DOI: 10.30727/0235-1188-ZQJQOL
Оригинальная исследовательская статья
Original research paper

Мыслят ли большие языковые модели? Технический ответ на философский вопрос

И.Ф. Михайлов
Институт философии Российской академии наук, Москва, Россия

Аннотация

В статье рассматривается вопрос о том, реализуют ли современные большие языковые модели процессы, функционально эквивалентные мышлению, и, если не реализуют, то в чем именно состоит разрыв между машинным и биологическим интеллектом. Центральный тезис состоит в том, что сознание не является ни онтологически необходимым условием мышления, ни эпистемологически достаточным критерием его верификации. Это положение, восходящее к Лейбницу и последовательно подтверждавшееся историей когнитивной науки, от Гельмгольца до Фристона, в статье распространяется на искусственные системы. Языковая модель обладает пониманием в функциональном смысле, что делает некорректным как отрицание этой компетенции на интроспективных основаниях, так и ее апологию. Архитектурный парадокс трансформера – рекурсивная продуктивность без рекурсивной архитектуры – эмпирически разрешает спор Хомского со Скиннером, не снимая теоретического вопроса о природе рекурсивной компетенции. Типологическое основание для этого вывода составляет структурное сходство двух операций: минимизации вариационной свободной энергии в предиктивном кодировании и вычисления взвешенного сходства в механизме внимания. В обоих случаях имеет место итеративное уточнение представления через редукцию расхождения между предсказанием и реальностью. Факт содержательного интеллектуального общения между людьми и языковыми моделями – совместного рассуждения, взаимной коррекции, продуктивного несогласия – был бы необъясним в отсутствие этого типологического родства. Разрыв между биологическим и искусственным интеллектом прослеживается на четырех уровнях: энергетическом, вычислительном, семантическом и темпоральном. На каждом из них обнаружено не количественное отставание, а качественно иной способ организации информационного процесса. Сделан вывод о том, что вопрос о природе машинного мышления требует переформулировки: оба процесса – трансформерный и

биологический – инстанции одного вычислительного типа – иерархической редукции неопределенности, – что составляет типологическую характеристику интеллекта как такового. Разрыв между двумя его видами реален, но располагается внутри одного рода.

Ключевые слова: эпистемология, языковые модели, трансформеры, предиктивное кодирование, принцип свободной энергии, рекурсия, марковские процессы, воплощенное познание, байесовский вывод, бессознательная когниция, искусственный интеллект.

Михайлов Игорь Феликсович – доктор философских наук, ведущий научный сотрудник сектора методологии междисциплинарных исследований человека, Институт философии РАН.
ifmikhailov@gmail.com
http://orcid.org/0000-0001-8511-8849

Для цитирования: *Михайлов И.Ф.* Мыслят ли большие языковые модели? Технический ответ на философский вопрос // Философские науки. 2026. Т. 69. EDN: ZQJQOL.
DOI: 10.30727/0235-1188-ZQJQOL

Do Large Language Models Think? A Technical Answer to a Philosophical Question

I.F. Mikhailov
Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia

Abstract

The article investigates whether contemporary large language models instantiate processes that are functionally equivalent to thinking and, if they do not, where exactly the gap between machine and biological intelligence lies. Its central claim is that consciousness is neither an ontologically necessary condition for thought nor an epistemologically sufficient criterion for establishing the presence of thought. This claim, which can be traced back to Leibniz and has been repeatedly reinforced in the history of cognitive science, from Helmholtz to Friston, is extended here to artificial systems. In a functional sense, a language model may be said to possess understanding, which makes both the denial of such understanding on introspective grounds and its uncritical apologetic defense inadequate. The architectural paradox of the transformer – recursive productivity in the absence of a recursive architecture – empirically settles the Chomsky–Skinner dispute

without resolving the theoretical question of the nature of recursive competence. The typological basis for this claim lies in the structural correspondence between two operations: the minimization of variational free energy in predictive coding and the computation of weighted similarity in the attention mechanism. In both cases, representations are iteratively refined through the reduction of the discrepancy between the model and the input data. The possibility of substantive intellectual exchange between humans and language models – joint reasoning, mutual correction, and productive disagreement – would be difficult to account for in the absence of this typological kinship. At the same time, the gap between biological and artificial intelligence can be traced across four levels: energetic, computational, semantic, and temporal. At each of these levels, what is at stake is not a merely quantitative lag but a qualitatively distinct mode of organizing informational processes. The article concludes that the question of machine thinking must therefore be reformulated: transformer-based and biological processes may be understood as different instantiations of a single computational type – the hierarchical reduction of uncertainty – which constitutes a typological feature of intelligence as such. The gap between these two forms of intelligence is real, but it lies within a common genus rather than between fundamentally different genera.

Keywords: epistemology, language models, transformers, predictive coding, free-energy principle, recursion, Markov processes, embodied cognition, Bayesian inference, unconscious cognition, artificial intelligence.

Igor F. Mikhailov – D.Sc. in Philosophy, Leading Research Fellow at the Department of Methodology of the Interdisciplinary Study of Man, Institute of Philosophy, Russian Academy of Sciences.

ifmikhailov@gmail.com

http://orcid.org/0000-0001-8511-884

For citation: Mikhailov I.F. (2026) Do Large Language Models Think? A Technical Answer to a Philosophical Question. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 69. EDN: ZQJQOL.
DOI: 10.30727/0235-1188-ZQJQOL

Введение

В 2023 году Н. Хомский с соавторами опубликовал в газете «New York Times» статью[1], в которой охарактеризовал большие

[1] *Chomsky N., Roberts I., Watumull J.* The False Promise of ChatGPT // The New York Times. March 8, 2023. – URL: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

языковые модели как «высокотехнологичный плагиат» и «избегание знания». Аргумент узнаваем: те же основания, по которым в 1959 году была отвергнута программа Б. Ф. Скиннера, применимы и к современным нейросетям, а именно – статистическая аппроксимация поверхности языка не есть овладение его глубинной структурой. Позиция вполне последовательна, но она игнорирует то обстоятельство, что за шестьдесят четыре года между двумя текстами произошло нечто, требующее не повторения старого аргумента, а его пересмотра.

