ДИСКУРС. 2026. Т. 12, № 5. С. 5–17
DISCOURSE. 2026, vol. 12, no. 5, pp. 5–17

Философия
Philosophy

Оригинальная статья
УДК 165
http://doi.org/10.32603/2412-8562-2026-12-5-5-17

ИИ как эпистемический авторитет: доверяем ли мы алгоритму больше, чем человеку?

Александр Михайлович Жаров

Института философии РАН, Москва, Россия
aleks.zharoff2016@yandex.ru, https://orcid.org/0000-0001-9082-3446

Введение. Статья посвящена исследованию феномена ИИ как нового типа эпистемического авторитета. Цель работы – систематический эпистемологический анализ доверия к алгоритму как источнику знания и выявление его последствия для познающего субъекта и общества. Актуальность исследования определяется стремительным распространением больших языковых моделей в качестве универсального советника по вопросам, прежде требовавшим обращения к эксперту-человеку.

Методология и источники. Исследование опирается на концептуальный аппарат социальной эпистемологии, когнитивной психологии и философии техники. Используются концепции рационального доверия к авторитету (Хардвиг: компетентность и благонамеренность как условия доверия; Загзебски: интеллектуальные добродетели), феномен автоматизационного уклона из области когнитивной психологии, а также критический анализ структурных предвзятостей ИИ-систем. Исследование носит теоретический характер и реализуется через последовательный анализ пяти тезисов об эпистемическом статусе ИИ.

Результаты и обсуждение. Установлено, что ИИ занимает позицию эпистемического авторитета на основании утилитарной эффективности, а не субъектных добродетелей. Иллюзия объективности и автоматизационный уклон обусловливают склонность доверять алгоритму в ущерб человеческой экспертизе. Фундаментальная сложность интерпретации (феномен «черного ящика») LLM существенно ограничивает классическое условие рационального доверия – возможность независимой верификации пути получения вывода. Систематическое делегирование мышления алгоритму ведет к атрофии критических способностей и деквалификации субъекта. Медиаторная концепция ИИ как агрегатора коллективного знания признается перспективной, однако ограниченной вследствие избирательности данных и рассеянной ответственности.

Заключение. ИИ формирует принципиально новый и более хрупкий тип эпистемических отношений. Ответом на этот вызов выступает «критическая алгоритмическая грамотность» – системная компетенция, которая становится необходимым условием сохранения когнитивной автономии в обществе, делегирующем знание алгоритму.

Ключевые слова: эпистемический авторитет, искусственный интеллект, социальная эпистемология, автоматизационный уклон, черный ящик, алгоритмическая грамотность, большие языковые модели

© Жаров А. М., 2026

Контент доступен по лицензии Creative Commons Attribution 4.0 License.
This work is licensed under a Creative Commons Attribution 4.0 License.

ДИСКУРС. 2026. Т. 12, № 5. С. 5–17
DISCOURSE. 2026, vol. 12, no. 5, pp. 5–17

Для цитирования: Жаров А. М. ИИ как эпистемический авторитет: доверяем ли мы алгоритму больше, чем человеку? // ДИСКУРС. 2026. Т. 12, № 5. С. 5–17. DOI: 10.32603/2412-8562-2026-12-5-5-17

Original paper

AI as an Epistemic Authority: do We Trust an Algorithm More than a Human?

Alexander M. Zharov

Institute of Philosophy of the RAS, Moscow, Russia
aleks.zharoff2016@yandex.ru, https://orcid.org/0000-0001-9082-3446

Introduction. The article is devoted to the study of the phenomenon of AI as a new type of epistemic authority. The purpose of the work is to conduct a systematic epistemological analysis of trust in the algorithm as a source of knowledge and to identify its consequences for the cognizing subject and society. The relevance of the research is determined by the rapid spread of large language models (LLMs) as a universal adviser on issues that previously required consulting a human expert.

Methodology and sources. The research is based on the conceptual framework of social epistemology, cognitive psychology and philosophy of technology. The study uses the concepts of rational trust in authority (Hardwig: competence and well-intentionedness as conditions of trust; Zagzebski: intellectual virtues), the phenomenon of automation bias from the field of cognitive psychology, as well as a critical analysis of the structural biases of AI systems. The research is theoretical in nature and is conducted through a consistent analysis of five theses on the epistemic status of AI.

Results and discussion. It has been established that AI occupies the position of epistemic authority based on utilitarian effectiveness rather than subjective virtues. The illusion of objectivity and the automation bias lead to a tendency to trust the algorithm to the detriment of human expertise. The fundamental difficulty of interpreting LLMs (the black box phenomenon) significantly limits the classical condition of rational trust – the possibility of independent verification of how a conclusion was reached. The systematic delegation of thinking to an algorithm leads to the atrophy of critical abilities and the deskilling of the subject. The mediation conception of AI as an aggregator of collective knowledge is considered promising, but limited due to the selectivity of data and dispersed responsibility.

Conclusion. AI forms a fundamentally new and more fragile type of epistemic relationship. The answer to this challenge is "critical algorithmic literacy", a systemic competence that becomes a prerequisite for maintaining cognitive autonomy in a society that delegates knowledge to an algorithm.

