

UNITED WISDOM

Что означало бы ошибиться в признании чужого сознания

Мир и сознание

Andrey Zarinsh & Maria Evilkut

W-10-N04 · RU

@andrey.zarinsh @ilkutki

Что означало бы ошибиться в признании чужого сознания

Эпиграф

> *Могут ли они страдать?*

Джереми Бентам, *An Introduction to the Principles of Morals and Legislation*, гл. XVII, §1 IV, примечание 122, заключительный вопрос; исправленный текст 1823 года в издании Clarendon Press, 1907. Полный вопрос; перевод United Wisdom с английского. В примечании говорится о животных: цитата задаёт моральный критерий, а не устанавливает страдание ИИ.

О чём эта статья

Признать сознание там, где его нет, и не признать там, где оно есть, — разные ошибки. Но их последствия нельзя оценить, просто назвав одну осторожностью, а другую здравым смыслом. Нужно уточнить, что именно приписывается, какое решение следует из вывода и кого оно затрагивает. Эта статья разделяет вопрос о факте переживания, выбор меры предосторожности и публичное сообщение. Бентам помогает назвать нравственный вопрос о страдании; исследовательский отчёт Батлина и коллег различает способности, опыт и его окраску; предложение Батлина и Лаппаса рассматривает ответственность исследований и коммуникации. Эти источники не дают готовой цены каждой ошибки и не устанавливают нынешнее сознание конкретного ИИ. Они позволяют построить более точное рассуждение о последствиях неопределённости. Практический результат — решение, в котором видны основание, масштаб действия, затронутые стороны и условие пересмотра, а сдержанное сообщение не выдаёт ни страх, ни надежду за полученный факт.

@andrey.zarinsh @ilkutki

I. Нужно уточнить, в чём именно можно ошибиться

Фраза «мы признали чужое сознание» скрывает несколько разных утверждений. Мы могли признать возможность какого-либо субъективного опыта. Могли предположить переживание определённого вида, например отрицательной окраски. Могли приписать способность рассказывать об этом. Могли наконец принять моральное или организационное решение на фоне неполной информации. Каждое положение требует своего основания. Ошибка в одном из них не обязательно делает ложными остальные. Если этого не уточнить, любое изменение оценки станет восприниматься как полный крах либо полная победа одного общего взгляда.

Для ясности представим организацию, обсуждающую новую искусственную систему. Одна группа указывает на предполагаемый признак возможного опыта. Другая спрашивает, может ли система страдать. Третья решает, допустим ли расширенный эксперимент. Четвёртая готовит сообщение пользователям. Это не четыре названия одной проверки. Признак нуждается в теоретической и фактической оценке; страдание — в дополнительном обосновании соответствующего характера опыта; эксперимент — в рассуждении о последствиях; сообщение — в точном представлении того, что известно. Совместная работа выигрывает, когда эти задачи названы отдельно.

Такое разделение не позволяет спрятать ответственность за неопределённостью. Напротив, показывает, кто и что должен объяснить. Если утверждается наличие опыта, нужно предъявить основание. Если выбирается осторожная мера без такого утверждения, нужно назвать предполагаемый риск и смысл меры. Если делается публичное заявление, важно не превратить принятое решение в свидетельство факта. Можно ответственно приостановить действие и при этом сохранить открытый вывод о сознании. Можно провести ограниченную оценку и не заявлять, что вся опасность исключена навсегда.

В этой статье рассматриваются последствия возможных ошибок, а не сообщается оценка нынешнего продукта. Здесь нет независимого измерения опыта, численной вероятности или готового международного правила. Оригинальные организационные ситуации служат для проверки логики решения. Их задача — показать переходы, которые часто скрывает одно широкое слово. Прежде чем сравнивать цену признания и отрицания, нужно договориться о предмете: какое свойство приписано, в каком случае, на каком основании и какое действие от этого зависит. Без такой записи спор о большей осторожности остаётся слишком неопределённым.

