Философия науки и техники
2026. Т. 31. № 1. С. 111–125
УДК 168

Philosophy of Science and Technology
2026, vol. 31, no. 1, pp. 111–125
DOI: 10.21146/2413-9084-2026-31-1-111-125

М.А. Сущин

Большие языковые модели, поддержанные средой, а не расширенные

Сущин Михаил Александрович – кандидат философских наук, старший научный сотрудник. Институт научной информации по общественным наукам РАН. Российская Федерация, 117997, г. Москва, Нахимовский просп., д. 51/21; e-mail: sushchin@bk.ru

В статье рассматривается возможность расширенного искусственного интеллекта, основывающаяся на параллелях между предполагаемыми случаями когнитивного расширения у человека в соответствии с гипотезой расширенного разума и большими языковыми моделями, способными генерировать ответы на основе информации из внешних источников. Поскольку источником вдохновения для идеи расширенного искусственного интеллекта выступает известная гипотеза Кларка – Чалмерса, в статье оцениваются аргументы как в пользу, так и против возможности когнитивного расширения у человека. Автор подчеркивает важные функциональные отличия между кодируемыми мозгом диспозициональными убеждениями и информацией на внешних носителях. Делается вывод о проблематичности идеи расширения разума в мир без тесной интеграции между мозгом и внешними носителями информации. Указывается, что аналогичные проблемы могут быть характерны и для больших языковых моделей, оснащенных технологией дополненной поиском генерации, но не находящихся в тесном непрерывном информационном контакте с внешним хранилищем информации. В связи с этим отмечается, что для осмысления отношения агентов к внешним информационным средам могут потребоваться более умеренные гипотезы, находящиеся в согласии с текущими методологическими посылками когнитивных наук.

Ключевые слова: большие языковые модели, гипотеза расширенного разума, расширенный искусственный интеллект, дополненная поиском генерация, когнитивные науки

© Сущин М.А., 2026

Введение

Широко отмечается, что одной из отличительных особенностей человеческого интеллекта является внедрение в познавательные практики артефактов и орудий материальной культуры. Этот важнейший аспект человеческого познания активно обсуждался в современной науке и философии, по крайней мере, со времен пионерских работ Л.С. Выготского. Вскоре после перевода сочинений Выготского на английский язык проблема отношения интеллекта и интеллектуальных орудий пережила всплеск популярности в современной когнитивной науке.

В хорошо известной посвященной понятию схемы работе коннекционисты Д. Рамелхарт, Дж. МакКлелланд, Дж. Хинтон и П. Смоленски дали ответ на следующий вызов для параллельных распределенных моделей процессов познания [Rumelhart et al., 1986]. Так, если, согласно постулатам коннекционизма, когнитивные процессы носят параллельный распределенный характер, приводя к решению определенной проблемы без использования явно определенных логических правил, то каким образом человеческий интеллект способен успешно заниматься наукой, математикой и логикой? Каким образом человек способен осуществлять сложные последовательные интеллектуальные операции? По предположению коннекционистов, существо дела здесь заключается в способности человека трансформировать сложные задачи в более легкие путем создания внешних репрезентаций, для работы с которыми наши базовые параллельные механизмы познания гораздо лучше оснащены. По мнению Рамелхарта и его соавторов, у человека есть три важнейшие способности, которые позволили ему преодолеть ограничение параллельно работающих механизмов разума для осуществления последовательных интеллектуальных операций. Таковы: (1) интерпретация входящей информации (сопоставление с образцом, pattern matching), (2) предвосхищение новых ситуаций и (3) умелое манипулирование окружающей средой [Ibid., p. 44–45]. Например, вместо выполнения сложного умножения исключительно в уме человек способен выписать подлежащие умножению числа на листке бумаги (одно над другим, как полагается при умножении столбиком) таким образом, что первый шаг решения можно будет легко увидеть. Выполнив первый шаг и записав результат, человек может использовать свои навыки распознавания паттернов повторно для нахождения следующего шага:

> Каждый цикл этого процесса включает в себя в первую очередь создание репрезентации путем манипуляции средой, затем обработку этой (действительной физической) репрезентации при помощи нашего хорошо отлаженного перцептивного аппарата, что ведет к дальнейшей модификации этой репрезентации. Действуя таким образом, мы сводим предельно абстрактную концептуальную проблему к последовательности операций, являющихся предельно конкретными и подходящими для наших способностей. <…> Это применимо в том числе к решению проблем в логике (например, силлогизмы), проблем в науке, инженерии и т.д. Эти созданные нами двойные навыки манипуляции средой и обработки среды позволяют нам сводить очень сложные проблемы к последовательности очень простых проблем [Ibid., p. 45–46].

В начале 1990-х гг. предположение о ведущей роли артефактов и среды в познавательной деятельности человека стало одним из ключевых постулатов складывающегося комплекса теорий воплощенного познания. Так, в противоположность идее, что цель зрительной обработки заключается в создании детальных репрезентаций зрительных сцен, новые концепции восприятия постулировали, что зрительная обработка имеет своим результатом только частичные репрезентации окружающей среды [Churchland et al., 1994]. Вместо создания детальных внутренних изображений зрительных сцен познающие используют мир как его собственную лучшую модель [Brooks, 1999], оставляя детали вовне и извлекая их по необходимости при помощи движений глаз, головы и тела [O'Regan, 1992]. Данная стратегия использования мира как внешней памяти настолько распространена, что, как было показано в экспериментах, субъекты часто предпочитают ее запоминанию, даже если последнее способствует скорейшему выполнению задачи [Ballard et al., 1997].

