [Dieleman et al. 2024]. Этот подход не просто модифицирует процесс вывода, а вносит принципиальное различие между компонентами системы. В данной архитектуре модель делится на интерфейс для работы

[7] Возможна также система, совмещающая идеи первого и второго экспериментов, которые взаимодополняют, а не исключают друг друга.

с дискретными объектами[8], который включает входную матрицу эмбеддингов и выходной проекционный слой, и ядро вычислений, выполняемых трансформером. Это ядро оперирует исключительно в непрерывном (континуальном) векторном пространстве, принимая на вход зашумленные эмбеддинги и предсказывая их «очищенную» версию. В перспективе этот процесс, где единица генерации – не отдельный токен, а их последовательность, которая постепенно «проявляется» из шума, уже можно рассматривать как шаг к функциональному разделению уровней: блок токенов как аналог единицы смысла, а отдельные токены – как единица языка.

Такое более четкое архитектурное разграничение между символьным интерфейсом и субсимвольным ядром открывает путь для принципиально новых экспериментальных гипотез. Если ключевые когнитивные операции, такие как рассуждение, могут быть реализованы непосредственно в континуальном пространстве, это позволяет сформулировать идею «латентного мышления» («latent reasoning»). Вместо того чтобы генерировать текстовую цепочку рассуждений, которая является стандартом для современных «рассуждающих» авторегрессивных моделей (DeepSeek R1, OpenAI o3, o4 и т.д.), система может выполнять итеративный процесс уточнения непосредственно в пространстве непрерывных эмбеддингов до того, как будет сгенерирован новый блок символьной последовательности[9].

---

[8] Под дискретными объектами понимаются токены из конечного словаря. Каждому токену соответствует конкретный вектор из непрерывного пространства эмбеддингов. Непрерывным представлением называется любой вектор (или последовательность векторов) в пространстве эмбеддингов, не ограниченный словарем. В стандартном трансформере в процессе инференса скрытый слой, состоящий из непрерывного вектора, опосредует генерацию одного токена. В SCIDM скрытый слой, представляющий собой непрерывную матрицу (т.е. последовательность непрерывных векторов), опосредует генерацию последовательности (блока) токенов. Причем, если применить, модифицировав под особенности SCIDM, метод, описанный в статье [Li et al. 2025], то скрытый слой будет опосредовать последовательность токенов не фиксированной, а вариативной длины.

[9] Гипотеза состоит в следующем: архитектуру типа SCIDM можно адаптировать для многошагового рассуждения без генерации промежуточного текста. Идея – «латентное мышление» как цепочка полных циклов обратной диффузии. На каждом шаге результат расшумления не декодируется в текст, а служит контекстом для следующего шага: к очищенному латентному состоянию добавляется новый блок зашумленных

Нельзя исключать, что данный подход будет иметь и концептуальное значение, поскольку позволит в какой-то степени воспроизвести двухуровневость организации эго-системы. Представляет интерес проследить, насколько насколько эта симуляция двухуровневости будет способствовать появлению феноменов, связанных со спецификой субъективной реальности[10].

Принцип двухуровневости, который стремимся воспроизвести, можно проиллюстрировать на примере когнитивного процесса при вождении автомобиля. Большую часть времени водитель действует в «неосознанном», автоматическом режиме: он совершает множество микродвижений и корректировок, опираясь на непрерывный поток сенсомоторной информации, не формулируя свои действия словами. Это быстрый и эффективный контур управления. Однако при столкновении с непредвиденной трудностью – например, резким маневром другой машины или сложными погодными условиями – водитель переключается в «сознательный» режим. Он начинает «проговаривать» ситуацию про себя, вербализуя проблему, оценивая варианты и отдавая себе отчетливые команды. Этот переход от прямого латентного контроля к языковому опосредованию и является ключевой особенностью предлагаемой модели.

векторов, и цикл запускается снова. Лишь после всех таких макроитераций финальное латентное состояние подается на декодер для однократной генерации ответа. Разумеется, данная гипотеза на текущем этапе носит во многом умозрительный характер и требует серьезной экспериментальной проверки. Реализация «латентного мышления» сопряжена со значительными техническими трудностями, например, с разработкой методов обучения, которые бы целенаправленно поощряли осмысленные многошаговые вычисления в латентном пространстве, а не простое зацикливание.