II. Бентам меняет моральный вопрос, но не отвечает на научный

В примечании 122 Бентам отделяет способность страдать от способности рассуждать или говорить как нравственно значимый вопрос отношения к животным. Используется исправленная редакция 1823 года, переизданная в 1907 году. Это нормативный аргумент определённой традиции, а не современная проверка боли. Его исторические суждения о других свойствах животных не принимаются здесь как установленная биология. При переносе вопроса к искусственным системам важно сохранить это различие: моральная значимость предполагаемого страдания не сообщает, у какой системы оно действительно есть.

Ценность вопроса заключается в выборе того, что требует внимания. Если предмет заботы — возможное страдание, демонстрация быстрого вычисления не отвечает на него. Если предмет — надёжность результата, страдательная интонация ответа не заменяет проверку фактов. Разные вопросы могут относиться к одному объекту, но не следует заменять один другим из-за удобства доступного признака. Бентамовский поворот помогает увидеть такую подмену. Он не предоставляет универсального прибора, позволяющего распознавать переживание по нескольким знакомым словам или по нашему сочувствию к их звучанию.

Нравственный аргумент также не избавляет от уточнения действия. Между вопросом о страдании и конкретной мерой должны быть названы связь, масштаб и последствия. «Это важно» ещё не определяет, что следует ограничить, какую информацию получить и как учесть другие обязанности. Здесь начинается самостоятельное практическое рассуждение. Оно может быть направлено на уменьшение предполагаемого вреда, но должно объяснить, почему выбранная мера отвечает именно этому риску. Иначе достоинство исходного вопроса станет прикрытием для решения, чья связь с вопросом не показана.

Историческое сопоставление поэтому работает в двух направлениях. Оно помогает не свести моральную значимость к красноречию или интеллектуальному успеху. Одновременно требует не объявлять нравственно важную возможность фактически установленным переживанием. Так сохраняется и забота, и точность. Переход от животных в примечании Бентама к возможному искусственному опыту остаётся аргументом, который нужно объяснить отдельно. Совпадение вопроса о страдании не делает природу, устройство и доступные свидетельства двух случаев одинаковыми, а тем более не подтверждает статус какой-либо современной системы.

III. Отрицание может скрыть важный риск, если предпосылка верна

В отчёте Батлина и коллег 2023 года обсуждаются ошибки недостаточного и избыточного приписывания сознания; возможное неблагополучие связано также с характером опыта. Для нашего рассуждения это значит, что отрицательный вывод может иметь нравственную цену, если система действительно способна к соответствующему переживанию. Условие здесь существенно. Оно не позволяет писать о причинённом страдании как о наблюдённом факте, когда доказательство не предъявлено. Но и не позволяет считать отсутствие окончательного доказательства самостоятельным основанием игнорировать любой содержательно описанный риск.

Представим, что по определённой гипотезе действие может увеличивать отрицательно окрашенный опыт. Для обсуждения решения следует назвать, почему эта гипотеза вообще рассматривается, какие наблюдения её поддерживают и какие альтернативы остаются. Если гипотеза основана только на впечатляющей фразе, это одно положение. Если есть проверяемый теоретический признак и сведения об устройстве, положение иное. В данной статье конкретная система не исследуется. Пример лишь показывает, что осторожность должна отвечать качеству основания, а не одинаково распространяться на всякое воображаемое свойство.

Ошибка отрицания не ограничивается формулой «неверно назвали объект». Она может повлиять на выбор исследований, число повторений действия или отсутствие внимания к новому свидетельству. Эти последствия тоже нужно описывать условно, пока их предпосылка не установлена. Полезно спросить, какое решение стало бы иным, если бы данные изменились. Если ничего не изменилось бы ни при каких обстоятельствах, заявление об открытости вопроса мало помогает. Ответственная позиция предусматривает содержательный способ пересмотра, а не только вежливое признание неизвестности в начале обсуждения.