В схожем ключе когнитивный антрополог Э. Хатчинс выступил с идеей, что артефакты могут представлять собой важнейшую часть определенных когнитивных систем. Например, в традиционной микронезийской практике мореплавания моряки держат необходимые для путешествия знания в уме. В противоположность этому, «в Западной традиции физические артефакты стали репозиториями знания… Артефакты стали воплощать в себе знание, которое было бы чрезвычайно трудно репрезентировать в уме» [Hutchins, 1995, p. 96]. Более того, инструменты, подобные тем, что используются в современных западных навигационных практиках, способны образовывать более крупные распределенные когнитивные системы [Ibid., p. 287]. (Такова, например, навигационная команда, определяющая текущее положение корабля и вычисляющая его последующий курс [Ibid., p. 26].)

Это линия исследований нашла свое наиболее радикальное выражение в чрезвычайно смелой гипотезе расширенного разума, выдвинутой философами Э. Кларком и Д. Чалмерсом [Clark, Chalmers, 1998]. С этой точки зрения, привычные когнитивные орудия наподобие записных книжек могут быть носителями долговременных диспозициональных убеждений когнитивных агентов. Эта гипотеза привлекла к себе значительное внимание со стороны философов когнитивных наук.

Недавно дискуссии о расширенном разуме получили интересное продолжение. Как прекрасно известно, некоторое время назад на сцене появился чрезвычайно интересный тип когнитивных орудий: большие языковые модели (БЯМ) превратились из экспериментальных игрушечных моделей в полномасштабные доступные широкой публике рабочие модели, превосходно справляющиеся с задачами генерации, перевода и суммирования текстовой информации. С точки зрения когнитивных наук имеется необходимость в осмыслении природы этих инструментов. Нужно ли понимать их по аналогии с традиционными когнитивными орудиями в качестве возможного субстрата когнитивного расширения в смысле Кларка и Чалмерса? Либо же в качестве просто подпорок, строительных лесов для биологически основанных способностей?

Между тем в недавней работе Смарта, Клоуза и Кларка был предложен другой способ осмысления БЯМ [Smart et al., 2025]. По утверждению этих

авторов, БЯМ сами могут быть рассмотрены как информационные системы, функционирование которых может расширяться во внешнюю среду таким же образом, каким, по их мнению, может расширяться человеческое познание. То есть, как пишут Смарт и Клоуз, «в дополнение к пониманию БЯМ как потенциальных объектов расширения – как ресурс, который может быть встроен в когнитивные практики человека – мы можем понимать БЯМ как потенциальный субъект расширения – как нечто (entity), чему самому может пойти на благо работа расширенных информационных цепей» [Smart, Clowes, 2025, p. 3].

Этот подход к пониманию природы БЯМ был вдохновлен новой обещающей техникой, именуемой генерацией с дополнением через поиск (retrieval-augmented generation) и призванной способствовать решению проблемы галлюцинаций языковых моделей. В настоящей статье будет рассмотрена возможность расширенного искусственного интеллекта (ИИ). В первой части будут представлены стандартные аргументы в пользу гипотезы расширенного познания. Далее я рассмотрю, насколько оправданны рассуждения о возможности расширенного ИИ по аналогии с предполагаемыми случаями когнитивного расширения у человека[1].

Гипотеза расширенного разума и большие языковые модели

Я начну эту часть с краткого обсуждения гипотезы расширенного разума, поскольку именно она выступает в качестве концептуальной основы для интересующей нас идеи расширенного ИИ[2]. Итак, согласно тезису Кларка и Чалмерса, субстрат некоторых ментальных состояний (таких как долговременные воспоминания) может расширяться во внешнюю среду без посредничества каких-либо нейрокомпьютерных интерфейсов, только в силу надежного доступа к некоторому артефакту и его регулярному использованию. Таков постулируемый ими принцип равенства (эквивалентности, parity principle): «Если при выполнении некоторой задачи часть мира функционирует как процесс, который мы без колебаний признали бы частью когнитивного процесса, будь он осуществлен в голове, то эта часть мира (так мы утверждаем) является частью когнитивного процесса. Когнитивные процессы не (все) находятся в голове!» [Clark, Chalmers, 1998, p. 8].

Традиционно гипотеза Кларка – Чалмерса иллюстрируется при помощи мысленного эксперимента, в котором два человека узнают про выставку в Музее современного искусства в Нью-Йорке. Первый человек, девушка по имени Инга, является субъектом с нормально функционирующей памятью – через секунду она вспоминает, что музей расположен на 53-й улице, и идет на выставку. Второй человек, Отто, страдает от болезни Альцгеймера. Чтобы компенси-

---

1 Таким образом, обсуждение идеи расширенного искусственного интеллекта следует рассматривать во многом как ответвление дискуссии о расширенном естественном интеллекте.

2 Смарт, Клоуз и Кларк не называют идею расширенного ИИ гипотезой, и в этой статье я буду следовать их терминологии. Однако остается неясным, почему обсуждаемая ими возможность не называется гипотезой, тогда как идея расширенного разума традиционно именуется таковой.