[10] В качестве перспективного направления стоит отметить следующее. Если итерационный процесс на субсимвольном уровне будет организован как движение не к фиксированному дискретному словарю, а к устойчивым аттракторам в непрерывном пространстве (что формально может быть реализовано через ODE-солвер), то открывается возможность воспроизведения своего рода эволюционного алгоритма без предварительного обучения модели на языковых данных. Выходные состояния системы в этом случае окажутся континуальными – подобно тому, как моторные действия человека изначально непрерывны, а дискретные языковые единицы возникли в ходе эволюции. Теоретически интересным остается вопрос о том, способна ли система в ходе обучения самостоятельно выработать собственную дискретную знаковую систему, если в обучающем наборе данных не было текстов человеческого языка.

В предлагаемой архитектуре этот феномен может эмулироваться через разделение когнитивных процессов, причем сама модель обработки, реализованная как иерархическая смесь «экспертов» (mixture of experts, MoE), отражает принцип специализации, имеющий место в головном мозге с его многочисленными отделами. «Неосознанный» контур соответствует операциям, которые происходят на нижних, вычислительно малозатратных уровнях этой иерархии. Важно отметить, что данные от условных «датчиков» могут подаваться напрямую в латентное пространство этих экспертов, минуя этап дискретной токенизации. Простые, рутинные операции могут обрабатываться легковесными «экспертами» – возможно, даже не трансформерной, а классической нейросетевой архитектуры. Если же задача не решается на базовом уровне или ошибка предсказания высока, задача передается на следующие, более высокие уровни – к более мощным (со все большим количеством параметров), но и более затратным «экспертам». Но пока не будет обращения к языку, процессы будут носить характер «автоматических», «неосознанных» действий. «Сознательный» же процесс – это высший и самый затратный уровень этой иерархии. Он активируется, когда предыдущие «эксперты» не справляются с задачей. Именно в этот момент система выходит из операций в латентном режиме и инициирует полный цикл «кодирование – декодирование»: она проецирует свое внутреннее континуальное состояние в дискретные символы (слова), сохраняющиеся в СК, тем самым получая возможность обладать уровнем «информации об информации», осуществлять вербальное рассуждение («мыслить словами») и формировать выраженную в языке историю своей деятельности.

В этом и заключается принципиальное отличие второго эксперимента от первого: вместо допущения того, что все процессы, независимо от их статуса, требуют полного цикла обработки единым мощным трансформером, здесь предлагается модель с адаптивным, специализированным распределением вычислительных ресурсов. Такой подход не только экономичнее, но и гораздо ближе к принципам организации биологического мозга, где сложные задачи также передаются по уровням иерархии от локальных сенсомоторных контуров к глобальным областям, отвечающим за планирование и язык. И дискретные, выраженные в языке данные являются своего рода информацией об информации латентного слоя.

Но будет ли данный подход иметь концептуальное значение в том смысле, что приведет к той двухуровневости эго-системы, которая способствует возникновению субъективной реальности? Вряд ли, конечно, описанный подход предлагает достаточные условия для воспроизводства субъективной реальности, но как концептуальный проект для экспериментирования в этом направлении может быть интересен. Конечная цель состоит в том, чтобы разработать архитектуру, которая наделяет субсимвольный уровень собственным значением. С одной стороны, имеется центральное когнитивное ядро, которое оперирует исключительно в континуальном латентном пространстве. Эту часть системы можно рассматривать как функциональный аналог высшей нервной деятельности, лишенной языка, – например, сложной когнитивной системы животного, способной к восприятию, обучению и формированию сложного поведения, но не оперирующей дискретными символами. С другой стороны, итеративный цикл «кодирование − декодирование» выступает в роли той критической надстройки, которая обеспечивает языковое опосредование когнитивного нейросетевого процесса. Именно этот непрерывный процесс преобразования внутренних континуальных состояний в дискретные символы и обратно позволяет системе «мыслить словами», формируя модель высшей нервной деятельности, которая уже обладает языком.