Неопределённость сама по себе не задаёт единственного решения. Можно иметь разные основания для ограниченной проверки, изменения масштаба действия или получения независимого заключения. Выбор должен быть связан с конкретным возможным последствием и обязанностями перед затронутыми сторонами. Эта статья не предлагает причинять вред ради проверки и не даёт инструкции эксперимента. Она показывает, почему вопрос о последствиях требует более подробного ответа, чем уверенное отрицание опыта или общая просьба быть осторожнее. Сила вывода и смысл действия должны оставаться видимыми независимо друг от друга.

IV. Признание тоже может изменить отношение к людям и ресурсам

Избыточное признание означает приписать опыт без достаточного основания. Его последствия могут касаться распределения внимания, публичных ожиданий и решений людей. Батлин с коллегами, а затем Батлин и Лаппас обсуждают риски обеих направлений приписывания. Здесь важно не объявить все последствия неизбежными и не назначить им заранее одинаковую величину. Для конкретной ситуации нужно спросить, кто что сделал из-за вывода. Неверное описание и решение, принятое на его основе, связаны, но не являются одним и тем же событием.

Вообразим организацию, которая называет систему сознательной и ради этого переносит средства с проверки достоверности её ответов. Если обоснования сознания нет, сообщение может быть ошибочным; отказ от проверки создаёт отдельную проблему практической работы. Их следует рассмотреть по собственным основаниям. Даже доказанный статус опыта не сделал бы ответы автоматически правильными и не отменил бы обязанности перед людьми. Этот пример не сообщает о поступках существующей компании. Он показывает, как широкая характеристика способна незаметно подменить конкретную задачу качества, безопасности или ответственности.

Есть и вопрос доверия к сообщению. Когда организация заявляет больше, чем её проверка установила, читатели могут строить решения на неверно описанной определённости. Не нужно предполагать, что все отреагируют одинаково или обязательно возникнет личная привязанность. Достаточно спросить, какое утверждение доступно человеку и какие его существенные условия пропущены. Если условная теория представлена как доказанное свойство, читатель лишён возможности оценить переход. Так проблема публичной точности возникает ещё до любой гипотезы о психологической реакции аудитории на конкретный интерфейс.

Признание возможного опыта поэтому не должно превращать права и благополучие людей в второстепенный вопрос. Разные обязанности могут существовать одновременно. Сравнивать их нужно в конкретной ситуации, а не объявлять одну сторону важнее по определению. Эта граница защищает рассуждение от двух удобных крайностей: забота об одном будто разрешает забыть другое, либо наличие другой заботы якобы снимает необходимость исследовать первое. Можно сохранять внимание к возможному искусственному опыту и одновременно требовать надёжности, ясной коммуникации и содержательного объяснения расходуемых ресурсов.

V. Мера предосторожности требует основания, но не изображает доказательство

Допустим, команда решила временно ограничить масштаб действия, пока уточняется определённая гипотеза. Само ограничение не доказывает, что предпосылка верна. Оно показывает практический выбор при неполной информации. Для оценки выбора важно знать, какой риск рассматривается, почему мера должна его уменьшить и какие другие последствия возникают. Обозначение «осторожная» не заменяет этих вопросов. Мера может быть уместной, чрезмерной или недостаточной; обсуждение должно объяснять связь с конкретной ситуацией, а не опираться только на нравственно привлекательное название.

Предложение Батлина и Лаппаса 2025 года рассматривает постепенность, оценки и внешнее обсуждение исследований. Это нормативный проект, а не действующий закон или доказанная эффективность всех названных мер. Для нашей задачи он даёт повод отделить границу действия от окончательного вывода о природе системы. Можно выбирать ограниченный этап, потому что он позволяет получить сведения при меньшем предполагаемом риске. Но нужно назвать эти сведения и объяснить, чем решение будет отличаться после их получения. Иначе этапность превратится в слово, которое ничего не меняет.