Это «нечто» не сводится к техническому прогрессу, к тому, что модели стали больше, быстрее и точнее. Речь идет о концептуальном сдвиге, который обнаруживается при внимательном анализе архитектуры трансформера и ее отношения к тому, что когнитивная наука называет мышлением. Системы, реализующие вероятностное предсказание следующего токена, освоили рекурсивную продуктивность – свойство, которое Хомский считал принципиально недостижимым для вероятностных моделей. Это не означает, что Хомский был неправ во всем. Это означает, что спор, который казался решенным в пользу дискретной, правилоуправляемой модели языка, оказался значительно сложнее – и значительно интереснее.

Настоящая статья не является ни апологией языковых моделей, ни их критикой. Ее предмет – попытка точного измерения дистанции между тем, что реализовано в современных архитектурах искусственного интеллекта, и тем, что биологическая эволюция реализовала в нейронных системах. Измерение такой дистанции требует одновременно философской и технической строгости: философской – поскольку понятие мышления нуждается в операционализации прежде, чем мы спросим, реализовано ли оно в данной системе; технической – поскольку ответ зависит от конкретных архитектурных решений, а не от общих интуиций.

Методологическим ориентиром служит тезис, восходящий к Лейбницу: мышление не таково, каким его наблюдает сознание. Бессознательные основания когнитивных процессов – то, что Лейбниц называл «petites perceptions», малыми восприятиями, недоступными рефлексии, – конституируют содержание сознательного опыта, не будучи сами сознательными. Это методологическое предупреждение работает в обе стороны: интроспективная очевидность («я знаю, что значит мыслить, потому что я мыслю») не является достаточным аргументом ни в пользу того,

что машина не мыслит, ни в пользу того, что она мыслит. Вопрос требует других оснований.

Статья организована следующим образом. В первом разделе сформулирована философская проблема: что именно ставится под вопрос с появлением языковых моделей и почему традиционные ответы (функционалистский и антифункционалистский) одинаково неудовлетворительны. Во втором разделе проанализированы архитектурный парадокс трансформера (рекурсивная продуктивность без рекурсивной архитектуры) и то, почему этот парадокс не разрешен ни технически, ни теоретически. В третьем разделе описывается геометрия разрыва между биологическим и искусственным интеллектом на четырех уровнях: энергетическом, вычислительном, семантическом и темпоральном. В заключении возвращаемся к исходному вопросу и формулируем его с той точностью, которую предшествующий анализ сделал возможной.

Вопрос «Мыслят ли языковые модели?» законен. Он не предполагает простого ответа. Но мы покажем, что он может быть поставлен с научной точностью, и это уже возможно считать философским результатом.

I. Мышление, каким его не видит сознание

В письме Христиану Гольдбаху от 17 апреля 1712 года Лейбниц формулирует тезис, которому предстоит долгая жизнь: «Musica est exercitium arithmeticae occultum nescientis se numerare animi» («Музыка есть тайное арифметическое упражнение души, не знающей, что она считает (перевод мой. – *И. М.*)») (цит. по: [Juschkewitsch, Kopelewitsch 1988, 182]). Душа вычисляет, но не знает об этом. Результат вычисления достигает сознания не как число, а как удовольствие от консонанса или неудовольствие от диссонанса: «ex multis enim congruentiis insensibilibus oritur voluptas» («из множества нечувствительных совпадений рождается наслаждение»). Процесс скрыт – явлен только его аффективный след.

Это не частное наблюдение о природе музыкального восприятия, а методологический манифест, направленный против картезианского тезиса о прозрачности сознания для самого себя. В этом же письме Лейбниц пишет: «Errant enim, qui nihil in anima fieri putant, cujus ipsa non sit conscia» («Ошибаются те, кто думает, что в душе не происходит ничего, о чем она сама не знает»). Лейбниц адресует этот упрек Декарту, Локку, Бейлю, всем, кто отождествил мышление с сознательным мышлением, кто положил апперцеп-

цию в основание когнитивной онтологии. Ошибка, по Лейбницу, не эмпирическая, а структурная, и состоит она в том, что за критерий мышления принимается его видимость изнутри.

Позже, в «Монадологии» [Лейбниц 1982], Лейбниц разворачивает симметричный аргумент, но уже в противоположном направлении. Представим, что мозг увеличен до размеров мельницы, можно войти внутрь и осмотреть все его механизмы. Мы увидим шестерни, рычаги, передачи, но нигде не обнаружим ничего, что объясняло бы восприятие. Если музыкальный тезис направлен от данности к ее основаниям (собственный механизм скрыт от сознания), то аргумент мельницы направлен противоположно: феноменальный опыт скрыт от внешнего наблюдателя механизма. Вместе они формулируют два симметричных ограничения, складывающихся в единый вывод: перспектива первого лица и перспектива третьего лица взаимно непереводимы. Сознание не видит своих оснований; наблюдатель оснований не видит сознания. Именно взаимная непереводимость перспектив делает вопрос «Мыслит ли система?» принципиально нетривиальным, независимо от того, идет ли речь о человеке или машине. Это не скептический парадокс и не повод для мистики. Это рабочая гипотеза, которую когнитивная наука последовательно подтверждала на протяжении двух с половиной веков, от гельмгольцевских «бессознательных умозаключений» до современного предиктивного[2] кодирования. Гельмгольц показал, что восприятие есть не пассивная регистрация, а активное построение гипотезы о причинах сенсорных данных посредством механизма «анализ через синтез»: нисходящие предсказания накладываются на восходящий сенсорный поток, и сознание получает результат этого согласования. То, что мы «видим», есть интерпретированный результат, а не сырой материал. Сознание получает не данные, а выводы.

К. Фристон радикализирует этот тезис [Friston et al. 2017]. В рамках *принципа свободной энергии* мозг непрерывно минимизирует расхождение между поступающими данными и выводами из своей априорной гипотезы. Ошибка предсказания – несовпадение ожидаемого с действительным – есть единственное, что пробивается вверх, по иерархии. Остальное подавляется как из-

[2] Некоторые отечественные авторы предпочитают переводить термин «predictive» как «прогнозирующий» или «предсказательный». Но мы полагаем, что русскоязычная научная терминология должна сохранять однозначную обратную переводимость.

быточное. Сознательный опыт в этой схеме не окно в реальность, а поверхность, на которую проецируется наиболее вероятная гипотеза о реальности. Вывод происходит в глубине иерархии; сознание фиксирует его следствия.

Казалось бы, спор о механизмах – это специальная проблема когнитивной науки, не затрагивающая философию мышления как такового. Однако появление больших языковых моделей сделало этот спор неотложным и придало ему новое измерение. Системы, не имеющие ни биологической нервной системы, ни тела, ни эволюционной истории, ни, по всей видимости, какого-либо внутреннего опыта, демонстрируют поведение, которое по внешним признакам неотличимо от того, что мы привыкли называть мышлением. Они выстраивают аргументы, обнаруживают аналогии, решают задачи, требующие многошаговых умозаключений, и порождают тексты, в которых специалисты с трудом опознают машинное происхождение.

