УДК 165:004.8

«ГАЛЛЮЦИНАЦИИ» ИИ КАК НОВАЯ ФОРМА ЭПИСТЕМИЧЕСКОЙ ОШИБКИ

А. А. Шевченко
Институт философии и права СО РАН (г. Новосибирск)
shev@philosophy.nsc.ru

Аннотация. В статье предлагается трактовка галлюцинаций в языковых моделях не как простого следствия статистической природы ИИ, а как новой формы эпистемической ошибки, активно порождающей ложные убеждения под видом достоверного знания. Автор критикует метафору «стохастического попугая», которая сводит деятельность ИИ к пассивному воспроизведению данных без претензии на истину, и демонстрирует, что продукты ИИ функционально эквивалентны суждениям как источники знания. Признание галлюцинаций формой ошибки позволяет перейти от узкотехнических проблем к вопросам эпистемической ответственности, структурных дефектов в производстве знания и социальных последствий доверия машинному «суждению».

Ключевые слова: галлюцинации ИИ, эпистемическая ошибка, стохастический попугай, теория суждения, эпистемическая ответственность, философия искусственного интеллекта.

Для цитирования: Шевченко, А. А. (2025). «Галлюцинации» ИИ как новая форма эпистемической ошибки. *Respublica Literaria*. Т. 6. № 4. С. 93-98. DOI: 10.47850/RL.2025.6.4.93-98

AI “HALLUCINATIONS” AS A NEW FORM OF EPISTEMIC MISTAKE

A. A. Shevchenko
Institute of Philosophy and Law SB RAS (Novosibirsk)
shev@philosophy.nsc.ru

Abstract. This article proposes reinterpreting hallucinations in language models not merely as a byproduct of AI’s statistical nature, but as a novel form of epistemic error – an active process that generates false beliefs disguised as reliable knowledge. The author critiques the dominant “stochastic parrot” metaphor, which reduces AI output to passive data repetition devoid of truth claims, and demonstrates that AI outputs functionally equate to judgments as sources of knowledge. Recognizing hallucinations as errors shifts the focus from narrow technical fixes to questions of epistemic responsibility, structural flaws in knowledge production, and the social consequences of trusting machine-generated “judgments.”

Keywords: AI hallucinations, epistemic error, stochastic parrot, theory of judgment, epistemic responsibility, philosophy of artificial intelligence.

For citation: Shevchenko, A. A. (2025). Ai “Hallucinations” as a New Form of Epistemic Mistake. *Respublica Literaria*. Vol. 6. No. 4. Pp. 93-98. DOI: 10.47850/RL.2025.6.4.93-98

С технической точки зрения, галлюцинации в языковых моделях возникают как следствие их статистической природы: они не получают знание из мира, а генерируют тексты, основанные на вероятностных паттернах. Такой продукт ИИ часто имеет все формальные признаки знания: он структурирован как утверждение, содержит аргументы

и изложен в уверенном, категоричном стиле. При этом искусственный интеллект не отличает истину от правдоподобия, поскольку не обладает ни субъективным опытом, ни ответственностью за свои «высказывания». Поэтому большинство исследований по галлюцинациям фокусируются на технических решениях – механизмах коррекции выводов или ранжирования источников по степени достоверности. Однако подобный подход игнорирует фундаментальную проблему: как относиться к подобным текстам, если стандартные процедуры проверки на истинность, такие как сопоставление с действительностью, не всегда возможны из-за масштабов автоматической генерации или ограниченного доступа к фактам?

Наиболее простой путь – объявить производителей таких текстов «эпистемически индифферентными», наподобие «стохастического попугая», бездумно и хаотично повторяющего чьи-то слова, ничего по-настоящему не утверждая. Стандартным считается следующее определение: «Языковая модель – это система, которая произвольно соединяет последовательности языковых форм, встречавшихся ей в огромных объемах обучающих данных, опираясь на вероятностные закономерности их сочетания, но без какой-либо связи со смыслом: стохастический попугай» [Bender et al., 2021, p. 617]. Хотя подход, в рамках которого система безразлична (не стремится) к истине доминирует в современной эпистемологии [см., напр.: Herrmann, Levinstein, 2025; Wachter et al., 2024], такую позицию вряд ли можно считать адекватной. Предположение, что галлюцинации ИИ представляют собой лишь результат «неудачного предсказания» и не имеют какого-либо отношения к подлинному знанию, может, на первый взгляд, показаться удобным решением, избавляя нас от необходимости вообще погружаться в сферу эпистемологии. Однако такой подход игнорирует реальность, в которой эти системы уже интегрированы в общественную жизнь – от политики и медицины до образования. Люди верят их утверждениям, принимают их за истину и действуют соответственно.