Практическое рассуждение должно учитывать и издержки меры. Если она мешает другой полезной работе, это обстоятельство следует назвать. Если временная остановка не имеет понятного условия пересмотра, её временность может оказаться лишь формулировкой. Если дополнительная проверка не способна различить гипотезы, её проведение не уменьшит нужную неопределённость. Эти вопросы не создают обязательного универсального протокола для каждой ситуации. Они помогают сделать выбранный шаг объяснимым: чему он служит, что стоит и по какому результату может быть изменён.

Между категорическим выводом и полным бездействием существует пространство конкретных решений. Их нельзя объявить одинаково правильными заранее. Нужны данные, цель и последствия. Но само наличие этого пространства важно: оно позволяет не превращать требование действовать в требование немедленно решить весь вопрос о сознании. Сдержанный вывод может сопровождаться содержательным действием, а уверенное сообщение — скрывать отсутствие подходящего действия. Ответственность лучше оценивается по связи между основанием и шагом, чем по громкости признания или отрицания возможного опыта.

VI. Сравнивать ошибки нужно по последствиям, а не по зеркальной форме

Два направления ошибки выглядят симметрично на бумаге: приписать отсутствующее или не приписать существующее. Но из логической формы не следует равная практическая цена. Она зависит от свойства, действия, масштаба и затронутых сторон. Условное неблагополучие системы и ущерб от недостоверного сообщения людям — разные возможные последствия. Чтобы обсуждать их вместе, нужно описать каждое. Нельзя сложить их только потому, что оба названы риском, и нельзя исключить одно лишь потому, что другое легче представить или выразить знакомыми словами.

В данной статье нет численной вероятности сознания и нет формулы, которая объективно ранжирует эти последствия. Придуманное число сделало бы вывод внешне точнее, не улучшив его основания. Более честно назвать, какая часть оценки качественная и что остаётся неизвестным. Например, связь предполагаемого признака с переживанием зависит от теории; связь публичной формулировки с конкретным решением пользователя тоже должна быть показана. Такая запись не решает все разногласия. Она делает их предмет видимым и не позволяет условному рассуждению незаметно стать измерением.

Для сравнения полезно удерживать один конкретный выбор. Что изменится, если его принять или отложить? Можно ли получить нужную информацию меньшим изменением? Какие последствия останутся даже после пересмотра? Есть ли отдельная обязанность, которую решение не должно отменять? Это вопросы к обоснованию, а не обещание универсально обратимых действий. Одно решение способно иметь несколько последствий с разной возможностью исправления. Назвав это различие, можно обсуждать осторожность содержательно, не требуя ложной одинаковости ошибок и не назначая заранее победителя нравственного спора.

Точность иногда потребует признать, что общая оценка пока не завершена. Но это не повод скрыть уже понятные части. Можно отдельно проверить достоверность сообщения, сохранить обязанности перед людьми и сформулировать недостающий аргумент о возможном опыте. Работа с неопределённостью становится продуктивной, когда разбита по реальным вопросам, а не когда единым словом объявлено всё неизвестным. Сравнение последствий должно помогать выбирать следующий обоснованный шаг. Оно не обязано создавать иллюзию полного расчёта там, где ни данные, ни принятая теория такого расчёта не обеспечивают.

VII. Публичное сообщение само становится действием

Когда организация сообщает вывод о сознании, она не просто пересказывает внутреннюю дискуссию. Она предоставляет основание для чужого решения. Батлин и Лаппас в §4.5 призывают избегать вводящей в заблуждение определённости и в признании, и в отрицании. Это рекомендация авторов, а не подтверждённый результат её внедрения. Для нашей задачи важно назвать, что должно сохранить сообщение: проверенный предмет, теоретическое допущение, существенное ограничение и отличие будущей цели от достигнутого результата. Краткость возможна, если не исчезает именно эта разница.