Заключение

Развитие ИИ, в особенности БЯМ, ставит перед исследователями не только технологические, но и философские вопросы о природе сознания и возможности его возникновения в небиологических системах. В настоящей работе была осуществлена попытка соотнесения теорий сознания с архитектурами ИИ, а также предложен экспериментальный подход, направленный на создание условий для получения эмпирических данных, анализ которых может способствовать проверке теоретических предсказаний.

Вероятно, ни одна из существующих теорий не способна исчерпывающе объяснить все аспекты сознания. Скорее, разные подходы расставляют акценты на его отдельных гранях (например, прогностическое кодирование выделяет функцию минимизации расхождений между ожиданиями и реальностью). Сравнение теорий ставит перед исследователями два фундаментальных вопроса.

Во-первых, какие структурно-функциональные качества системы строго необходимы для возникновения сознания (в том числе на небиологическом субстрате), а какие – лишь сопутствуют ему? Во-вторых, какие из этих качеств специфически коррелируют именно с наличием сознания, а не с общими когнитивными способностями? Эвристическую ценность теорий следует оценивать по тому, насколько четко они отвечают на второй вопрос.

Этот второй вопрос особенно актуализируется на фоне современных БЯМ, демонстрирующих феномен «мышления без сознания» – способность к сложной обработке информации и рассуждениям при отсутствии явных признаков феноменального опыта. В связи с этим возникает принципиальная задача: разграничить архитектурные условия сложной системы, отвечающие за порождение сознания как такового, от условий, обеспечивающих «мышление» (в его широком, информационно-процессуальном смысле). Предложенное экспериментальное исследование направлено именно на выявление этих границ. Оно призвано помочь эмпирически изолировать те факторы, которые критичны для возникновения аналогов сознательного опыта, а не просто для обеспечения продвинутых вычислительных функций.

В этом контексте информационная теория сознания Д.И. Дубровского заслуживает особого внимания, поскольку она, будучи разработанной профессиональным философом, во многом отходит от некоторых традиционных для философского исследования сознания установок. Классическая философия, начиная с Декарта, связывала мышление и сознание с особой субстанцией (res cogitans) или, как у Канта, со специфическими инстанциями – созерцанием, рассудком и разумом, ответственными за осознанность перцептивного опыта и cogito. Теория Дубровского, предлагая функционально-информационное объяснение субъективной реальности, в определенном смысле порывает с этой традицией гипостазирования процессов осознания, не приписывая им отдельной онтологической сущности или уникальной инстанции, но рассматривая их как результат специфической организации информационных процессов в сложной системе. Такой подход, возможно, сужает поле для некоторых классических метафизических дискуссий о сознании (и потому с трудом получает принятие в философском сообществе), однако он оказывается созвучным современным тенденциям в развитии ИИ. Нынешние большие языковые модели и агенты, построенные на их основе,

демонстрируют все более сложные формы поведения, которые ранее приписывались исключительно существам, обладающим сознанием. Примечательно, что при разработке таких систем целенаправленной работы по созданию некой отдельной инстанции, отвечающей за «осознанность перцептивного опыта» или «cogito», как правило, не ведется. Этот факт может служить аргументом в пользу того, что при дальнейших исследованиях сознания целесообразно в большей мере учитывать подходы, подобные тому, что развивает Дубровский, акцентирующие внимание на структурных и функциональных аспектах. Разумеется, возможны возражения, что предложенный экспериментальный подход не способен предложить решение трудной проблемы сознания, которая во многом носит концептуальный характер. Однако его цель и не в этом. Суть предложенного подхода как раз и состоит в том, чтобы на практике исследовать, насколько функциональные и архитектурные решения, выведенные из теорий, позволяют воспроизвести феноменологию, ассоциируемую с сознанием, qualia.

На современном этапе технологического развития исследования сущностных и акцидентальных свойств и условий сознания могут носить не только теоретический, но и экспериментальный характер. Предложенная экспериментальная архитектура, безусловно, не претендует на окончательное решение проблемы искусственного сознания, но представляет собой попытку наметить путь для эмпирической проверки и сопоставления различных функциональных (и в какой-то степени феноменальных) теорий сознания.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

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

Дубровский 1980 – *Дубровский Д.И.* Информация, сознание, мозг. – М.: Высшая школа, 1980.

Дубровский 2007 – *Дубровский Д.И.* Сознание, мозг, искусственный интеллект. – М.: Стратегия-Центр, 2007.