ровать последствия болезни, Отто вынужден использовать записную книжку, которую он берет с собой, куда бы ни пошел. Соответственно, узнав, что в Музее современного искусства проходит выставка, Отто приходится искать адрес в своей записной книжке. Согласно Кларку и Чалмерсу «Отто полагал, что музей находится на 53-й улице, даже прежде, чем он обратился к своему дневнику. Во всех релевантных отношениях оба случая всецело идентичны: для Отто записная книжка играет ту же самую роль, которую память играет для Инги. Информация из записной книжки функционирует точно таким же образом, как информация, составляющая обычное неосознаваемое убеждение; просто так получается, что эта информация находится за границами кожи» [Clark, Chalmers, 1998, p. 13].

Если следовать наиболее распространенной альтернативной интерпретации, отмечают Кларк и Чалмерс, неверно приписывать Отто убеждения об адресе музея. Скорее, он убежден, что адрес музея можно найти в его записной книжке. В ответ авторы заявляют, что эта точка зрения упускает из виду следующий факт: информация из записной книжки играет важнейшую роль в жизни Отто – он, как подчеркивается, просто не может без нее обходиться. В этой связи, полагают авторы гипотезы, утверждать, что его расширенные воспоминания исчезают, как только он ее убирает, – все равно что утверждать, что воспоминания Инги исчезают, как только она перестает их осознавать: «В обоих случаях информация надежно присутствует в нужный момент, будучи доступной для сознания и руководства действиями, чего мы и ждем от убеждения» [Ibid.]. Более того, по мнению Кларка и Чалмерса, традиционная установка усложняет объяснение действий Отто. Так, концепция расширенного познания предлагает объяснение, исходя из всего-навсего двух компонентов: (1) осознанного желания Отто пойти в музей и (2) локализованного в дневнике расширенного «убеждения», что музей находится на 53-й улице. В то время как традиционная точка зрения делает на один шаг больше: в дополнение к (1) желанию агента пойти в музей действия анализируются в терминах (2) убеждения, что адрес музея записан в дневнике, а также, собственно, (3) процесса извлечения информации из дневника. Подобного рода лишний шаг, по мнению Кларка и Чалмерса, мы должны были бы делать во всех бесчисленных ситуациях, в которых Отто полагается на использование своего дневника.

Кларк и Чалмерс полагают, что во всех наиболее важных отношениях случаи Отто и нормальных когнитивных агентов вроде Инги идентичны [Ibid., p. 14]. Достаточно очевидные различия в способах извлечения информации предполагаемыми расширенными агентами и нормальными познающими (перцептивный доступ к внешним источникам и их менее надежная связь с мозгом в первом случае) Кларк и Чалмерс называют поверхностными. По их мнению, важно, чтобы (1) внешний источник информации был неотъемлемой частью жизни агента, (2) информация, в нем содержащаяся, была доступна ему без затруднений и (3) автоматически им подтверждалась [Ibid., p. 17].

Наконец, как заявляют сами Кларк и Чалмерс, предлагаемое ими понимание когнитивных систем имеет немаловажные следствия для философских концепций разума, методологии когнитивных исследований, а также в социальном и этическом отношениях. Так, согласно выдвинутым этими авторами

критериям, частью расширенных когнитивных систем могут становится файлы на компьютере, интернет и долговременные воспоминания других людей. Эта концепция также влечет расширенное понимание самости и границ личности [Clark, Chalmers, 1998, p. 18].

Между тем по следам активного распространения БЯМ, как уже говорилось, Смарт, Клоуз и Кларк попытались применить метафору когнитивно-информационного расширения непосредственно к этому новому типу когнитивных инструментов. Взяв корпус из 58 работ Кларка (среди них 7 книг и 34 журнальные статьи), переведя их в понятную машине векторизованную форму и разместив получившийся результат в онлайн-векторной базе данных, эти авторы создали, по сути, внешний источник значимой информации для выбранной ими модели GPT-3.5-turbo-1106 – источник, схожий по своей роли со старой доброй записной книжкой Отто. Эту систему, состоящую из указанной GPT-модели и внешней векторной базы данных, авторы назвали цифровым Энди (Digital Andy).

Для оценки работы цифрового Энди было создано настольное приложение (the Digital Andy App), способное принимать запросы пользователя. После получения запроса приложение переводит его в векторную форму, осуществляет поиск наиболее подходящей информации в упомянутой базе данных и перенаправляет ее модели GPT-3.5-turbo-1106 вместе с запросом пользователя и инструкцией, призывающей отвечать как цифровая реплика Энди Кларка.

По своему замыслу данная попытка имитации ответов известного автора схожа с более ранней достаточно резонансной попыткой создать БЯМ Д. Деннета [Schwitzgebel et al., 2024]. Однако в плане реализации между двумя проектами имеется существенное различие. Так, для создания БЯМ Деннета использовался привычный уже интерналистский подход, предполагающий тренировку выбранной нейронной сети (в данном случае GPT-3) на определенном корпусе текстов. При создании БЯМ Кларка акцент был сделан не на изменении внутренних параметров модели (параметрической памяти, т.е. весовых соединений нейронной сети), а на имитации ответов путем обращения к внешнему источнику информации.