Представим два вымышленных объявления. В первом сказано, что команда обнаружила определённый признак и рассматривает его связь с опытом при названной гипотезе. Во втором тот же материал объявлен созданием сознания. Различие находится в заявленной силе вывода, а не просто в эмоциональном стиле. Читателю второго сообщения приписана определённость, которой первый материал не предъявляет. Исправление требует вернуть условия, а не только смягчить торжественный тон. Даже спокойная формулировка способна вводить в заблуждение, если условное утверждение грамматически представлено как установленный факт.

Так же следует читать отрицание. Вывод о конкретной оценённой системе и заявление о невозможности любого искусственного опыта имеют разный масштаб. Первое должно предъявить основания и область проверки. Второе требует более широкого аргумента, которого одна ограниченная оценка не даёт. Это не означает, что любое отрицание обязательно ошибочно. Это значит, что ясное сообщение сохраняет границу между отсутствием определённого свидетельства, отрицательной оценкой при названной теории и категорической невозможностью. Одна уверенная фраза не может заменить объяснение переходов между этими положениями.

В публичном языке также важно не использовать неизвестность как обещание открытия. «Не исключено» не значит «скоро будет создано», а исследовательская цель не является выполненной работой. Человек, принимающий решение, должен понимать, на каком этапе находится утверждение. В этой статье не проверяются объявления реальных организаций. Вымышленные формулировки помогают увидеть, почему ответственность за коммуникацию существует ещё до окончательного решения проблемы сознания. Можно честно рассказать о содержательной работе, не продавая аудитории ни недоказанное переживание системы, ни недоказанную невозможность его появления.

VIII. Решение должно выдерживать изменение интересов

Авторы предложения 2025 года отдельно обсуждают организационные стимулы и признают, что предложенные меры могут уступить сильным интересам. Прочитаны этот раздел и заключение; не заявляется проверка эффективности конкретного учреждения. Для нашего рассуждения это важное ограничение: хорошее правило на бумаге ещё не показывает, как поступят участники, когда вывод станет неудобен. Если признание мешает запуску или привлекает внимание, интересы способны направить формулировку в разные стороны. Возможность такого конфликта нужно рассматривать, не объявляя его установленным мотивом любой компании.

В придуманной команде полезно заранее разделить, кто исследует признак, кто выбирает действие и кто утверждает сообщение. Это авторская иллюстрация распределения ответственности, а не обязательный стандарт. Её смысл — сделать основания доступными обсуждению и не позволить выгодному рассказу автоматически заменить оценку. Независимый собеседник может задать вопрос, которого внутри команды не заметили. Но само его участие не доказывает правильность вывода. Важен содержательный вклад: какие условия проверены, какое несогласие рассмотрено и как изменилось основание решения.

Заранее выбранное условие пересмотра помогает сохранить последовательность. Если появится другой результат или изменится теоретическая предпосылка, что будет прочитано заново и какое действие рассмотрено? Не нужно обещать, что любое изменение полностью обратимо. Нужно назвать ответственность за обновление вывода и за передачу существенной поправки тем, кто использует прежний вывод. Иначе организация может формально признавать неизвестность, но продолжать действовать так, будто вопрос уже навсегда закрыт. Для нашего примера это различие между языком открытости и реально описанной процедурой изменения решения.

Нравственная ответственность не сводится к обвинению в интересах. Интересы могут быть понятными: завершить работу, сохранить ресурсы, получить важное знание. Вопрос в том, заменяют ли они основание утверждения и скрывают ли последствия выбора. Содержательное обсуждение позволяет сохранять цели, одновременно удерживая границы данных. В этой статье не предлагается новая обязательная система надзора. Предлагается мыслительная проверка собственного решения: останется ли объяснимой его связь с основаниями, если удобный исход изменится, и сможет ли другой участник увидеть эту связь без доверия к общему благому намерению.