Keywords: epistemic authority, artificial intelligence, social epistemology, automation bias, black box, algorithmic literacy, large language models

For citation: Zharov, A.M. (2026), "AI as an Epistemic Authority: do We Trust an Algorithm More than a Human?", *DISCOURSE*, vol. 12, no. 5, pp. 5–17. DOI: 10.32603/2412-8562-2026-12-5-5-17

Введение. История знания – это в значительной мере история эпистемического авторитета. Просвещение произвело здесь фундаментальную революцию: жреца и оракула сменил ученый и врач. Новый авторитет был основан на методе – воспроизводимом, проверяемом, открытом для критики. Большинство из нас не способно самостоятельно верифицировать прописанный антибиотик или расчет нагрузки на мост, но мы доверяем системе, которая производит и верифицирует экспертов [1, p. 335]. Эксперт заслуживал доверия именно по-

тому, что его знание получено через надежную процедуру и могло быть проверено другими специалистами.

Сегодня мы наблюдаем новый, возможно, столь же принципиальный сдвиг. Миллионы людей ежедневно обращаются за ответами к большим языковым моделям (LLM) – ChatGPT, Gemini, Claude по вопросам, прежде требовавшим обращения к эксперту-человеку: о здоровье и лечении, праве и финансах, моральных дилеммах. ИИ превращается в универсального советника и тем самым в новый тип эпистемического авторитета. Является ли этот авторитет легитимным? Основан ли он на тех же критериях, что и доверие к человеческим экспертам, или мы имеем дело с принципиально иным и, возможно, опасным типом эпистемических отношений?

Цель статьи – систематический эпистемологический анализ феномена ИИ как эпистемического авторитета, опираясь на аппарат социальной эпистемологии [2; 3], когнитивной психологии и философии техники. Исследование структурировано вокруг пяти тезисов:

1) ИИ смещает основание авторитета от интенциональности к эффективности;

2) иллюзия объективности и автоматизационный уклон обусловливают чрезмерное доверие алгоритму;

3) фундаментальная сложность интерпретации (феномен «черного ящика») LLM существенно ограничивает классическое условие рационального доверия;

4) передача авторитета ИИ ведет к деквалификации познающего субъекта;

5) ИИ является не автономным авторитетом, а особым типом знаниевого посредника.

Методология и источники. Исследование опирается на концептуальный аппарат социальной эпистемологии, когнитивной психологии и философии техники. Методологическую основу составляют: концепция рационального доверия к авторитету, разработанная Дж. Хардвигом [1; 4] (компетентность и благонамеренность как базовые условия рационального доверия); теория интеллектуальных добродетелей Л. Загзебски [5; 6] (интенциональная ориентация на истину как конститутивный признак эпистемического авторитета); феномен автоматизационного уклона из области когнитивной психологии и человеко-машинного взаимодействия (Л. Скитка [7], Р. Парасураман [8]); критический анализ структурных предвзятостей ИИ-систем в рамках философии техники (К. Кроуфорд [9]).

Исследование носит теоретический характер и реализуется через последовательный концептуальный анализ пяти тезисов об эпистемическом статусе ИИ. В качестве эмпирического материала привлекаются результаты экспериментальных исследований в области когнитивной психологии и человеко-машинного взаимодействия [10, 8], данные о пользовательских практиках взаимодействия с LLM, а также работы по технической архитектуре языковых моделей [10–12]. Источниковую базу составляют работы в области социальной эпистемологии [2; 3; 13], когнитивной науки и философии техники.

Результаты и обсуждение. В ходе исследования установлено, что ИИ занимает позицию эпистемического авторитета на принципиально иных основаниях, чем традиционные человеческие авторитеты: его легитимность строится на утилитарной эффективности сервиса, а не на добродетелях субъекта – честности, ответственности, ориентации на истину. Это означает переход от субъектно-нормативной к инструментально-прагматической модели доверия, что создает структурно более хрупкую эпистемическую конструкцию.

Показано, что иллюзия объективности и феномен автоматизационного уклона обусловливают склонность людей доверять алгоритму больше, чем эксперту-человеку, вопреки рациональным основаниям. Невидимая системная предвзятость LLM оказывается эпистемически опаснее видимой человеческой субъективности, поскольку против нее не выработано культурных иммунных механизмов. Фундаментальная сложность интерпретации (феномен «черного ящика») LLM существенно ограничивает классическое условие рационального доверия – возможность независимой верификации пути получения вывода, превращая алгоритм в оракула, а не в эксперта.

Систематическое делегирование познавательных функций алгоритму ведет к атрофии критического мышления и формирует деструктивную петлю обратной связи: чем больше пользователь доверяет ИИ, тем хуже он способен оценивать качество его ответов. Медиаторная концепция ИИ как агрегатора коллективного человеческого знания признается концептуально продуктивной, однако принципиально ограниченной вследствие избирательности обучающей выборки, трансформации при синтезе и практического коллапса распределенной ответственности.