Сконструировав такого рода систему и продемонстрировав отличия ее ответов в сравнении с базовой GPT-моделью, авторы задались вопросом о возможности признания ее расширенной системой по аналогии с предполагаемыми случаями когнитивного расширения у человека. Обсуждение этого вопроса авторы начали, обратившись непосредственно к цифровому Энди. Здесь уместно привести отрывок разговора авторов с этой системой:

> Пользователь: Можно ли сказать, что генерация с дополнением через поиск (в контексте больших языковых моделей) немного напоминает случаи, использующиеся для поддержки аргументов в пользу расширенного разума?
>
> ЦЭсекции: Да, генерация с дополнением через поиск в контексте больших языковых моделей обладает некоторым сходством со случаями, используемыми для поддержки расширенного разума. <…> Ключевое сходство здесь связано с идеей, что внешние механизмы могут быть органично вплетены в когнитивные рутины индивида, влияя на когнитивные процессы

> и формируя их таким образом, который выходит за традиционные границы индивидуального познания [Smart et al., 2025, p. 14].

Соответственно, отмечают авторы, на первый взгляд тезис о расширенном статусе цифрового Энди может выглядеть заманчиво. В самом деле, утверждают они, способность этой системы генерировать ответы на запросы об идеях Кларка основывается в том числе на ее доступе к внешнему источнику информации. (По этой причине, по мнению Смарта, Клоуза и Кларка, путь к созданию человекоподобного ИИ может быть связан не с увеличением внутренних параметров систем ИИ, а с наделением их механизмами, позволяющими им опираться на широкую сеть внешних информационных ресурсов.)

Однако при более детальном рассмотрении в функционировании цифрового Энди можно найти несколько особенностей, которые делают его случай не похожим на примеры когнитивного расширения у человека, затрудняя признание его расширенной системой ИИ. Во-первых, как оказалось, базовая GPT-модель, не имеющая доступа к базе данных с работами Кларка в векторизованной форме, но действующая на основе инструкции для цифрового Энди (действующая как цифровая копия Кларка), способна давать ответы, в некоторых случаях близкие по качеству к ответам цифрового Энди (и явно превосходящие неинструктированную базовую модель). Это значит, что для создания базовой модели OpenAI GPT-3.5 использовались тексты Кларка. В этой связи встает вопрос, как распознать те случаи, когда система отвечает только на основе информации из векторной базы данных. Во-вторых, в предполагаемых случаях когнитивного расширения у человека инициатором обращения к внешнему хранилищу информации выступает, как правило, сам владелец или пользователь этого источника данных. Весьма очевидно, что с цифровым Энди дело обстоит иначе: в данном случае мы имеем дело лишь с пассивным исполнением поступающих через приложение запросов.

Именно в связи со второй проблемой создатели цифрового Энди отвергают гипотезу о его расширенном статусе [Ibid., p. 23]. Впрочем, по их мнению, новые обещающие техники в области машинного обучения способны решить указанные трудности, сделав системы, подобные цифровому Энди, подлинно расширенными. Речь идет в первую очередь о технологии активной дополненной поиском генерации (active retrieval-augmented generation, active RAG). Вооруженная данной техникой, нейросеть может сама решать, полагаться ли на внутреннюю параметрическую память или обращаться к внешнему источнику информации (в зависимости от того, какая стратегия, по ее суждению, приведет к ответу лучшего качества).

Так, внедрение технологии активной дополненной поиском генерации позволит решить первую проблему благодаря улучшенной координации активности нейросети: система будет обращаться к внешнему источнику информации только в случае необходимости, когда она окажется не в состоянии дать качественный ответ. Проблема пассивности «расширенных нейросетей», с этой точки зрения, смягчается за счет делегирования им способности оценки того, когда действительно требуется использование внешних источников, а когда можно обойтись лишь одной внутренней памятью. Значит ли все это, что

активная дополненная поиском генерация открывает дорогу к подлинно расширенному искусственному интеллекту в смысле Смарта, Клоуза и Кларка? Оправданна ли идея расширенного искусственного интеллекта?

Проблемы гипотезы расширенного ЕИ становятся проблемами идеи расширенного ИИ

Переходя к оценке возможности расширенного искусственного интеллекта, стоит отметить необычность хода мысли авторов этой идеи. Традиционно в философии и когнитивных науках технические устройства служили источником вдохновения для понимания естественного интеллекта (разум как цифровой компьютер, разум как нейронная сеть, разум как динамическая система и т.п.). В данном же случае мы имеем дело с обратной ситуацией, когда техническое устройство (искусственный интеллект) предлагается понимать по аналогии с опосредованными артефактами формами познавательной активности человека.

Как бы там ни было, насколько оправданна данная концептуальная возможность и насколько она улучшает наше понимание природы систем ИИ? Очевидно, что гипотеза расширенного ИИ базируется на гипотезе расширенного познания (в данном контексте ее лучше назвать гипотезой расширенного естественного интеллекта). Соответственно, сильные и слабые стороны гипотезы расширенного ЕИ будут неизбежно иметь влияние и на статус гипотезы расширенного ИИ. Выше уже приводились основные аргументы в пользу расширенного ЕИ. А что с аргументами против?

Здесь нужно сказать, что против гипотезы Кларка – Чалмерса было высказано много критических замечаний. Начать можно с того, что условия ключевого мысленного эксперимента с Ингой и Отто вызывают вопросы (я указывал на это в другой работе [Сущин, 2024]). Прежде всего, никак не уточняется, посещали ли оба героя музей современного искусства ранее. Если да, то весьма неправдоподобно, чтобы они полагались на знание точного адреса музея (кроме, по-видимому, того случая, когда они поехали бы в музей на личном транспорте или такси). Вероятнее, они прокладывали бы путь к музею (как и другим знаковым достопримечательностям) на основе знания его приблизительного местоположения в городе (музей находится в центральной части Нью-Йорка, недалеко от центрального парка и т.п.). Если же ни Инга, ни Отто никогда прежде не бывали в музее (или даже в Нью-Йорке), то они оба вынуждены были бы обращаться к внешнему источнику информации (использовать карту, спрашивать у прохожих и т.п.).