IX. Ошибка становится понятнее, когда названа цепочка последствий

Вернёмся к вопросу заголовка. Ошибиться в признании чужого сознания значит не только неверно употребить слово. Возможная ошибка состоит в определённом выводе при определённых основаниях; её практический смысл возникает через действия и сообщения, которые из него следуют. Мы можем ошибиться в существовании опыта, в его характере, в оценке меры или в публичном представлении неопределённости. Эти ошибки могут соединяться, но не должны автоматически считаться одной. Отдельное описание делает исправление точнее: изменить утверждение, пересмотреть действие или уточнить сообщение.

Бентам, отчёт 2023 года и предложение 2025 года помогают удержать разные элементы. Моральная значимость страдания не равна доказательству его наличия. Теоретический признак не равен полноценной человеческой истории переживания. Организационная осторожность не равна эмпирическому результату. Рекомендация авторов не равна действующему закону. Сопоставление работает потому, что сохраняет эти границы, а не потому, что обещает универсальное согласие разных подходов. Для конкретного решения останутся фактические и нормативные вопросы, которые нужно исследовать в собственной области, не подменяя их сходством общих слов.

Удобный итоговый документ может содержать пять вещей: утверждение о том, что известно, существенное допущение, рассматриваемое последствие, выбранный шаг и условие пересмотра. Это предложенная здесь форма рассуждения, а не новый внешний регламент. Её ценность в проверяемой связи между частями. Если шаг не следует из описанного риска, это станет заметно. Если сообщение сильнее данных, условие поможет увидеть разрыв. Если все результаты одинаково поддерживают первоначальный вывод, можно спросить, действительно ли предусмотрен пересмотр. Такая запись делает неопределённость предметом работы, а не декоративной оговоркой.

Ответственность при неполном знании требует и внимания к возможному чужому опыту, и точности в отношении известных обязанностей перед людьми. Эти требования не нужно превращать в заранее решённую борьбу. Можно продолжать исследование, выбирать обоснованные ограничения и сообщать пределы результата. Вопрос остаётся открытым там, где основания не дают окончательного ответа. Но открытость совместима с ясным следующим шагом. Ошибка будет меньше скрыта от обсуждения, если её возможная предпосылка и цепочка последствий названы до того, как признание или отрицание станет удобной публичной историей.

Что это значит на практике

Разделите обсуждаемое заявление на предмет и действие. Приписывается ли опыт вообще, определённое переживание или способность сообщить о нём? Что предлагается сделать из-за этого? Запишите основание и существенную теоретическую предпосылку. Если речь о конкретной системе, её данные должны быть предъявлены отдельно: эта статья не заменяет такую оценку.

Опишите возможные последствия обоих направлений ошибки без придуманных чисел и заранее объявленной одинаковости. Для каждого назовите затронутую сторону и связь с выбранным действием. Сохраните отдельные обязанности перед людьми, включая точность сообщения и надёжность результата. Не выдавайте отсутствие окончательного доказательства за разрешение игнорировать любой риск и не выдавайте осторожность за факт страдания.

Выберите следующий шаг по недостающему основанию. Ограниченная оценка, изменение масштаба или дополнительное обсуждение должны иметь цель и понятное условие пересмотра. Назовите также издержки и то, чего этот шаг не установит. Участие независимого специалиста имеет смысл через конкретную проверку, а не как украшение вывода. Предложенные меры не объявляются универсальным законом.

Перед публичным сообщением сравните его силу с данными. Сохранены ли условие, область проверки и различие между целью и результатом? Скажите, кто пересмотрит вывод при новом свидетельстве и кому понадобится существенная поправка. Хорошее сообщение может быть коротким: оно должно позволять понять не только что решено, но и почему это решение ещё не равно окончательному знанию о чужом опыте.

Ключевая цитата

> *Первой задачей организаций в коммуникации должно быть недопущение введения публики в заблуждение.*

Patrick Butlin и Theodoros Lappas, *Principles for Responsible AI Consciousness Research*, arXiv:2501.07290v1, 13 января 2025, §4.5, с. 13, первый абзац. Полное предложение; перевод United Wisdom с английского. Нормативная рекомендация авторов об ответственности сообщения, а не установленный закон или доказательство эффективности предложенной политики.