Трансформация оснований авторитета: от интенциональности к эффективности. *Классическая модель эпистемического авторитета.* Социальная эпистемология давно занимается вопросом рациональности доверия авторитетам [2; 3; 13]. Дж. Хардвиг показал, что в современном обществе большая часть знания является «опосредованной»: нам известно многое, но не потому, что мы сами провели исследование, а потому, что приняли свидетельство других [1, p. 336]. Это не слабость, а структурная необходимость: разделение познавательного труда является условием возможности науки и экспертизы [4, p. 694]. Рациональное доверие рационально тогда, когда есть основания считать, что свидетель обладает компетентностью (доступом к знанию через надежную процедуру) и благонамеренностью (стремлением точно передать это знание).

Л. Загзебски добавляет к этой модели важное измерение: подлинный авторитет предполагает не просто компетентность, но правильную эпистемическую ориентацию – стремление к истине как самоценности, «интеллектуальные добродетели»: честность, открытость, смирение, мужество в отстаивании истины [5, p. 15; 6, p. 7]. Именно эта интенциональная ориентация делает авторитет эпистемически надежным партнером. Данная модель конститутивно ориентирована на человеческого субъекта: авторитет является авторитетом потому, что он – субъект с определенными намерениями, ценностями и ответственностью.

Как ИИ разрушает классическую модель. ИИ принципиально не вписывается в описанную модель – не в смысле «плохо соответствует», а в смысле принципиально иной природы отношения. Когда пользователь спрашивает у ChatGPT о симптомах болезни, он не оценивает благонамеренность системы, и не может. Понятие «хочет ли ИИ меня обмануть?» бессмысленно применительно к языковой модели так же, как бессмысленно спрашивать, хочет ли нас обмануть калькулятор. У системы нет целей, ценностей, ориентации на истину [10; 14]. Это касается любой LLM независимо от ее возможностей.

Тем не менее люди доверяют ответам ИИ, и это доверие основано на ином принципе: эффективности. Пользователь доверяет системе, потому что она «обычно дает правильные ответы», «не подводила раньше». Это прагматическое, утилитарное доверие к инструменту,

а не к агенту. Этот сдвиг принципиален: ошибка инструмента – технический сбой, а не предательство. Нарушение нормативного контракта между знающим субъектом и доверяющим ему не происходит – значит, не активируются и механизмы социальной коррекции: порицание, потеря репутации, институциональные санкции. Переход к инструментальному доверию денормативизирует отношения с источником знания.

Утилитарный поворот и его риски. Утилитарная легитимация авторитета – «доверяю, потому что работает» – сама по себе не является пороком. Если система стабильно дает надежные результаты, рациональность ее использования не вызывает сомнений. Однако она становится проблематичной при условиях, характерных для LLM: когда надежность системы неоднородна и непредсказуема. Термометр либо работает, либо нет. LLM работает отлично в одних областях и систематически ошибается в других, не подавая никакого сигнала о смене режима [11]. Это создает специфический риск «трансфера доверия»: пользователь, выработавший доверие на основе успешных взаимодействий в одном домене (запросов о рецептах или рекомендациях фильмов), переносит его на принципиально иные области – медицинские или юридические советы. ИИ-система воспринимается как универсально компетентная, что нарушает нормальные механизмы калибровки доверия: классический авторитет имеет четкие границы компетентности (врач – в медицине, но не в юриспруденции), тогда как LLM таких границ не маркирует.

Иллюзия объективности и автоматизационный уклон. *Откуда берется иллюзия объективности.* Одним из наиболее устойчивых и опасных когнитивных заблуждений в отношении ИИ-систем является иллюзия объективности – убеждение в том, что ответ алгоритма лишен субъективных искажений. Психологические корни иллюзии хорошо изучены. Во-первых, машинность: у людей существует укорененная ассоциация между «машинным» и «точным» – механические часы не ошибаются, компьютер считает без арифметических ошибок. Эта ассоциация переносится на LLM, хотя принципиально неприменима: языковые модели – вероятностные генераторы, а не детерминированные вычислительные машины. Во-вторых, отсутствие видимых признаков субъективности: LLM не проявляет усталости, раздражения, энтузиазма по поводу любимой теории, ее ответы выглядят одинаково ровно и уравновешенно, что создает впечатление бесстрастности, ассоциируемой с объективностью.

Между тем иллюзия объективности скрывает принципиальную предвзятость иного рода. LLM не является нейтральным зеркалом реальности – она артефакт, встроенный в конкретный социальный, культурный и экономический контекст. Выбор обучающих данных, их состав, предобработка, функции потерь при обучении, системы безопасности, наложенные разработчиком, – все это определяет, какие точки зрения представлены в ответах. К. Кроуфорд детально документирует механизмы структурной предвзятости: гендерные и расовые перекосы в обучающих данных, отражение доминирующих культурных нарративов, экономические интересы корпораций-разработчиков [9, p. 14]. ИИ не объективнее человека – он предвзят по-другому, но эта предвзятость скрыта за фасадом ровного уверенного текста без видимых маркеров субъективности.