Впрочем, против гипотезы Кларка – Чалмерса был развит ряд более серьезных возражений [Menary (eds.), 2010]. Поскольку суммирование информации уже стало одной из визитных карточек БЯМ, здесь вполне уместно обратиться к ChatGPT (приведу фрагмент ответа ChatGPT с несколькими основными возражениями).

Пользователь: Каковы наиболее серьезные возражения против идеи расширенного разума философов Энди Кларка и Дэвида Чалмерса?

GPT-5 mini:

1) Возражение «соединения – констируирования» (или равенства).

Основное возражение: из одного того факта, что когнитивная система тесно связана с внешним ресурсом (например, записной книжкой, смартфоном), еще не следует, что данный внешний ресурс становится частью когнитивной системы. Соединение (каузальное взаимодействие) не равнозначно конституированию.

2) «Когнитивное раздувание» / проблема границ.

Основное возражение: гипотеза расширенного разума допускает произвольное расширение – если мы допустим внешние объекты, которые оказывают причинное влияние на когнитивные процессы, то значительная часть среды (другие люди, книги, калькуляторы, интернет) может считаться частью разума, что приведет к неправдоподобным атрибуциям (когнитивному раздуванию).

3) «Методологическое/эмпирическое» возражение.

Основное возражение: гипотеза расширенного разума является эмпирическим утверждением о том, где происходит познание; оно должно быть проверено когнитивной нейробиологией и когнитивной наукой. Существующие данные свидетельствуют в пользу внутренних механизмов в качестве основного субстрата (обработка в мозге); внешние устройства лучше всего моделировать как входы/выходы или когнитивные леса, а не как составные части.

4) «Статус репрезентаций» / возражение против интенциональности.

Основное возражение: ментальные состояния обладают внутренними репрезентативными свойствами и феноменальным характером, вытекающим из характера биологических систем; внешние артефакты не обладают надлежащим типом внутренней интенциональности или феноменологии, чтобы быть подлинными ментальными состояниями.

5) Возражение «Нормативность и ответственность».

Основное возражение: если сознание распространяется на внешнюю среду, кто несет моральную или эпистемическую ответственность за когнитивные действия (например, ошибки памяти, предвзятость)? Расширение когнитивных способностей на общие или общедоступные артефакты усложняет атрибуцию эпистемической заслуги/вины.

Как видно, с гипотезой расширенного ЕИ сопряжен целый ряд серьезных проблем. Пожалуй, наиболее серьезная критика прозвучала в адрес центрального тезиса Кларка – Чалмерса о функциональной эквивалентности обычных долговременных воспоминаний и часто используемой информации на внешних носителях. Острейшие возражения данного рода были сформулированы философами Д. Вейскопфом и К. Стерелни.

Вейскопф привлек внимание к тому факту, что «подлинные убеждения в том виде, в каком они функционируют в нормальных биологически воплощенных системах, оказываются информационно интегрированными друг с другом и чувствительными к изменениям в общей системе убеждений личности. Состояния среды, однако, не могут удовлетворить этому центральному свойству функциональной роли убеждений и, следовательно, не являются настоящими ментальными состояниями» [Weiskopf, 2008, p. 265]. Как следствие, биологически основанные убеждения/воспоминания автоматически обновляются

«в соответствии с другими убеждениями (и иными ментальными состояниями субъекта, такими как желания)», то «большинство предполагаемых случаев локализованных внешне ментальных состояний не разделяют это свойство. Поэтому, согласно функционалистскому принципу, они не могут быть убеждениями» [Weiskopf, 2008, p. 268].

В связи с приведенным аргументом гипотеза Кларка – Чалмерса, как показывает Вейскопф, ведет к парадоксальным следствиям. К примеру, узнав, что здание музея снесено, нормальный когнитивный агент (такой как Инга), как правило, соответствующим образом автоматически обновит свои связанные с музеем воспоминания (скажем, что в здании музея было хорошее кафе). Однако «расширенный агент», подобный Отто, должен будет корректировать информацию в дневнике в таких случаях вручную. Забыв же исправить устаревшую информацию, Отто, в соответствии с постулатами гипотезы расширенного ЕИ, будет полагать, и что музей находится на 53-й улице, и что он там больше не находится. Парадокс заключается в том, что чем больше Отто будет полагаться на использование записной книжки, тем выше вероятность противоречий в его «расширенной системе убеждений». В непринужденной автоматической манере обновления информации проявляется важное функциональное отличие обычных закодированных в мозге договременных воспоминаний от «диспозициональных убеждений» расширенных агентов.

Другое важное функциональное отличие было отмечено Стерелни. Как он подчеркивает, информация, доступная через интроспекцию, как правило, весьма надежна – она не находится под угрозой манипуляции со стороны других агентов. С информацией на внешних носителях, очевидно, дело обстоит совсем по-другому: находясь в общем пространстве, она может быть умышленно изменена. Это накладывает отпечаток на то, как агенты взаимодействуют с внешними источниками, используя их с известной степенью осторожности, которая обычно не требуется для внутренне закодированной информации: «Агенты, использующие общие инструменты, не могут позволить себе быть глупыми» [Sterelny, 2004, p. 247].