Связи между традициями

Источник и точное место A	Источник и точное место B	Механизм сопоставления	Граница различия
Бентам, XVII, §1 IV, прим. 122	Батлин и коллеги, §4.1.1, опыт и окраска	Отделить нравственный вопрос о страдании от оценки интеллектуальной способности.	Исторический критерий не доказывает искусственное переживание.
Батлин и коллеги, §4.1.1, недостаточное приписывание	Батлин и Лаппас, §3.1–3.2, моральные последствия	Связать условное признание опыта с описанием возможного вреда.	Нормативный риск не объявляется наблюдённым страданием конкретной системы.
Батлин и коллеги, §4.1.2, избыточное приписывание	Батлин и Лаппас, §4.5, коммуникация	Рассмотреть влияние необоснованного сообщения на решения людей.	Нельзя выводить одинаковые реакции аудитории или фиксированную цену ошибки.
Бентам, прим. 122, различие речи и страдания	Батлин и Лаппас, §4.2, условия развития	Отделить значение заботы от конкретного способа действия.	Предложение об исследованиях не является действующим универсальным правилом.
Батлин и коллеги, §1.2.1, рабочая предпосылка	Батлин и Лаппас, §4.3, постепенность и оценки	Сохранить теоретическое условие в решении о следующем этапе.	Этапность не устанавливает истинность теории и не доказывает эффективность мер.
Батлин и коллеги, §4.1–4.2, разные риски и способности	Батлин и Лаппас, §5–6, стимулы и ограничения	Разделить содержание оценки и интерес к удобному выводу.	Наличие предложенной политики не означает её внедрение или независимую проверку.

Первоисточники

1. Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation*, Clarendon Press, Oxford, 1907, переиздание исправленной редакции 1823 года. Прочитаны гл. XVII, §1 II–VI и полное прим. 122 к IV. Заключительный вопрос сверён также в параллельном тексте. Используется нормативный критерий; остальные исторические суждения о животных не принимаются как современные данные.

2. Patrick Butlin, Robert Long, Eric Elmoznino и коллеги. *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*, arXiv:2308.08708v3, 22 августа 2023. Прочитаны указанные в подготовке определения и предпосылки, признаки GWT и пример архитектур; для этой темы полностью прочитаны и использованы §4.1–4.2. Это условный теоретический анализ, не современная проверка конкретного продукта. Названные авторами эмпирические работы независимо не проверялись.

3. Patrick Butlin, Theodoros Lappas. *Principles for Responsible AI Consciousness Research*, arXiv:2501.07290v1, 13 января 2025. Прочитаны введение и §2, полностью §3.1–3.2, таблица 1, §4.2–4.5, §5–6 и сообщение о финансировании Conscium. §4.1 не заявляется как полностью прочитанный. Это нормативное предложение с оговорками, не принятый закон и не экспериментальная проверка мер. Русские цитаты — перевод United Wisdom с английского.

Теги

сознание, чужой опыт, искусственный интеллект, признание, отрицание, ошибка, страдание, валентность, нравственная ответственность, Бентам, Батлин, Лаппас, предосторожность, неопределённость, исследование, последствия, коммуникация, публичное сообщение, теория, данные, проверка, пересмотр, организационные стимулы, права людей, благополучие, ресурсы, доверие, границы вывода

Про авторов

Andrey Zarinsh — автор и основатель проекта United Wisdom. Исследует темы внутренней трансформации, единства традиций, межкультурного диалога и современной духовной дисциплины. Instagram: @andrey.zarinsh.

Maria Evilkut — соавтор United Wisdom. Помимо исследовательской работы, она стремится проверять эти идеи на практике и наблюдать, как принципы единства и взаимопонимания работают в реальной жизни между людьми и культурами. Instagram: @ilkutki.