Автоматизационный уклон: психология чрезмерного доверия. Иллюзия объективности получает психологическое подкрепление через феномен автоматизационного уклона (automation bias) – систематическую тенденцию переоценивать рекомендации автоматизированных систем и недооценивать собственное суждение, когда оно противоречит машин-

ному [7, p. 991]. Феномен описан в работах Л. Скитка и коллег, исследовавших поведение операторов в системах высокой автоматизации: в авиации, медицине, управлении атомными станциями. Операторы систематически следовали рекомендациям автоматизированных систем, даже когда те были ошибочными: пилоты следовали ошибочным командам автопилота; врачи принимали неверные диагнозы, игнорируя клинические наблюдения [8, p. 381].

Механизм уклона включает несколько компонентов. Когнитивная экономия: обработка машинной рекомендации требует меньше ресурсов, чем самостоятельный анализ, – в условиях нагрузки мозг предпочитает «передать» принятие решения машине. Диффузия ответственности: если машина рекомендует X, ответственность за ошибку субъективно «переходит» к ней. Эффект ореола: зарекомендовавшая себя система распространяет репутацию надежности на новые применения, принципиально отличающиеся по достоверности. В контексте LLM уклон особенно остр: система производит не просто сигнал, а убедительно аргументированный текст, организованный как авторитетное экспертное свидетельство. Такой текст активирует доверие не только через механизм автоматизации, но и через нарративное убеждение – он «звучит» как эксперт. Кроме того, диалоговый формат взаимодействия с LLM активирует те же когнитивные схемы, что и общение с человеком: собеседник интерпретируется как агент с намерениями и ответственностью.

Паттерны автоматизационного уклона в контексте LLM. Применительно к LLM автоматизационный уклон порождает три характерных паттерна. Первый – некритическое принятие: получив убедительно оформленный ответ, пользователь, как правило, не проверяет его по независимым источникам, даже когда речь идет о принципиально важных для принятия решений вопросах. Второй – обесценивание человеческой экспертизы: если ответ эксперта расходится с ответом LLM, пользователи нередко принимают сторону алгоритма – врач, чья оценка расходится с «рекомендацией» ИИ-диагностики, может испытывать давление следовать машинной рекомендации. Третий – избирательная проверка: ответы LLM проверяются, когда расходятся с ожиданиями, но не когда приятны или подтверждают имеющиеся убеждения. Это создает риск «комфортной дезинформации». В совокупности три паттерна формируют эпистемическую экосистему высокого риска, требующую системного образовательного ответа, а не просто индивидуальных предупреждений.

Эпистемический «черный ящик»: авторитет без прозрачности. *Прозрачность как условие рационального доверия.* В классической модели эпистемического авторитета прозрачность играет ключевую роль не как требование лично проверять каждый шаг эксперта, а как принципиальная возможность объяснения. Эксперт способен изложить резоны своего суждения; они поддаются критике; при выявлении ошибки возможна коррекция. Хардвиг подчеркивает: рациональное доверие основано на убеждении, что авторитет является рациональным агентом, способным дать отчет в своих суждениях [1, p. 340]. Достаточно знать, что такой отчет в принципе возможен. Именно поэтому институты научной экспертизы включают требования к методологической документации, рецензированию, воспроизводимости – они создают инфраструктуру потенциальной прозрачности. Знание, которое нельзя в принципе проверить и воспроизвести, не является знанием в строгом смысле.

ИИ как новый оракул. LLM существенно затрудняет выполнение условия потенциальной прозрачности в ее классическом, процедурном понимании. Хотя область Explainable AI (XAI)

активно развивается и предлагает методы для визуализации зон внимания трансформеров или построения упрощенных суррогатных моделей, в случае с генерацией свободного текста на сотни миллиардов параметров эти объяснения остаются вероятностными картами, а не логической реконструкцией рассуждения. Даже разработчики, имея полный доступ к весам модели, не могут дать однозначного ответа на вопрос, почему в данном конкретном случае модель предпочла факт А факту Б, а не сгенерировала галлюцинацию Х. Таким образом, «черный ящик» в контексте LLM – это не столько полное отсутствие доступа, сколько отсутствие семантически интерпретируемой причинно-следственной цепи, что и ограничивает классическое условие рационального доверия [12, p. 5]. Нейронная сеть с миллиардами параметров – не алгоритм в классическом смысле, проследить ее шаги невозможно. Авторитет без возможности объяснения – это не экспертный авторитет Просвещения, а нечто ближе к авторитету оракула. Аналогия содержательна: Дельфийский оракул принимался как авторитет не потому, что его слова были прозрачны, а потому что воспринимались как исходящие из источника, превосходящего человеческое. Нечто аналогичное происходит с ИИ: обучение на колоссальных массивах человеческих текстов воспринимается как источник трансцендентной компетентности, а непрозрачность становится не проблемой, а признаком превосходства.

К. Кроуфорд описывает ИИ-системы как «черные ящики власти»: они производят решения, влияющие на жизни людей, не давая никакого объяснения и не оставляя пути для апелляции [9, p. 212]. В эпистемологическом смысле это авторитет без ответственности – возможно, самая опасная форма авторитета.