Стандартная реакция на это явление делится на два полюса, одинаково неудовлетворительных. Первый: «это не мышление, а имитация, статистическая компрессия текстового корпуса, лишенная понимания». Второй: «поскольку поведение неотличимо, значит, это и есть мышление» (функционализм как методологический принцип). Оба ответа предполагают, что мы знаем, что такое мышление, и спорим лишь о том, подпадает ли под это определение данный случай. Но именно это знание и находится под вопросом.

Лейбниц предлагает третий путь, менее комфортный, но более продуктивный. Если сознание – ненадежный свидетель собственных оснований, то интроспективная очевидность («я мыслю, следовательно, понимаю, что значит мыслить») не может служить критерием. Мышление может быть устроено иначе, чем оно себя переживает. Более того, по всей видимости, оно и устроено иначе. «Petites perceptions» Лейбница – малые восприятия, недоступные сознанию, но конституирующие его содержание. Это не метафора, а структурное утверждение: между вычислением и переживанием нет прямого и прозрачного соответствия. Душа считает и не знает, что считает. Результат счета она принимает за данность.

Из этого следует методологический сдвиг, принципиальный для данной статьи. Вопрос «Мыслят ли языковые модели?» некорректно ставить как вопрос о субъективном опыте. Мы не имеем инструментов для его верификации даже применительно к другим

людям, не говоря о машинах. Корректная постановка иная. Какие вычислительные процессы необходимы и достаточны для того, что мы функционально называем мышлением, и реализованы ли эти процессы в данной архитектуре? Это вопрос технический, но имеющий философские последствия, поскольку ответ на него неизбежно уточняет само понятие.

Историю когнитивной науки можно рассматривать как постепенный переход от линейно-символического понимания когнитивных процедур к биологически реалистическому. Это движение шло и продолжает идти по трем независимым, но сопряженным осям: от детерминизма к вероятности (мышление не выводит, оно взвешивает); от научения к предсказанию (мышление не накапливает, оно моделирует); наконец, от дискретного к непрерывному (мышление не манипулирует символами по правилам, оно интерполирует в многомерном пространстве вероятностей). Ни один из этих переходов не завершен. Каждый из них был полем особых исторических споров, но совокупно они описывают контур того, чем мышление является, в отличие от того, чем оно себе кажется.

Языковые модели реализуют все три оси движения в чистом, незамутненном виде, без биологии, без воплощения, без эволюционного багажа. Не в том смысле, что в них буквально происходит взвешивание вероятностей или построение моделей, но в том смысле, что математические операции трансформера реализуют такие же функциональные отношения, которые на уровне описания соответствуют этим переходам. Именно поэтому они представляют собой не просто технологический феномен, но философский инструмент: зеркало, в котором структура мышления как такового отражается без субъекта. Что именно показывает зеркало и где оно искажает – это и есть предмет настоящей статьи.

II. Парадокс без теоремы

В 1959 году Н. Хомский опубликовал рецензию на книгу Б.Ф. Скиннера «Вербальное поведение» [Chomsky 1959]. Это стало одним из редких примеров в истории науки того, что критический текст уничтожил не просто книгу, но целую исследовательскую программу. Однако бихевиоризм Скиннера был лишь ближайшей мишенью. Атака была шире: Хомский отвергал саму идею, что язык поддается вероятностному описанию. Аргумент сформулирован с намеренной резкостью. Рассмотрим два предложения:

«Бесцветные зеленые идеи яростно спят»; «Яростно спите идеи зеленый бесцветных». Хомский утверждал, что статистические модели его времени обращались бы с обоими как с одинаково невероятными, поскольку ни то, ни другое не представлено в языковых корпусах. Однако первое предложение построено грамматически правильно, в отличие от второго. Значит, грамматическая правильность не является вопросом вероятности. Никакая система, оперирующая статистическими связями между элементами, не способна уловить то, что отличает первое предложение от второго, – иерархическую синтаксическую структуру, управляемую конечным набором рекурсивно применяемых правил.

Приведенный аргумент имел прямые последствия для судьбы вероятностной лингвистики на несколько десятилетий вперед. Марковские модели языка – системы, предсказывающие следующее слово по нескольким предыдущим, – были дискредитированы не как несовершенный инструмент, нуждающийся в улучшении, а как принципиально несостоятельный подход, неспособный по своей природе справиться с иерархическими зависимостями синтаксиса.

Чтобы понять, почему это важно, необходимо пояснить, что именно означает марковское свойство. Система обладает марковским свойством, если для предсказания ее следующего состояния достаточно знать только текущее, предшествующая история несущественна. Применительно к языку это означает: следующее слово предсказывается только по нескольким непосредственно предшествующим словам, без учета сказанного в начале предложения, абзаца или разговора. Марковская модель языка успешно справляется с локальными статистическими закономерностями (типичными сочетаниями слов, характерными идиомами), но принципиально не улавливает зависимости между элементами, разделенными произвольным расстоянием. Приведем в качестве примера высказывание: «Человек, которого я встретил вчера на улице, где мы гуляли с детьми, оказался моим старым другом». В этом предложении согласование подлежащего и сказуемого происходит через длинную цепочку придаточных. Именно такая дальняя зависимость невидима для марковской модели с фиксированным локальным контекстом. Хомский был прав: это системное ограничение, а не дефект реализации.

Шестьдесят лет спустя история ответила, но не теоремой. Она ответила эмпирически, что значительно неудобнее: системы, в

основе которых находится именно вероятностное предсказание следующего токена, освоили рекурсивную продуктивность без рекурсивной архитектуры. Генеративный предобученный трансформер (generative pre-trained transformer, GPT) порождает синтаксически корректные вложенные структуры произвольной глубины, не имея в своей архитектуре ничего, что соответствовало бы хомскианским правилам перезаписи. Это не победа Скиннера – его программа действительно несостоятельна по основаниям, на которые указал Хомский. Речь идет о более странной ситуации, об эмпирическом снятии спора без теоретического объяснения того, как именно это стало возможным.

Чтобы понять природу такого парадокса, необходимо понять архитектуру трансформера, и не детали инженерной реализации, а концептуальное устройство. Ключевое различие, которое в данном случае представляется существенным, состоит в том, что трансформер является немарковским по контексту, но марковским по архитектуре слоев.