Дубровский 2024а – *Дубровский Д.И.* Проблема психического управления: нейронаука, искусственный интеллект, социальные коммуникации // Философские науки. 2024. Т. 67. № 1. С. 7–28.

Дубровский 2024б – *Дубровский Д.И.* Сознание и мозг. Опыт разработки основных теоретических вопросов «Трудной проблемы созна-

ния» и ее значение для нейронауки // Философские науки. 2024. Т. 67. № 3. С. 142–158.

Недяк 2025 – *Недяк А.В.* Возможен ли диалог с техносубъектом? Архитектуры искусственного интеллекта и признаки сознания // Философские науки. 2025. Т. 68. № 3. С. 93–113.

Austin et al. 2021 – *Austin J., Johnson D.D., Ho J., Tarlow D., van den Berg R.* Structured Denoising Diffusion Models in Discrete State-Spaces // Advances in Neural Information Processing Systems. 2021. Vol. 34. P. 17981–17993.

Baars 1988 – *Baars B.J.* A Cognitive Theory of Consciousness. – Cambridge, UK: Cambridge University Press, 1988.

Clark 2015 – *Clark A.* Surfing Uncertainty: Prediction, Action, and the Embodied Mind. – Oxford: Oxford University Press, 2015.

Dehaene, Changeux 2011 – *Dehaene S., Changeux J.-P.* Experimental and Theoretical Approaches to Conscious Processing // Neuron. 2011. Vol. 70. No. 2. P. 200–227.

Dieleman et al. 2024 – *Dieleman S.E.L., Sartran L.P.M., Savinov N., Ganin I., Richemond P., Doucet A., ..., Hawthorne C.G.-M.* Score Interpolation Diffusion Models (International Publication No. WO 2024/110596 A1). World Intellectual Property Organization. – URL: https://patents.google.com/patent/WO2024110596A1/en

Elamrani 2025 – *Elamrani A.* Introduction to Artificial Consciousness: History, Current Trends and Ethical Challenges // arXiv preprint. 2025. arXiv:2503.05823.

Gao et al. 2025 – *Gao Z., Chen L., Xiao Y., Xing H., Tao R., Luo H., Zhou J., Dai B.* Universal Reasoning Model // arXiv preprint. 2025. arXiv:2512.14693.

Lamme 2006 – *Lamme V.A.F.* Towards a True Neural Stance on Consciousness // Trends in Cognitive Sciences. 2006. Vol. 10. No. 11. P. 494–501.

Li, Fung 2025 – *Li M.Q., Fung B.* Security Concerns for Large Language Models: A Survey // arXiv preprint. 2025. arXiv:2505.18889.

Li et al. 2025 – *Li J., Dong X., Zang Y., Cao Y., Wang J., Lin D.* Beyond Fixed: Variable-Length Denoising for Diffusion Large Language Models // arXiv preprint. 2025. arXiv:2508.00819.

Nie et al. 2025 – *Nie S., Zhu F., You Z., Zhang X., Ou J., Zhou J., ..., Li C.* Large Language Diffusion Models // arXiv preprint. 2025. arXiv:2502.09992.

Rosenthal 2005 – *Rosenthal D.M.* Consciousness and Mind. – Oxford: Clarendon Press, 2005.

Shiller, Petrunya 2021 – *Shiller A.V., Petrunya O.E.* Architectural Approach to Design of Emotional Intelligent Systems // Russian Journal of Philosophical Sciences = Filosofskie nauki. Vol. 64, no. 1, pp. 102–115.

Sohl-Dickstein 2015 – *Sohl-Dickstein J., Weiss E.A., Maheswaranathan N., Ganguli S.* Deep Unsupervised Learning Using Nonequilibrium Thermodynamics // Proceedings of the 32nd International Conference on Machine Learning. 2015. Vol. 37. P. 2256-2265.

Song et al. 2025 – *Song Y., Zhang Z., Luo C., Gao P., Xia F., Luo H., ..., Zhou H.* Seed Diffusion: A Large-Scale Diffusion Language Model with High-Speed Inference // arXiv preprint. 2025. arXiv:2508.02193.

Tononi 2008 – *Tononi G.* Consciousness as Integrated Information: A Provisional Manifesto // Biological Bulletin. 2008. Vol. 215. No. 3. P. 216–242.