Тем не менее было бы неверно утверждать, что LLM остаются абсолютно непрозрачными. Современные исследования постепенно открывают «черный ящик», предлагая инструменты, которые позволяют если не полностью понять, то хотя бы частично проинтерпретировать поведение модели. Так, методы анализа внимания дают возможность отследить, какие именно фрагменты входного текста оказали наибольшее влияние на сгенерированный ответ. Другой подход связан с техниками стимуляции нейронов: исследователи учатся автоматически выявлять интерпретируемые паттерны в активациях сети, что позволяет заглянуть во внутренние представления, которые модель формирует о мире. Наконец, методы контрастного объяснения показывают, почему модель предпочла один ответ другому, выявляя ключевые различия между альтернативами.

Таким образом, вопрос об объяснимости ИИ не сводится к дихотомии «полностью прозрачный–абсолютно непрозрачный». Скорее, речь идет о спектре возможностей, которые постоянно расширяются. Для рационального эпистемического доверия достаточно необязательно полной прозрачности, а принципиальной возможности получить осмысленное объяснение в конкретном случае. По этому критерию современные LLM занимают промежуточное положение: они не являются «оракулами» в строгом смысле, но и не соответствуют идеалу просвещенческого эксперта. Это требует от пользователя более сложной калибровки доверия с учетом того, для какого типа решения и какого качества объяснения доступны.

Невозможность «перекрестного допроса». К проблеме непрозрачности примыкает проблема невозможности подлинного критического диалога с ИИ. В классической эпистемологии уточняющий допрос – ключевой инструмент верификации авторитета: дополнительные вопросы раскрывают структуру знания эксперта, выявляют пробелы, проверяют устойчи-

вость позиции к контраргументам. С LLM этот инструмент работает неоднозначно. С одной стороны, к 2026 г. ситуация, описанная в обзорах трехлетней давности [11], претерпела значительные изменения. Современные модели (включая новейшие версии Gemini, GPT-5 и Claude) демонстрируют значительно более высокую когерентность в длинных диалогах и снабжены механизмами верификации (например, поисковые агенты на базе LLM или встроенные модули фактчекинга). Это частично возвращает пользователю инструмент перекрестного допроса: запрос «объясни шаг № 3 подробнее» или «приведи источник для этого утверждения» уже не является имитационным, а ведет к реальной детализации или самокоррекции.

Однако эпистемическая проблема не снимается полностью, а смещается на метауровень. Во-первых, «место разлома» по-прежнему не маркируется самой системой: в областях с низкой плотностью обучающих данных (редкие языки, узкоспециальные научные дискуссии) модель может генерировать правдоподобный, но ложный ответ с той же интонационной уверенностью, что и в проверенных областях. Во-вторых, возникает риск «иллюзии верификации»: пользователь, получивший ссылку на источник, сгенерированный самой же LLM, или краткую выжимку из поиска, может ошибочно принять процедурную активность системы за содержательную проверку истинности. Таким образом, перекрестный допрос становится возможным технически, но требует от пользователя значительно большей эпистемической бдительности, чем при диалоге с экспертом-человеком, чья компетентность и границы некомпетентности социально маркированы. Современные методы взаимодействия (например, техника «цепочки рассуждений», повторные запросы с переформулированием, использование LLM для самопроверки) позволяют частично повысить результативность такого «допроса», выявляя некоторые внутренние противоречия. Однако эти приемы не устраняют фундаментальной проблемы: у LLM нет связной и стабильной системы убеждений, которую можно было бы проверить на когерентность так же, как у человека. Поэтому «перекрестный допрос» к ИИ остается имитацией, а не полноценной эпистемической процедурой.

Необходимо признать: за последние три года достигнут заметный прогресс в снижении частоты галлюцинаций. Тем не менее даже самые современные LLM продолжают галлюцинировать, особенно в областях, требующих актуальных, узкоспециализированных или логически сложных рассуждений. Проблема не решена кардинально, а ее частичное смягчение не отменяет эпистемической опасности «черного ящика».

Эпистемическая лень и деквалификация познающего субъекта. *Экстернализация мышления и ее когнитивные последствия.* История цивилизации – это во многом история экстернализации когнитивных функций: от письменности, вынесшей память за пределы головы, до калькуляторов, взявших на себя арифметику [15, p. 7]. Каждый шаг сопровождался тревогами об атрофии соответствующей способности – и эти тревоги отчасти оправдывались. Устная традиция с ее феноменальными мнемоническими практиками постепенно деградировала; устный счет сократился. Но каждый раз изменения компенсировались освобождением когнитивных ресурсов для задач более высокого уровня. Этот компенсаторный механизм хорошо документирован: экстернализация памяти в цифровые хранилища [16] высвобождает ресурсы для концептуального мышления [17].

Вопрос о том, что происходит при экстернализации мышления через LLM, актуален именно потому, что LLM берет на себя функции, компенсация которых по описанному об-

разцу невозможна или крайне затруднена. Речь идет не о памяти или арифметике, а о рассуждении, аргументации, критическом анализе, синтезе. Если эти функции экстернализированы в алгоритм, что остается «внутри»?

Атрофия критического мышления. Наиболее очевидный риск – атрофия критического мышления как практической навыковой способности. Критическое мышление – не врожденный дар, а приобретенный навык: умение формулировать вопросы, оценивать качество аргументов, выявлять логические ошибки, проверять факты по первоисточникам. Все эти умения деградируют без практики. Передача задач, требующих критического мышления, алгоритму ведет к снижению частоты практики и постепенному ослаблению навыка. Исследования фиксируют снижение способности студентов к самостоятельному аргументированию и анализу первоисточников в условиях широкой доступности ИИ-инструментов [17].