Механизм внимания, предложенный А. Васвани и соавторами в 2017 году [Vaswani et al. 2017], решает проблему дальних зависимостей радикально: при вычислении представления каждого токена модель одновременно учитывает все предшествующие токены в контексте, взвешивая их по степени релевантности. Не фиксированное окно нескольких соседних слов, а динамически вычисляемое внимание ко всей доступной истории. Т. Парр, Дж. Пеццуло и К. Фристон формулируют это точно: трансформеры успешны именно потому, что выходят за пределы марковского приближения, они используют расширенный контекст в случаях, в которых марковская модель была бы вынуждена его усечь [Parr et al. 2025a; Parr et al. 2025b]. Зависимость между подлежащим и сказуемым через длинную цепочку придаточных – это именно тот случай, в котором механизм внимания делает то, чего не могла сделать марковская модель.

Однако внутри каждого слоя трансформера вычисления строго локальны и детерминированы. Каждый слой получает векторные представления токенов и производит над ними преобразования по фиксированным правилам, передавая результат следующему слою. Обратной связи между слоями нет, информация движется только вперед. В этом смысле архитектура остается марковской: каждый следующий слой обрабатывает только текущее состояние представлений, не обращаясь к истории вычислений как самостоя-

тельному объекту[3]. Немарковский охват контекста реализован не через рекурсию и не через память в техническом смысле, а через механизм внимания, который на каждом шаге (слое) заново вычисляет релевантность всего доступного прошлого.

Чтобы ошибочно не принять за основу рекурсивности архитектуру предшественников трансформеров – рекуррентных нейросетей, – необходимо разграничить два понятия, которые в литературе нередко смешивают: *рекуррентность* и *рекурсивность*. Рекуррентные нейронные сети – архитектурный тип, в котором скрытое состояние сети после каждого шага обработки передается на следующий шаг как дополнительный вход. Шаг в данном случае – один элемент входной последовательности: одно слово, один токен, один символ. Обрабатывая первое слово, сеть порождает скрытое состояние – вектор, несущий память о прочитанном; обрабатывая второе слово, она принимает на вход и само слово, и это скрытое состояние; и таким образом обрабатывается вся последовательность. Обратная связь в данном случае темпоральна: она осуществляется не между слоями сети в пространстве архитектуры, а между шагами обработки во времени. Это архитектурный принцип субстрата.

Рекурсия в понимании Хомского – принципиально иное. Это свойство формального описания, правило, применимое к собственному результату и порождающее тем самым вложенные структуры произвольной глубины (см.: [Mikhailov 2024a]). Придаточное предложение внутри придаточного – рекурсия в этом смысле. Стандартные трансформеры не рекуррентны в первом смысле[4]: в них нет темпоральной обратной связи между шагами обработки. Но они освоили рекурсивную продуктивность во втором смысле, как функциональное свойство порождать и распознавать вложенные структуры. Субстрат и функция разошлись, и именно этого Хомский не предвидел.

---

[3] Слои трансформера соединены остаточными связями (residual connections): вход слоя или подслоя напрямую добавляется к результату его преобразования. Поэтому информация о более ранних состояниях присутствует в накопленном виде – как часть векторного представления на входе каждого слоя, но не как отдельный объект обработки. Это не отменяет принципиального отличия от рекуррентной архитектуры, в которой скрытое состояние является явным объектом обработки.

[4] Существуют варианты с рекуррентностью, в частности Universal Transformer [Dehghani et al. 2019]. Однако они остаются за пределами доминирующей архитектурной парадигмы.

Его критика языковых моделей 2023 года воспроизводит аргументы 1959 года без поправки на изменившиеся обстоятельства. Языковые модели, как он утверждает, не имеют подлинной грамматической компетенции, они лишь статистически аппроксимируют ее поверхностные проявления. Это может оказаться верным, но именно как философское утверждение о природе компетенции, а не как техническое утверждение об ограничениях архитектуры. Техническое утверждение опровергнуто практикой.

Однако парадокс остается. И важно быть точным в том, что именно остается открытым и почему. Можно выделить два способа интерпретации эмпирического успеха трансформеров. Первый: рекурсивная продуктивность не требует рекурсивной архитектуры, а значит, Хомский переоценил роль детерминирующих грамматических правил, и мышление устроено иначе, чем он полагал. Второй: трансформеры не обладают подлинной рекурсивной компетенцией, они лишь имитируют ее в пределах обучающей выборки, и за ее границами эта имитация систематически разрушается.

Попытки разрешить данный вопрос эмпирически были и продолжаются. Тесты на систематическую генерализацию, такие как SCAN и COGS [Lake, Baroni 2018; Kim, Linzen 2020], проверяют способность модели применить выученные правила к принципиально новым комбинациям, не встречавшимся при обучении. Результаты неоднозначны: трансформеры справляются значительно лучше ранних архитектур, но систематически проваливаются на определенных классах задач, требующих переноса структурного правила в новый контекст. Однако полная верификация наталкивается на принципиальную методологическую трудность. Мы не можем гарантировать того, что тестовые примеры действительно находятся за пределами обучающей выборки в смысловом, а не только лексическом отношении. Модель, обученная на большом корпусе, могла усвоить структурное правило не через его понимание, а через исчерпывающее покрытие его частных случаев. И тест, призванный это разграничить, сам оказывается под вопросом. Это не тупик, а продуктивная методологическая проблема: ее решение требует синтетических тестов с контролируемым распределением обучающих данных – области, которая сегодня активно развивается.

Показательно, что аналогичным тестам подвергались и люди. Результаты оказались неудобными для обеих сторон спора. Б.М.

Лейк и М. Бэрони, создавая SCAN [Lake, Baroni 2018], тестировали людей параллельно: люди справлялись значительно лучше моделей, но в случае наиболее сложных комбинаций, с глубокой вложенностью, тоже ошибались. Г.Ф. Маркус и соавторы в классической серии экспериментов с искусственными грамматиками показали, что и семимесячные младенцы, и взрослые обобщают абстрактные правила на новый материал [Marcus et al. 1999], но способность к обобщению зависит от поверхностного сходства стимулов и не является абсолютной. Иными словами, разрыв между большими языковыми моделями и человеком на этих тестах количественный, а не качественный; вопрос о природе этого количественного разрыва остается открытым. В данном случае принципиально: если систематическая генерализация у человека тоже опирается не на встроенный символический процессор, а на статистически усвоенные абстрактные паттерны, то аргумент в пользу врожденной грамматики теряет главную опору.