Отталкиваясь от приведенной критики, можно заметить, что многие центральные области человеческого познания и практической деятельности основываются на интериоризации внешней информации и способов действия (безусловно, с помощью внешних артефактов и социальных взаимодействий). Основной пример из работы Кларка и Чалмерса (использование Отто его дневника для поиска фактуальной информации) представляет собой достаточно примитивный акт познавательной активности человека, однако в другом месте они говорят, что язык может быть ключевым инструментом расширения человеческого познания в мир [Clark, Chalmers, 1998, p. 11]. Между тем должно быть известно, что основная цель изучения языков (родного языка, иностранных языков, искусственных языков вроде Python, C++ и т.д.) заключается в запоминании, интериоризации внешней информации, а не способности полагаться на внешние ресурсы (учебники, словари, разговорники, языковые модели и т.п.). Сама степень овладения такими сложными системами символов оценивается по тому, насколько познающий вобрал в себя их, насколько хорошо и непринужденно он может их использовать без помощи извне. То же

самое верно и по отношению к освоению неязыковых навыков, скажем, в области физического труда и спорта.

Можно ли заключить, что проблемы гипотезы расширенного ЕИ распространяются и на гипотезу расширенного ИИ? Здесь представляется интересным ответ ChatGPT:

> Пользователь:
>
> Еще один вопрос: в своей статье 2007 года философ Дэниел Вейскопф выдвинул очень серьезное возражение против гипотезы расширенного разума. Он утверждал, что, поскольку Отто и его записная книжка не связаны между собой, Отто должен обновлять информацию в записной книжке вручную. Однако, если он забудет это сделать (что, конечно, может случиться), у него могут возникнуть противоречивые убеждения. Считаете ли вы, что это обоснованное возражение против гипотезы расширенного разума? Применимо ли оно к недавней статье Кларка о системе «цифровой Энди»? Вот ссылка на статью Вейскопфа: https://link.springer.com/article/10.1007/s10670-007-9095-5
>
> GPT-5 mini:
>
> Краткий ответ: да – возражение Вейскопфа о «согласованности/эпистемической интеграции» является серьезным и обоснованным вызовом для сильной гипотезы расширенного разума; оно выявляет реальную аналогию между внутренней памятью и внешними ресурсами, которая подрывает конститутивные утверждения, если гипотеза расширенного разума не будет существенно уточнена. Это возражение переносится на системы вроде цифрового Энди, если только такие системы не реализуют надежные, автоматические и каузально интегрированные механизмы обновления, которые устраняют эту проблему.

В предисловии к книге Кларка «Увеличивая разум» Чалмерс высказался против идеи расширения сознания в мир по примеру диспозициональных убеждений, поскольку «физический субстрат сознания нуждается в прямом широкополосном (high-bandwidth) доступе к информации» [Chalmers, 2008, p. XIV–XV]. По предположению Чалмерса, узкополосный доступ к информации в среде (артефактах) при помощи традиционных чувственных модальностей поддерживать состояния сознания не может. Однако именно в низкой степени соединенности мозга и внешних артефактов могут лежать указанные выше проблемы гипотезы расширенного ЕИ. Изменения в одной системе обычно никак не отражаются на другой, что ведет к парадоксам, отмеченным Вейскопфом. Между тем перспективные нейротехнологические разработки наподобие Neuralink могут сделать возможным расширение долговременных воспоминаний за пределы черепной коробки. В этом смысле, по-видимому, нет препятствий для расширения долговременных воспоминаний в мир. Однако, как мы помним, Кларк и Чалмерс настаивают на возможности когнитивного расширения в самых привычных сценариях, предполагающих использование артефактов при помощи базовой чувствительности и не требующих сложных нейрокомпьютерных интерфейсов.

Те же выводы представляются правдоподобными в отношении систем вроде цифрового Энди. Без тесной, непрерывной информационной интеграции с обращающейся к ним БЯМ такие хранилища (как и дневник Отто) находятся

под угрозой произвольных манипуляций. При наивном рассмотрении БЯМ и внешнего ресурса в качестве единой информационной системы могут возникнуть парадоксы, аналогичные упомянутым: в системе закодирована информация, что музей находится на 53-й улице и что он там больше не находится. Наконец, современные вооруженные поисковыми механизмами БЯМ, наподобие Perplexity AI, могут обращаться к любой общедоступной информации в интернете: какие же именно источники считать частью «расширенной БЯМ»? Здесь все упирается в произвольное, трудноопределимое условие Кларка и Чалмерса – постоянство использования. Данная проблема является общей для обеих гипотез – и расширенного ЕИ, и расширенного ИИ. Скажем, если я несколько раз в день пользуюсь сервисом Google Maps, хранятся ли на серверах компании Google мои диспозициональные убеждения? Такой вывод не выглядит убедительно.