Эта ситуация создает опасный парадокс: чем больше люди доверяют ИИ, тем хуже они способны оценивать качество его ответов. Компетентный специалист распознает галлюцинацию LLM быстрее и точнее чем неофит именно потому, что у него есть независимая система знания, с которой можно сравнивать. Систематическое использование LLM вместо самостоятельного мышления постепенно лишает пользователя собственного критического мышления – петля обратной связи с деструктивным потенциалом.

Централизация знания и хрупкость эпистемического сообщества. Научное знание всегда создавалось и поддерживалось сообществом: ошибка одного участника сети корректируется другими, монополия на истину невозможна, знание устойчиво к сбоям отдельных узлов [2, p. 32]. Доминирование нескольких LLM как источников знания для миллионов людей создает принципиально иную, централизованную и хрупкую архитектуру. Концентрация эпистемического влияния несет риски цензуры, манипуляции, системных ошибок, одновременно распространяющихся на весь массив пользователей. Эпистемическое неравенство при этом усугубляет несправедливость: те, кто лишен доступа к альтернативным источникам, оказываются в полной зависимости от алгоритмического «канона» [18, p. 1].

Ситуация структурно аналогична историческим монополиям на знание. Средневековый контроль церкви над книгопечатанием создавал именно такую хрупкую конструкцию: ошибка, вошедшая в канон, не корректировалась – у общества не было механизмов ее исправления [19, p. 138]. Реформация стала отчасти ответом на эту монополию; распределенная природа современной науки является институциональным ответом на тот же урок. Возврат к централизации через доминирование LLM воспроизводит те же уязвимости в принципиально новом технологическом контексте.

ИИ как посредник: компромиссная концепция. *ИИ как агрегатор коллективного знания.* После развернутой критики необходимо рассмотреть альтернативную концептуальную рамку. ИИ-система обучена на колоссальном корпусе текстов, созданных людьми. Когда LLM отвечает на вопрос, она не создает знание самостоятельно – она извлекает, синтезирует и перефразирует то, что было написано людьми. В этом смысле доверие ответу LLM есть опосредованное доверие коллективному интеллекту авторов, чьи тексты стоят за ответом. Это позволяет переформулировать природу эпистемического доверия к LLM: не доверие к алгоритму как к агенту, а доверие к алгоритму как к медиуму – посреднику, обеспечивающему доступ к распределенному человеческому знанию. Концепция ближе к роли высоко-

квалифицированного аналитика-медиатора, обрабатывающего огромный массив источников и извлекающего релевантное содержание.

Более того, медиаторная модель не учитывает важное эмерджентное свойство современных LLM, которое отличает их от «продвинутого T9». Если бы модель была лишь статистическим усреднением популярных ответов, она систематически воспроизводила бы распространенные заблуждения и мифы. Однако эмпирические данные (особенно актуальные на 2026 г.) показывают обратное: благодаря процедурам тонкой настройки (fine-tuning) и обучению с подкреплением на основе обратной связи (RLHF), модели демонстрируют прогресс в детекции ошибочных, хотя и частотных, взглядов. Например, на запрос о вреде прививок или эффективности гомеопатии современные LLM склонны генерировать научно-обоснованный ответ, а не наиболее часто встречающуюся в интернете риторику. Это свидетельствует о наличии в архитектуре их «знания» не только статистического паттерна, но и иерархии авторитетности источников, привитой в ходе антропогенного обучения. Следовательно, ИИ в роли посредника не просто отражает коллективный разум, но и избирательно его фильтрует в соответствии с определенной (западной, академической) эпистемической нормой, что одновременно является и преимуществом в борьбе с дезинформацией, и риском гомогенизации знания.

Проблемы медиаторной модели. Медиаторная концепция сталкивается с тремя серьезными ограничениями. Во-первых, проблема репрезентативности: обучающая выборка LLM перекошена в сторону английского языка и западных культурных контекстов, недопокрывает знание в устной форме или в специализированных профессиональных практиках [9, p. 89]. Когда мы «доверяем человечеству через LLM», мы на деле доверяем избирательному и предвзятому срезу человеческой письменности.

Во-вторых, проблема трансформации при синтезе: LLM трансформирует тексты, производя новый текст на основе статистических закономерностей. Этот процесс непрозрачен – он может смешивать четко противопоставленные позиции, нивелировать принципиальные нюансы, производить синтезы, которые ни один из авторов не одобрил бы. В-третьих, проблема скрытой авторизации: пользователи не воспринимают ответ LLM как синтез чужих текстов – они воспринимают его как высказывание системы. Авторство стягивается к алгоритму, а не распределяется между людьми-авторами, что прямо противоречит медиаторной логике.

Распределенная ответственность и ее практический коллапс. Медиаторная концепция предполагает «распределенный» авторитет LLM: за каждым ответом стоит ответственность разработчиков, инженеров, кураторов данных, авторов исходных текстов. Ответственность, распределенная без механизмов ее атрибуции, в конкретном случае функционально эквивалентна отсутствию ответственности. Если LLM выдала ложное медицинское утверждение, кто несет ответственность? Разработчик? Куратор данных? Автор первоисточника? Ответ «все и никто» – институциональный провал эпистемической системы.