Между тем существует типологическое основание, позволяющее выйти за пределы агностицизма, не претендуя на закрытие вопроса. И трансформер, и мозг в рамках предиктивного кодирования реализуют одну и ту же абстрактную операцию: итеративное уточнение представления через минимизацию расхождения между предсказаниями модели и входными данными в многомерном пространстве. У Фристона это минимизация вариационной свободной энергии – функционала, включающего дивергенцию Кульбака – Лейблера между вариационным распределением и генеративной моделью как меру их расхождения. В трансформере механизм внимания сопоставляет запрос с ключами в пространстве векторных представлений и преобразует полученные оценки совместимости в нормированное распределение весов по элементам контекста. Эти веса используются для построения контекстуально обусловленного представления каждого элемента. Таким образом, внимание реализует принципиально то же самое: оценивает релевантность доступных данных и уточняет представление, минимизируя расхождение между модельным предсказанием и входными данными. Это операция того же рода, что и дивергенция Кульбака – Лейблера у Фристона, – локальное сравнение распределений, но без явного обращения к понятию свободной энергии. Парр, Пеццуло и Фристон прямо указывают на то, что авторегрессионные модели и глубокие временны́е иерархии математически эквивалентны по

итоговому распределению при условии маргинализации скрытых переменных, и разница между ними состоит не в модели мира, а в способе осуществления *инференса*[5] [Parr et al. 2025b]. Если принять это как типологическое основание, то оба процесса суть инстанции одного вычислительного типа – иерархической редукции неопределенности. И вопрос «Мыслит ли языковая модель?» получает не ответ, но точную переформулировку. Достаточна ли редукция неопределенности без воплощенной верификации для того, что мы называем мышлением?

Этот парадокс философски значим, а не просто технически интересен, поскольку связан с тезисом, сформулированным в предыдущем разделе. Если мышление не таково, каким его наблюдает сознание, то аргумент от интроспекции («я знаю, что такое подлинная грамматическая компетенция, потому что я ею обладаю») не имеет доказательной силы. Возможно, то, что мы называем рекурсивной компетенцией, и есть именно функциональное свойство (способность порождать и распознавать вложенные структуры), а не некий внутренний механизм особого рода, доступный только биологическим системам. Возможно, нет. Но для разрешения этого вопроса нужны аргументы иного рода. Именно к ним мы обращаемся в следующем разделе.

III. Геометрия разрыва

Сравнение мозга и языковой модели легко превращается в риторический прием: или в апологию машины («смотрите, как похоже»), или в ее дискредитацию («смотрите, как непохоже»). Во избежание обоих соблазнов необходимо зафиксировать разрыв точно, не метафорически, а структурно. Разрыв существует на четырех уровнях: энергетическом, вычислительном, семантическом и темпоральном. На каждом из них обнаруживается не просто количественное различие, но качественно иной способ организации информационного процесса.

Энергетический уровень

Человеческий мозг потребляет около 20 ватт. Обучение большой языковой модели требует мегаватт электроэнергии на протяжении недель, и это без учета инфраструктуры охлаждения и поддержки. Даже *инференс* – генерация одного ответа уже обученной модели

[5] Инференс (от англ. *inference* – «вывод») – генерация ответа обученной моделью.

– потребляет на несколько порядков больше энергии, чем эквивалентная когнитивная операция у человека. Это не инженерная досада, поддающаяся оптимизации, а симптом принципиального архитектурного различия.

Биологическая энергоэффективность достигается через два механизма, которые в цифровых архитектурах отсутствуют. Первый – *спайковое кодирование*. Нейрон генерирует электрический импульс только в случае, если его мембранный потенциал превышает пороговое значение; в остальное время он молчит. Вычисление реализовано через редкую, асинхронную активность: в каждый момент времени активна лишь малая доля нейронов сети. Этот факт сочетается с гипотезой, что мозг вычисляет только отклонения от ожидаемого (ошибки предсказания), а не входной поток целиком. Второй механизм – *массивный параллелизм* без центрального процессора. В мозге нет узкого горлышка фон Неймана, на компьютерах архитектуры которого пока еще реализуются нейросети и в котором разделяются память и вычисление. Синаптическая пластичность мозга означает, что память и обработка реализованы в одном и том же субстрате, локально и одновременно.

Цифровая архитектура устроена противоположным образом. Трансформер при каждом шаге генерации активирует фиксированный объем вычислений, будь то все веса сети в плотных архитектурах или их подмножество в архитектурах «смесь экспертов» (mixture-of-experts), независимо от того, насколько тривиален входной сигнал. Обновление весов при обучении через обратное распространение ошибки требует глобального прохода по всей сети: локальное обновление, которое было бы биологически правдоподобным, в современном генеративном ИИ пока не реализовано. Все это накладывается на физическую архитектуру фон-неймановских компьютеров, с ее разделением памяти и обработки по разным блокам. Нейроморфные архитектуры, в частности чипы Loihi компании Intel, TrueNorth компании IBM, предпринимают попытку воспроизвести спайковую разреженность и асинхронность в кремниевом субстрате, но пока остаются далеко за пределами масштаба, необходимого для языковых задач. Мемристоры – элементы с изменяемым сопротивлением, сохраняющим свое значение после снятия напряжения, – представляют собой аппаратный аналог синаптической пластичности (память и вычисление в одном компоненте). Их интеграция в нейроморфные

схемы теоретически могла бы сократить энергетический разрыв на несколько порядков. Однако это пока область активных исследований, а не готовых решений.

Энергетический разрыв важен не сам по себе, но как индикатор: он свидетельствует о том, что мозг решает задачу принципиально иным способом. Не плотными матричными умножениями по всем параметрам, а точечной реакцией на отклонение от модели. Это различие вычислительное, и оно напрямую связано со следующим уровнем анализа.

Вычислительный уровень

Предиктивное кодирование, разработанное Р. Рао и Д. Баллардом [Rao, Ballard 1999] и реинтерпретированное Фристоном в рамках принципа свободной энергии, предполагает, что мозг есть иерархическая машина предсказаний (см.: [Mikhailov 2024b]). Каждый уровень иерархии порождает предсказание о состоянии уровня ниже и получает обратно только ошибку предсказания, т.е. расхождение между ожидаемым и действительным. Если предсказание оправдывается, сигнал подавляется: нет необходимости передавать то, что уже известно. Высший уровень не пересчитывает предсказание при каждом входном сигнале, а удерживает текущую гипотезу и обновляет ее лишь при поступлении сигнала об ошибке от низшего. Вычислительная экономия достигается именно тем, что в большинстве случаев ошибка мала: верхние уровни иерархии остаются относительно «тихими», пока нижние поглощают предсказуемый поток. Вверх по иерархии движется только отличное от ожидаемого.