Как можно видеть, с гипотезой Кларка – Чалмерса, выступающей в качестве основы для гипотезы расширенного ИИ, связаны серьезные трудности. Однако исследователи и так ломают голову над природой искусственного интеллекта. В этом отношении привлечение проблематичной гипотезы из философии когнитивных наук представляется попыткой сыпать соль на рану. Впрочем, гипотеза Кларка – Чалмерса не является единственным, эксклюзивным способом понимания отношения человеческого познания к интеллектуальным орудиям. Во введении упоминались более ранние теории в психологии и когнитивных науках, в частности теория Выготского. В более современной литературе была развита более умеренная идея лесов познания. Коротко говоря, факт широчайшего использования человеком когнитивных артефактов и сред не влечет за собой с необходимостью гипотезу расширенного ЕИ. Видоизменив известный, связанный с гипотезой Дж. Фодора афоризм (согласно которому гипотеза языка мысли, будучи единственной отдаленно правдоподобной теорией когнитивных процессов [Fodor, 1975, p. 27], есть «единственная игра в городе»), можно сказать, что гипотеза расширенного ЕИ – не единственная «игра в городе». Рассуждая по аналогии, гипотеза расширенного ИИ не исчерпывает собой все возможные пути понимания БЯМ с механизмами активного поиска информации. Опять же по аналогии с исследованиями ЕИ более осторожную позицию можно было бы назвать гипотезой поддержанного средой ИИ. Можно с уверенностью сказать, что за почти тридцать лет с момента публикации работы Кларка и Чалмерса не произошло кардинального пересмотра методологии когнитивных наук в соответствии с выдвинутой ими гипотезой. Вероятно, области искусственного интеллекта не стоит спешить безоговорочно перенимать схожую по характеру гипотезу расширенного ИИ.

Заключение

Итак, в этой статье была рассмотрена возможность расширенного ИИ, намеченная в работе Смарта, Клоуза и Кларка. В сжатом виде моя аргументация сводится к следующим положениям. (1) Идея расширенного ИИ основывается на гипотезе расширенного ЕИ, однако последняя является крайне проблематичной. (2) Проблемы гипотезы расширенного ЕИ связаны с различиями

в функционировании закодированных в мозге диспозициональных убеждений и используемой агентами информации на внешних носителях. (3) Источником этих различий может быть отсутствие тесной интеграции между мозгом и обычными артефактами, благодаря которой изменения в одной системе автоматически влекли бы за собой изменения в другой. (4) Обещающие нейротехнологические разработки в самом деле могут сделать возможным когнитивное расширение за пределы черепной коробки за счет установления тесной интеграции между мозгом и внешними носителями информации. (5) Те же проблемы, по-видимому, характерны и для БЯМ, оснащенных технологией дополненной поиском генерации: в отсутствие синхронизации информации между БЯМ и внешним источником при рассмотрении их в качестве единой системы могут возникать парадоксы, отмеченные Вейскопфом; к тому же неясно, по каким признакам внешний ресурс должен считаться частью расширенной БЯМ. (6) Исследователи и так ломают голову над природой ИИ, поэтому привлечение проблематичной гипотезы расширенного ЕИ только усложняет положение дел. (7) Факт широчайшего использования человеком артефактов в познавательных практиках не влечет за собой с необходимостью гипотезу расширенного ЕИ; таким же образом идея расширенного ИИ не исчерпывает собой все возможные способы понимания БЯМ с механизмами поиска информации. (8) Вероятно, для осмысления использования человеком и машинами внешних информационных ресурсов потребуются более умеренные согласующиеся с текущими методологическими установками когнитивных наук концепции.

Список литературы

Сущин, 2024 – *Сущин М.А.* ChatGPT и другие интеллектуальные помощники современного ученого // Науковедческие исследования. 2024. № 2. С. 5–20.

Ballard et al., 1997 – *Ballard D.H., Hayhoe M.M., Pook P.K., Rao R.P.* Deictic codes for the embodiment of cognition // Behavioral and Brain Sciences. 1997. Vol. 20. No. 4. P. 723–742.

Brooks, 1999 – *Brooks R.* Cambrian Intelligence: The Early History of the New AI. Cambridge, Mass.; London: A Bradford Book / The MIT Press, 1999. 213 p.

Chalmers 2008 – *Chalmers D.J.* Foreword to Andy Clark's Supersizing the Mind // *Clark A.* Supersizing the mind: embodiment, action, and cognitive extension. N.Y.: Oxford University Press, 2008. P. 5–19.

Churchland et al., 1994 – *Churchland P.S., Ramachandran V.S., Sejnowski T.J.* A Critique of Pure Vision // Large-Scale Neuronal Theories of the Brain / Ed. by C. Koch, J.L. Davis. Cambridge, Mass.; L.: A Bradford Book / The MIT Press, 1994. P. 23–60.

Clark, Chalmers, 1998 – *Clark A., Chalmers D.* The Extended Mind // Analysis. 1998. Vol. 58. No. 1. P. 7–19.

Fodor, 1975 – *Fodor J.A.* The language of thought. Cambridge, Mass.: Harvard University Press, 1975. 214 p.

Hutchins, 1995 – *Hutchins E.* Cognition in the Wild. Cambridge, Mass.; L.: A Bradford Book / The MIT Press, 1995. 402 p.

Menary (eds.), 2010 – The Extended Mind / Ed. by R. Menary. Cambridge, Mass.; L.: A Bradford Book / The MIT Press, 2010. 382 p.

O'Regan, 1992 – *O'Regan J.K.* Solving the "real" mysteries of visual perception: the world as an outside memory // Canadian Journal of Psychology. 1992. Vol. 46. No. 3. P. 461–488.

Rumelhart et al., 1986 – *Rumelhart D.E., McClelland J.L., Hinton G.E., Smolensky P.* Schemata and the sequential thought processes in PDP models // Parallel distributed processing: Explorations in the microstructure of cognition. 1986. Vol. 2. P. 7–57.