Классический человеческий авторитет был надежен в том числе потому, что ответственность была концентрирована и атрибутируема: врач, давший неверный совет, несет профессиональную ответственность; ученый, опубликовавший фальсификацию, сталкивается с ретракцией и репутационными последствиями. Эта концентрированная ответственность создавала стимулы к качеству. Рассеянная ответственность LLM этих стимулов не создает – или создает их в принципиально ослабленном виде.

Заключение. Выполненный анализ позволяет сформулировать пять взаимосвязанных выводов о феномене ИИ как эпистемическом авторитете.

Во-первых, ИИ занимает позицию эпистемического авторитета в современной культуре на принципиально иных основаниях, чем традиционные авторитеты: его легитимность строится на утилитарной эффективности сервиса, а не на добродетелях субъекта. Это прагматический поворот в структуре доверия, создающий более хрупкую конструкцию.

Во-вторых, иллюзия объективности и автоматизационный уклон обусловливают склонность доверять алгоритму больше, чем эксперту-человеку, вопреки рациональным основаниям. Невидимая системная предвзятость LLM опаснее видимой человеческой субъективности, против которой выработаны культурные иммунные механизмы.

В-третьих, непрозрачность ИИ как «черного ящика» ограничивает классическое условие рационального доверия – возможность независимой верификации пути получения вывода. Мы имеем дело с авторитетом без ответственности, структурно аналогичным оракулу, а не эксперту. Невозможность «перекрестного допроса» делает верификацию знания, полученного от LLM, задачей принципиально иного уровня сложности.

В-четвертых, систематическое делегирование познавательных функций алгоритму ведет к атрофии критического мышления, замыкая петлю деградации: чем больше мы доверяем ИИ, тем хуже мы способны его проверять. Концентрация эпистемического авторитета в нескольких LLM-системах воспроизводит хрупкость исторических монополий на знание.

В-пятых, медиаторная концепция ИИ как агрегатора коллективного знания концептуально продуктивна, но не снимает принципиальных проблем: перекоса обучающей выборки, трансформации при синтезе и практического коллапса распределенной ответственности [20, p. 148].

Итоговый нормативный вывод: нарастающее использование ИИ как источника знания требует формирования «критической алгоритмической грамотности» – системного набора навыков и установок, позволяющих использовать ИИ продуктивно, не попадая в ловушки автоматизационного уклона, иллюзии объективности и эпистемической лени. Эта грамотность включает знание механизмов работы LLM и их ограничений, умение верифицировать ответы по независимым источникам, осознание системных предвзятостей ИИ, способность отличать область надежной работы системы от области высокого риска ошибок. Она должна стать частью образования – так же, как медиаграмотность стала ответом на вызовы информационного общества.

СПИСОК ЛИТЕРАТУРЫ

1. Hardwig J. Epistemic Dependence // The J. of Philosophy. 1985. Vol. 82, № 7. P. 335–349. DOI: https://doi.org/10.2307/2026523

2. Goldman A. I. Knowledge in a Social World. Oxford: Clarendon Press, 1999.

3. Foley R. Intellectual Trust in Oneself and Others. Cambridge: Cambridge Univ. Press, 2001.

4. Hardwig J. The Role of Trust in Knowledge // The J. of Philosophy. 1991. Vol. 88, № 12. P. 693–708. DOI: https://doi.org/10.2307/2027007

5. Zagzebski L. T. Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief. Oxford: Oxford Univ. Press, 2012. DOI: 10.1093/acprof:oso/9780199936472.001.0001

6. Zagzebski L. T. Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge. Cambridge: Cambridge Univ. Press, 1996.

7. Skitka L. J., Mosier K. L., Burdick M. Does Automation Bias Decision-Making? // Int. J. of Human-Computer Studies. 1999. Vol. 51, № 5. P. 991–1006. DOI: 10.1006/ijhc.1999.0252

8. Parasuraman R., Manzey D. H. Complacency and Bias in Human Use of Automation // Human Factors. 2010. Vol. 52, № 3. P. 381–410. DOI: 10.1177/0018720810376055

9. Crawford K. Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven: Yale Univ. Press, 2021.

10. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? / E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell // Conf. on Fairness, Accountability, and Transparency (FAccT '21), Virtual Event, Canada, 3–10 March 2021 / Canada, 2021. P. 610–623. DOI: https://doi.org/10.1145/3442188.3445922

11. Survey of Hallucination in Natural Language Generation / Z. Ji, N. Lee, R. Frieske et al. // ACM Computing Surveys. 2023. Vol. 55, iss. 12: 248. DOI: https://doi.org/10.1145/3571730