Это означает, что вычислительная нагрузка распределена весьма неравномерно: подавляющая часть сенсорного потока поглощается на нижних уровнях иерархии и не достигает вершины иерархии. То, что достигает, – результат многоступенчатой фильтрации, сжатия и интерпретации. Мозг не обрабатывает реальность; он обрабатывает отклонения своей модели от реальности.

Обратное распространение ошибки в нейронных сетях поверхностно напоминает этот принцип: в обоих случаях ошибка служит сигналом для обновления модели. Однако имеется принципиальное различие. Обратное распространение ошибки (backpropagation) – глобальная процедура: ошибка на выходе сети должна быть распространена обратно через все слои и исправлена одновременно, что требует хранения промежуточных

активаций и глобального прохода. В биологическом предиктивном кодировании обновление локально: каждый уровень иерархии обновляет свои параметры на основе локальной ошибки предсказания, без необходимости знать о том, что происходит на других уровнях. Это не просто вычислительное удобство, а условие возможности обучения в реальном времени: мозг обновляет модель непрерывно, в процессе взаимодействия с миром, а не в отдельной фазе предшествующего обучения, как искусственные нейросети.

Трансформер предварительно обучается на фиксированном корпусе, и затем его применяют к новым данным без обновления весов. Это означает, что его модель мира статична в момент применения: обновления в реальном времени в ответ на новый опыт не происходит. Мозг, напротив, находится в состоянии непрерывной калибровки: каждое взаимодействие с миром уточняет модель.

Активный вывод в концепции Фристона – расширение предиктивного кодирования на моторику – описывает действие как способ подтверждения предсказаний: организм не только обновляет модель в ответ на сенсорные данные, но и активно изменяет мир таким образом, чтобы привести его в соответствие с ожиданиями. Это замкнутый цикл восприятия и действия, принципиально отличный от разомкнутой архитектуры языковой модели.

Семантический уровень

Трансформер оперирует эмбеддингами – непрерывными векторами высокой размерности, посредством которых слова и токены представлены как точки многомерного пространства. Пространство задано архитектурой. Это, как правило, евклидово пространство фиксированной размерности. Однако расположение токенов в нем определяется в процессе обучения таким образом, что семантически близкие единицы оказываются геометрически близкими векторами. Непрерывное векторное пространство свидетельствует о движении когнитивной науки по третьей оси приближения к биологически правдоподобным моделям – от дискретных символов к непрерывным представлениям. Смысл перестает быть атомарным, он становится геометрическим отношением на многомерном континууме. Джеффри Хинтон настаивал именно на этом: распределенное представление принципиально отличается от символьного тем, что смысл единицы определяется

ее положением в пространстве относительно остальных единиц, а не ее тождеством себе[6].

Однако возникает разрыв, который геометрия эмбеддингов не закрывает. Семантика языковой модели замкнута на текстовый корпус: отношения между векторами отражают статистические закономерности совместной встречаемости слов в текстах, порожденных людьми. «Яблоко» для модели – вектор, определенный своими отношениями с векторами «фрукт», «красный», «сад», «Ньютон», «укус» и ничем иным. Для человека «яблоко» – это также вкус, запах, тяжесть в руке, специфическое ощущение при надкусывании, воспоминание о конкретном дереве в конкретном месте. Семантика укоренена в сенсомоторном опыте, в воплощенном взаимодействии с миром.

Различие между семантикой из корреляций и семантикой из воплощения имеет верифицируемые функциональные следствия. Языковые модели систематически ошибаются в задачах, требующих пространственного, физического или каузального рассуждения, укорененного в опыте взаимодействия с материальным миром. Они компетентны в пространстве слов о вещах и менее компетентны в ситуации, если слова отсылают к свойствам вещей, недоступным через текст. Это не случайные ошибки, поддающиеся устранению через масштабирование, а структурный след архитектурного решения.

В этом контексте полезно обратиться к аргументу Сёрля [Searle 1980], но не для того, чтобы его принять, а для того, чтобы указать на его логическую уязвимость. «Китайская комната» доказывает, что человек внутри комнаты, манипулирующий символами по инструкции, не понимает китайского языка. Это верно. Но аргумент совершает скачок: из того, что элемент системы его не понимает, делается вывод, что система в целом его не понимает. Между тем понимание, если оно вообще существует как функциональное свойство, является свойством системы в целом, а не ее отдельных компонентов. Каждый отдельный нейрон человеческого мозга тоже не понимает ничего, ни китайского языка, ни того, что такое языковая модель. Понимание, если оно реализовано в нейронном субстрате, обнаруживается на уровне системной организации, а не на уровне элемента. Сёрль, таким образом, демонстрирует не то,

[6] *Hinton G.E.* What Is Understanding? Keynote address at the inaugural IASEAI conference (Paris, February 6, 2025) // International Association for Safe and Ethical AI. July 8, 2025. – URL: https://youtu.be/6fvXWG9Auyg.

что система не понимает, а то, что понимание не локализовано в элементе, что само по себе не является ни новым, ни опровергающим тезисом. Языковая модель как система обладает пониманием в функциональном смысле; семантический разрыв, описанный выше, указывает не на его отсутствие, а на его специфическую природу: это понимание, укоренено в корреляциях текстового корпуса, а не в воплощенном взаимодействии с миром. Разрыв реален, но это разрыв в способе референции, а не в наличии или отсутствии понимания как такового.

Темпоральный уровень

Четвертый уровень разрыва наименее очевиден, но, возможно, наиболее фундаментален. Мозг существует во времени не только как физический объект, но как вычислительная система. Его модель мира непрерывно обновляется; каждый момент восприятия изменяет параметры системы. Биография организма вписана в структуру его нейронных связей в буквальном, нейропластическом смысле. Память реализуется не как архив, а как измененная структура: прошлое присутствует не как сохраненная запись, а как модифицированная готовность реагировать.