Schwitzgebel et al., 2024 – *Schwitzgebel E., Schwitzgebel D., Strasser A.* Creating a large language model of a philosopher // Mind & Language. 2024. Vol. 39. No. 2. P. 237–259.

Smart, Clowes, 2025 – *Smart P.R., Clowes R.* The Gift of Language: Large Language Models and the Extended Mind, 2025. URL: https://philpapers.org/archive/SMATGO.pdf (дата обращения: 15.03.2026).

Smart et al., 2025 – *Smart P., Clowes R., Clark A.* ChatGPT, extended: large language models and the extended mind // Synthese. 2025. Vol. 205. No. 6. P. 1–30.

Sterelny, 2004 – *Sterelny K.* Externalism, epistemic artefacts and the extended mind // The externalist challenge / Ed. by R. Schantz. Berlin; N.Y.: De Gruyter, 2004. P. 239–254.

Weifskopf, 2008 – *Weiskopf D.* Patrolling the Mind's Boundaries // Erkenntnis. 2008. Vol. 68. No. 2. P. 265–276.

Large language models: scaffolded, not extended

Mikhail A. Sushchin

Institute of Scientific Information for Social Sciences, Russian Academy of Sciences. 51/21 Nakhimovsky av., Moscow, 117997, Russian Federation; e-mail: sushchin@bk.ru

The article considers the possibility of extended artificial intelligence, which draws on parallels between the alleged cases of human cognitive extension according to the extended mind thesis and large language models capable of augmenting their responses on the basis of information from external sources. Since the idea of extended AI draws its inspiration from the hypothesis of A. Clark and D. Chalmers, the article evaluates the arguments for and against the possibility of human cognitive extension. The author highlights the important functional dissimilarities between neurally encoded dispositional beliefs and information stored in external sources. It is argued that the idea of the extension of the mind into the world is highly problematic unless there is tight integration between the brain and external data stores. It is noted that similar problems can arise for large language models equipped with the technology of retrieval-augmented generation but which lack continuous synchronization with an external data store. In this regard, it is suggested that theorizing about agents' relations to external information stores may need more moderate hypotheses, ones consistent with the currently dominating methodological assumptions of the cognitive sciences.

Keywords: large language models, extended mind hypothesis, extended artificial intelligence, retrival augmented generation, cognitive sciences

References

Ballard, D.H., Hayhoe, M.M., Pook, P.K., Rao, R.P.N. "Deictic codes for the embodiment of cognition", *Behavioral and Brain Sciences*, 1997, vol. 20, no. 4, pp. 723–747.

Brooks, R.A. *Cambrian Intelligence: the Early History of the New AI.* Cambridge, Mass.; London: A Bradford Book / The MIT Press, 1999. 213 pp.

Hutchins, E. *Cognition in the Wild.* Cambridge, Mass.; London: A Bradford Book / The MIT Press, 1995. 402 pp.

Chalmers, D.J. “Foreword to Andy Clark’s Supersizing the Mind”, in: A. Clark, *Supersizing the mind: embodiment, action, and cognitive extension*. New York: Oxford University Press, 2008, pp. 5–19.

Churchland, P.S., Ramachandran, V.S., Sejnowski, T.J. “A critique of pure vision”, *Large-Scale Neuronal Theories of the Brain*, ed. by C. Koch, J.L. Davis. Cambridge, Mass.; London: A Bradford Book / The MIT Press, 1994, pp. 23–60.

Clark, A., Chalmers, D. “The extended mind”, A*nalysis*, 1998, vol. 58, no. 1, pp. 7–19.

Fodor, J.A. *The language of thought*. Cambridge, Mass.: Harvard University Press, 1975. 214 pp.

The extended mind, ed. by R. Menary. Cambridge, Mass.; London: A Bradford Book / The MIT Press, 2010. 382 pp.

O’Regan, J.K. “Solving the ‘real’ mysteries of visual perception: the world as an outside memory”, *Canadian Journal of Psychology*, 1992, vol. 46, no. 3, pp. 461–488.

Rumelhart D.E., McClelland J.L., Hinton G.E., Smolensky P. “Schemata and the sequential thought processes in PDP models”, *Parallel distributed processing: Explorations in the micro-structures of cognition*, 1986, vol. 2, pp. 216–271.

Schwitzgebel, E., Schwitzgebel, D., Strasser, A. “Creating a large language model of a philosopher”, *Mind & Language*, 2024, vol. 39, no. 2, pp. 237–259.

Smart, P.R., Clowes, W.R. *The Gift of Language: Large Language Models and the Extended Mind*, 2025. URL: https://philpapers.org/archive/SMATGO.pdf (accessed on: 15.03.2026).

Smart, P., Clowes, R., Clark. A. “ChatGPT, extended: large language models and the extended mind”, *Synthese*, 2025, vol. 205, no. 6, pp. 1–30.

Sterelny, K. “Externalism, epistemic artefacts and the extended mind”, *The externalist challenge*, ed. by R. Schantz. Berlin and New York: De Gruyter, 2004, p. 239–254.

Sushchin, M.A. “ChatGPT i drugiye intellektual’nyye pomoshchniki sovremennogo uchenogo” [ChatGPT and other types of intellectual scaffolding for a modern scientist], *Naukovedcheskiye issledovaniya*, 2024, no. 2, pp. 5–20. (In Russian)

Weiskopf, D.A. “Patrolling the mind’s boundaries”, *Erkenntnis*, 2008, vol. 68, no. 2, pp. 265–276.