12. Russell S., Norvig P. Artificial Intelligence: A Modern Approach. 4th ed. Hoboken, NJ: Pearson, 2020.

13. Coady C. A. J. Testimony: A Philosophical Study. Oxford: Clarendon Press, 1992.

14. Turing A. M. Computing Machinery and Intelligence // Mind. 1950. Vol. 59, № 236. P. 433–460.

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

16. Carr N. The Shallows: What the Internet Is Doing to Our Brains. NY: W. W. Norton, 2010.

17. Sparrow B., Liu J., Wegner D. M. Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips // Science. 2011. Vol. 333, iss. 6043. P. 776–778. DOI: 10.1126/science.1207745

18. Fricker M. Epistemic Injustice: Power and the Ethics of Knowing. Oxford: Oxford Univ. Press, 2007.

19. Popper K. R. The Open Society and Its Enemies. London: Routledge, 1945.

20. Quine W. V. O., Ullian J. S. The Web of Belief. 2nd ed. NY: Random House, 1978.

Информация об авторе

Жаров Александр Михайлович – младший научный сотрудник сектора междисциплинарных проблем научно-технического развития Института философии РАН, Гончарная ул., д. 12, стр. 1, Москва, 109240, Россия. Автор 20 научных публикаций. Сфера научных интересов: социальная эпистемология, онтология и теория познания, эволюционная эпистемология.

О конфликте интересов, связанном с данной публикацией, не сообщалось.
Поступила 20.02.2026; принята после рецензирования 21.03.2026; опубликована онлайн 17.11.2026.

REFERENCES

1. Hardwig, J. (1985), "Epistemic dependence", *The J. of Philosophy*, vol. 82, no. 7, pp. 335–349. DOI: https://doi.org/10.2307/2026523

2. Goldman, A.I. (1999), *Knowledge in a social world*, Clarendon Press, Oxford, UK.

3. Foley, R. (2001), *Intellectual trust in oneself and others*, Cambridge Univ. Press, Cambridge, UK.

4. Hardwig, J. (1991), "The role of trust in knowledge", *The J. of Philosophy*, vol. 88, no. 12, pp. 693–708. DOI: https://doi.org/10.2307/2027007

5. Zagzebski, L.T. (2012), *Epistemic authority: a theory of trust, authority, and autonomy in belief*, Oxford Univ. Press, Oxford, UK. DOI: 10.1093/acprof:oso/9780199936472.001.0001

6. Zagzebski, L.T. (1996), *Virtues of the mind: an inquiry into the nature of virtue and the ethical foundations of knowledge*, Cambridge Univ. Press, Cambridge, UK.

7. Skitka, L.J., Mosier, K.L. and Burdick, M. (1999), "Does automation bias decision-making?", *Int. J. of Human-Computer Studies*, vol. 51, no. 5, pp. 991–1006. DOI:10.1006/ijhc.1999.0252

8. Parasuraman, R. and Manzey, D.H. (2010), "Complacency and bias in human use of automation", *Human Factors*, vol. 52, no. 3, pp. 381–410. DOI: 10.1177/0018720810376055

9. Crawford, K. (2021), *Atlas of AI: power, politics, and the planetary costs of artificial intelligence*, Yale Univ. Press, New Haven, USA.

10. Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021), "On the dangers of stochastic parrots: can language models be too big?", *Conf. on Fairness, Accountability, and Transparency (FAccT '21), Virtual Event*, Canada, 3–10 March 2021, pp. 610–623. DOI: https://doi.org/10.1145/3442188.3445922

11. Ji, Z., Lee, N., Frieske, R. et al. (2023), "Survey of hallucination in natural language generation", *ACM Computing Surveys*, vol. 55, iss. 12: 248. DOI: https://doi.org/10.1145/3571730

12. Russell, S. and Norvig, P. (2020), "Artificial intelligence: a modern approach", 4th ed., Pearson, Hoboken, NJ, USA.

13. Coady, C.A.J. (1992), *Testimony: a philosophical study*, Clarendon Press, Oxford, UK.

14. Turing, A.M. (1950), "Computing machinery and intelligence", *Mind*, vol. 59, no. 236, pp. 433–460.

15. Clark, A. and Chalmers, D.J. (1998), "The extended mind", *Analysis*, vol. 58, no. 1, pp. 7–19.

16. Carr, N. (2010), *The shallows: what the internet is doing to our brains*, W.W. Norton, NY, USA.

17. Sparrow, B., Liu, J. and Wegner, D.M. (2011), "Google effects on memory: cognitive consequences of having information at our fingertips", *Science*, vol. 333, iss. 6043, pp. 776–778. DOI: 10.1126/science.1207745

18. Fricker, M. (2007), *Epistemic injustice: power and the ethics of knowing*, Oxford Univ. Press. Oxford, UK.

19. Popper, K.R. (1945), *The open society and its enemies*, Routledge, London, UK.

20. Quine, W.V.O. and Ullian, J.S. (1978), *The web of belief*, 2nd ed., Random House, NY, USA.

Information about the author

Alexander M. Zharov – Research Assistant at the Sector of Interdisciplinary Problems of Scientific and Technological Development, Institute of Philosophy of the RAS, 12/1 Goncharnaya st., Moscow 109240, Russia. Author of 20 scientific publications. Area of expertise: social epistemology, ontology and theory of knowledge, evolutionary epistemology.

No conflicts of interest related to this publication were reported.
Received 20.02.2026; adopted after review 21.03.2026; published online 17.11.2026.