Языковая модель, напротив, атемпоральна в момент применения. Ее веса фиксированы, дата отсечения обучающих данных есть горизонт ее «биографии». В пределах одной сессии она способна удерживать контекст разговора, но это не обучение и не обновление модели, а использование контекстного окна как временной рабочей памяти, которая обнуляется по завершении сессии. Парр, Пеццуло и Фристон проводят важное различие: трансформер кеширует все прошлые события в пределах контекстного окна [Parr et al. 2025a; Parr et al. 2025b]. Это охват без глубины. Мозг строит иерархически сжатые резюме на разных временных шкалах, от миллисекунд до десятилетий. Это глубина без необходимости хранения необработанных данных.

Указанное различие напрямую связано с природой галлюцинаций языковых моделей. Галлюцинация, или генерация уверенных, но фактически ложных утверждений, есть статистическая неизбежность неагентной системы, т.е. лишенной механизма верификации через взаимодействие с реальностью. Мозг (или агентная система) располагает таким механизмом: тело взаимодействует с миром, ошибка предсказания корректирует модель. Языковая модель оперирует в замкнутом пространстве статистических

отношений между текстами, и, если наиболее вероятное продолжение не совпадает с реальным положением дел, модель этого не знает. Перед нами не дефект реализации, а структурное следствие атемпоральности и бестелесности архитектуры.

Совокупность четырех уровней разрыва позволяет сформулировать точный ответ на вопрос о дистанции между биологическим и искусственным интеллектом. Это не дистанция по одной шкале («умнее/глупее» или «ближе/дальше к человеку»), а многомерное расхождение в способе организации информационного процесса: синхронное против асинхронного, локальное против глобального, воплощенное против дистиллированного, темпоральное против атемпорального. Трансформеры достигли впечатляющих результатов, двигаясь принципиально иным путем, чем биологическая эволюция. Назвать такой путь имитацией интеллекта или интеллектом иного рода – значит принять определение, которое в любом случае будет спорным. Но вопрос определения не является самой важной проблемой. Важнее другое: тот факт, что люди оказываются способны к содержательному интеллектуальному общению с языковыми моделями (к совместному рассуждению, взаимной коррекции, продуктивному несогласию), был бы необъясним, если бы между двумя типами интеллекта не существовало типологического родства. Это родство, как мы показали, состоит в общей операции иерархической редукции неопределенности. Разрыв реален, но он находится внутри одного рода, а не между разными.

Заключение

Изложенное выше подчинено движению по одной траектории: от философской проблемы – через архитектурный парадокс – к структурному разрыву. Что именно на этом пути установлено?

Первый вывод – дескриптивный. Разрыв между биологическим и искусственным интеллектом не является одномерным и не сводится к количественному отставанию по единой шкале. Он структурирован на четырех уровнях: энергетическом, вычислительном, семантическом и темпоральном. На каждом из них можно обнаружить не просто различие в степени, но качественно иной способ организации информационного процесса: синхронное локальное обновление против асинхронного глобального, воплощенная референция против замкнутой статистической, темпоральная глубина против охвата без глубины. Измерение этого разрыва –

предварительное условие любого содержательного разговора о природе машинного интеллекта, и без него вопрос «Мыслит ли языковая модель?» остается риторическим.

Второй вывод – философский. Сознание не является ни онтологически необходимым условием мышления, ни эпистемологически достаточным критерием его верификации в биологических системах, а тем более в искусственных. Лейбниц сформулировал это как методологический принцип: перспектива первого лица и перспектива третьего лица взаимно непереводимы, и ни одна из них не дает привилегированного доступа к тому, что такое мышление по существу. Когнитивная наука в разные исторические периоды, от Гельмгольца до Фристона, последовательно подтверждала этот принцип применительно к биологическим системам. Появление языковых моделей распространяет его на системы искусственные: интроспективная очевидность не может служить ни аргументом против машинного мышления, ни аргументом в его пользу. Вопрос требует других оснований – функциональных и архитектурных, а не феноменологических.

Третий вывод – технический, но с философскими последствиями. Рекурсивная продуктивность является функциональным следствием иерархически организованного предсказания, а не свойством конкретной архитектуры субстрата. Именно поэтому трансформеры – системы, не имеющие ни встроенной рекурсивной грамматики, ни рекуррентной обратной связи, – освоили порождение и распознавание вложенных структур произвольной глубины. Это снимает спор Хомского со Скиннером, и не в пользу одной из сторон, а как вопрос, поставленный в неверных координатах: противопоставление правило-управляемой рекурсии и статистической аппроксимации оказывается ложной дихотомией, если иерархически организованное предсказание реализует рекурсивную продуктивность третьим путем, не предусмотренным ни одной из сторон. Более того, этот третий путь, по-разному реализованный в трансформерной архитектуре искусственного интеллекта и в биологических системах, демонстрирует один и тот же вычислительный тип, а именно – иерархическую редукцию неопределенности, что делает вопрос о машинном мышлении не риторическим, а содержательно открытым: достаточна ли такая редукция без воплощенной верификации для того, чтобы считаться мышлением?

Но остается и еще один вопрос, который не нашел решения в пределах настоящей статьи. Однако его лучше сформулировать , чем оставить имплицитным. Не является ли понимание мышления как иерархической редукции неопределенности редукционистским в неприемлемом смысле? Разве функция творческого, эстетического, религиозного мышления сводится к минимизации расхождения между предсказанием и реальностью? Принцип свободной энергии предлагает утвердительный ответ, при условии, что «реальность» для биологической системы включает в себя собственные психологические состояния, социальные ожидания и культурные горизонты как часть генеративной модели мира. Художник, создающий произведение, минимизирует расхождение между тем, что есть, и тем, что должно быть согласно его модели; зритель, воспринимающий его, обновляет собственную. Редукция неопределенности в этом смысле описывает не содержание мышления, а его вычислительную форму, хотя граница между формой и содержанием может оказаться эпистемической, а не онтологической. Типологической эта характеристика является не потому, что форма отделена от содержания, а потому, что она инвариантна относительно субстрата: один и тот же вычислительный тип реализуется в нейронном субстрате и в кремниевом, с воплощением и без него. Именно эта инвариантность делает иерархическую редукцию неопределенности характеристикой интеллекта как такового, а не его частной биологической формы.