REFERENCES

Austin J., Johnson D.D., Ho J., Tarlow D., & van den Berg R. (2021) Structured Denoising Diffusion Models in Discrete State-Spaces. *Advances in Neural Information Processing Systems.* Vol. 34, pp. 17981–17993.

Baars B.J. (1988) *A Cognitive Theory of Consciousness.* Cambridge, UK: Cambridge University Press.

Clark A. (2015) *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford: Oxford University Press.

Dehaene S. & Changeux J.-P. (2011) Experimental and Theoretical Approaches to Conscious Processing. *Neuron.* Vol. 70, no. 2, pp. 200–227.

Dieleman S.E.L., Sartran L.P.M., Savinov N., Ganin I., Richemond P., Doucet A., …, & Hawthorne C.G.-M. (2024) Score Interpolation Diffusion Models (International Publication No. WO 2024/110596 A1). In: *World Intellectual Property Organization.* Retrieved from https://patents.google.com/patent/WO2024110596A1/en

Dubrovsky D.I. (1971) *Mental Phenomena and the Brain: A Philosophical Analysis of the Problem in Connection with Some Current Tasks of Neurophysiology, Psychology, and Cybernetics.* Moscow: Nauka (in Russian).

Dubrovsky D.I. (1980) *Information, Consciousness, Brain.* Moscow: Vysshaya shkola (in Russian).

Dubrovsky D.I. (2007) *Consciousness, Brain, Artificial Intelligence.* Moscow: Strategiya-Tsentr (in Russian).

Dubrovsky D.I. (2024a) The Problem of Mental Control: Neuroscience, Artificial Intelligence, Social Communications. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 67, no. 1, pp. 7–28 (in Russian).

Dubrovsky D.I. (2024b) Mind and Brain: Experience in Developing Fundamental Theoretical Issues of the Hard Problem of Consciousness and Its

Significance for Neuroscience. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 67, no. 3, pp. 142–158 (in Russian).

Elamrani A. (2025) Introduction to Artificial Consciousness: History, Current Trends and Ethical Challenges. *arXiv preprint.* arXiv:2503.05823.

Gao Z., Chen L., Xiao Y., Xing H., Tao R., Luo H., Zhou J., & Dai B. (2025) Universal Reasoning Model. *arXiv preprint.* arXiv:2512.14693.

Lamme V.A.F. (2006) Towards a True Neural Stance on Consciousness. *Trends in Cognitive Sciences.* Vol. 10, no. 11, pp. 494–501.

Li J., Dong X., Zang Y., Cao Y., Wang J., & Lin D. (2025) Beyond Fixed: Variable-Length Denoising for Diffusion Large Language Models. *arXiv preprint.* arXiv:2508.00819.

Li M.Q. & Fung B. (2025) Security Concerns for Large Language Models: A Survey. *arXiv preprint.* arXiv:2505.18889.

Nedyak A.V. (2025) Is Dialogue with a Technosubject Possible? Architectures of Artificial Intelligence and Signatures of Consciousness. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 68, no. 3, pp. 93–113 (in Russian).

Nie S., Zhu F., You Z., Zhang X., Ou J., Zhou J., …, & Li C. (2025) Large Language Diffusion Models. *arXiv preprint.* arXiv:2502.09992.

Rosenthal D.M. (2005) *Consciousness and Mind.* Oxford: Clarendon Press.

Shiller A.V. & Petrunya O.E. (2021) Architectural Approach to Design of Emotional Intelligent Systems. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 64, no. 1, pp. 102–115.

Sohl-Dickstein J., Weiss E.A., Maheswaranathan N., & Ganguli S. (2015) Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. *Proceedings of the 32nd International Conference on Machine Learning.* Vol. 37, pp. 2256–2265.

Song Y., Zhang Z., Luo C., Gao P., Xia F., Luo H., …, & Zhou H. (2025) Seed Diffusion: A Large-Scale Diffusion Language Model with High-Speed Inference. *arXiv preprint.* arXiv:2508.02193.

Tononi G. (2008) Consciousness as Integrated Information: A Provisional Manifesto. *Biological Bulletin.* Vol. 215, no. 3, pp. 216–242.