За пределами настоящего исследования остаются вопросы, которые оно делает более точными и операционализируемыми. Является ли воплощение необходимым условием референции или лишь одним из возможных путей к ней? При каких архитектурных условиях воплощенное взаимодействие с миром и непрерывное обновление модели могут быть реализованы в искусственных системах? Если рекурсивная продуктивность не требует ни рекурсивной архитектуры, ни сознания, чего именно она требует, т.е. что является сущностным ядром биологических и искусственных генеративных моделей? Эти вопросы остаются открытыми, но не по причине их неразрешимости, а потому что ответы могут и должны быть получены математически и эмпирически. История когнитивной науки дает основания полагать, что математики по дороге к ответу сформулируют их точнее, чем философия формулировала их прежде.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Лейбниц 1982 – *Лейбниц Г.В.* Монадология / пер. с фр. Е.А. Боброва // *Лейбниц Г.В.* Сочинения: в 4 т. Т. 1 / ред. и сост. В.В. Соколов. – М.: Мысль, 1982. С. 413–429.

Chomsky 1959 – *Chomsky N.* A Review of B.F. Skinner's Verbal Behavior // Language. 1959. Vol. 35. No. 1. P. 26–58.

Dehghani et al. 2019 – *Dehghani M., Gouws S., Vinyals O., Uszkoreit J., Kaiser Ł.* Universal Transformers // Proceedings of the 7th International Conference on Learning Representations. 2019. arXiv:1807.03819.

Friston et al. 2017 – *Friston K., Parr T., de Vries B.* The Graphical Brain: Belief Propagation and Active Inference // Network Neuroscience. 2017. Vol. 1. No. 4. P. 381–414.

Helmholtz 1867 – *Helmholtz H. von.* Handbuch der physiologischen Optik. – Leipzig: Leopold Voss, 1867.

Juschkewitsch, Kopelewitsch 1988 – *Juschkewitsch A.P., Kopelewitsch Ju.Ch.* La correspondance de Leibniz avec Goldbach /// Studia Leibnitiana. 1988. Bd. 20. H. 2. S. 175–189.

Kim, Linzen 2020 – Kim N., Linzen T. COGS: A Compositional Generalization Challenge Based on Semantic Interpretation // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020. P. 9087–9105. arXiv:2010.05465.

Lake, Baroni 2018 – Lake B.M., Baroni M. Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks // Proceedings of the 35th International Conference on Machine Learning. 2018. Vol. 80. P. 2873–2882. arXiv:1711.00350.

Marcus et al. 1999 – *Marcus G.F., Vijayan S., Bandi Rao S., Vishton P.M.* Rule Learning by Seven-Month-Old Infants // Science. 1999. Vol. 283. No. 5398. P. 77–80.

Mikhailov 2024a – *Mikhailov I.F.* The Concept of Recursion in Cognitive Studies. Part I: From Mathematics to Cognition // Philosophical Problems of IT and Cyberspace. 2024. No. 1. P. 58–76.

Mikhailov 2024b – *Mikhailov I.F.* The Concept of Recursion in Cognitive Studies. Part II: From Turing to Bayes to Consciousness // Philosophical Problems of IT and Cyberspace. 2024. No. 2. P. 4–22.

Parr et al. 2025a – *Parr T., Pezzulo G., Friston K.* Beyond Markov: Transformers, Memory, and Attention // Cognitive Neuroscience. 2025. Vol. 16. No. 1–4. P. 5–23.

Parr et al. 2025b – *Parr T., Pezzulo G., Friston K*. Closing the Box // Cognitive Neuroscience. 2025. Vol. 16. No. 1–4. P. 43–48.

Rao, Ballard 1999 – *Rao R.P.N., Ballard D.H.* Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects // Nature Neuroscience. 1999. Vol. 2. No. 1. P. 79–87.

Searle 1980 – *Searle J.R.* Minds, Brains, and Programs // Behavioral and Brain Sciences. 1980. Vol. 3. No. 3. P. 417–424.

Vaswani et al. 2017 – *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I.* Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.

REFERENCES

Chomsky N. (1959) A Review of B.F. Skinner's Verbal Behavior. *Language.* Vol. 35, no. 1, pp. 26–58.

Dehghani M., Gouws S., Vinyals O., Uszkoreit J., & Kaiser Ł. (2019) Universal Transformers. In: *Proceedings of the 7th International Conference on Learning Representations.* arXiv:1807.03819.

Friston K., Parr T., & de Vries B. (2017) The Graphical Brain: Belief Propagation and Active Inference. *Network Neuroscience.* Vol. 1, no. 4, pp. 381–414.

Helmholtz H. von (1867) *Handbuch der physiologischen Optik*. Leipzig: Leopold Voss (in German).

Juschkewitsch A.P. & Kopelewitsch Ju.Ch. (1988) La correspondance de Leibniz avec Goldbach. *Studia Leibnitiana.* Vol. 20, no. 2, pp. 175–189 (in French and Latin).

Kim N. & Linzen T. (2020) COGS: A Compositional Generalization Challenge Based on Semantic Interpretation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 9087–9105). arXiv:2010.05465.

Lake B.M. & Baroni M. (2018) Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In: *Proceedings of the 35th International Conference on Machine Learning.* Vol. 80, pp. 2873–2882. arXiv:1711.00350.

Leibniz G.W. (1982) Monadology (E.A. Bobrov, Trans.). In: Leibniz G.W. *Works in 4 Vols.* (Vol. 1, pp. 413–429). Moscow: Mysl' (Russian translation).

Marcus G.F., Vijayan S., Bandi Rao S., & Vishton P.M. (1999) Rule Learning by Seven-Month-Old Infants. *Science.* Vol. 283, no. 5398, pp. 77–80.

Mikhailov I.F. (2024a) The Concept of Recursion in Cognitive Studies. Part I: From Mathematics to Cognition. *Philosophical Problems of IT and Cyberspace.* No. 1, pp. 58–76.

Mikhailov I.F. (2024b) The Concept of Recursion in Cognitive Studies. Part II: From Turing to Bayes to Consciousness. *Philosophical Problems of IT and Cyberspace.* No. 2, pp. 4–22.

Parr T., Pezzulo G., & Friston K. (2025a) Beyond Markov: Transformers, Memory, and Attention. *Cognitive Neuroscience.* Vol. 16, no. 1–4, pp. 5–23.

Parr T., Pezzulo G., & Friston K. (2025b) Closing the Box. *Cognitive Neuroscience.* Vol. 16, no. 1–4, pp. 43–48.

Rao R.P.N. & Ballard D.H. (1999) Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects. *Nature Neuroscience.* Vol. 2, no. 1, pp. 79–87.

Searle J.R. (1980) Minds, Brains, and Programs. *Behavioral and Brain Sciences.* Vol. 3, no. 3, pp. 417–424.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., & Polosukhin I. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems.* Vol. 30, pp. 5998–6008.