Именно это дает основания рассматривать продукты ИИ как особую форму «знания» – если не по сути, то по функции. Хотя ИИ не агент в классическом смысле (без убеждений, стремления к истине или ответственности), он занимает позицию агента: производит утверждения, претендующие на истину. Независимо от онтологической природы, он вовлечен в эпистемические отношения, которые мы не можем игнорировать. Как показывает В. Барасси, модель участвует в «суждении» в функциональном смысле: она выполняет операцию приписывания предиката субъекту на основе обучения на корпусе данных. Это не сознательное действие, но акт, имеющий эпистемические последствия [Barassi, 2024]. Метафора «суждения» здесь не онтологическая, а нормативная: она позволяет применять к ИИ критерии, разработанные для оценки человеческого познания, без необходимости приписывать ему сознание.

Когда мы приписываем моделям безразличие к истине, то уходим от проблемы, которая заключается не в пассивном безразличии, а в том, что модель активно производит ложные убеждения, маскирующиеся под знание. В этой статье предлагается считать галлюцинацию не формой индифферентности, а новым видом эпистемической ошибки: это когнитивно активный акт, порождающий ложное убеждение под видом достоверного знания. Подобно иррациональному обоснованию или когнитивному диссонансу в человеческом познании, модель производит суждение, логически оформленное, но лишенное соответствия миру. Это позволяет перенести на искусственные системы классические критерии оценки познания – обоснованность, когерентность, надежность познавательного процесса.

Такая постановка вопроса позволяет обратиться к давней традиции размышления об ошибке – от кантовского различения мнения, веры и знания до глубоко проработанной эпистемологии индийской школы Ньяя. Уже в античности сторонники этой школы четко разводили простое отсутствие знания (например, не знать, что есть в комнате) и активное, но ложное убеждение (например, думать, что в комнате змея, когда там веревка). Важно, что здесь ошибка – не дефицит информации, а ее искажение: один объект квалифицируется как другой, с полной уверенностью в истинности суждения. Это активный, хотя и дефектный, когнитивный акт, а не простое молчание незнания. Условия такой ошибки включают: контакт с органом чувств, дефектное восприятие, неверную квалификацию и убежденность [Matilal, 1986, pp. 209-213]. Все четыре элемента присутствуют в случае ИИ: входные данные, их искаженная обработка, ложная квалификация и уверенная формулировка. Та же интуиция лежит в основе кантовского анализа ошибки в «Критике чистого разума». Кант указывает, что чувственность сама по себе не ошибается, поскольку не судит. Ошибка возникает в суждении, когда разум связывает представления в утверждение и выдает его за истину. «… чувства не ошибаются, однако не потому, что они всегда правильно судят, а потому, что они вообще не судят. Следовательно, и истина, и ошибка, а значит, и видимость, вводящая в заблуждение, имеют место только в суждении, т. е. только в отношении предмета к нашему рассудку» [Кант, 1964, с. 336].

Структурно аналогично это происходит и с генеративными моделями. Когда ИИ утверждает, что ученый X опубликовал работу Y в 2023 г. (хотя ни X, ни Y не существуют), он не «додумывает» из-за нехватки данных. Он конструирует суждение: приписывает объекту свойство, формируя ложное, но целостное представление. У него нет разума в кантовском смысле, нет субъективного опыта. Но логическая форма высказывания – субъект, предикат, притязание на истину – воспроизводит структуру суждения. Мы имеем дело не с «фантазией», а с ошибкой в «производстве» знания. Конечно, ИИ не способен судить, поскольку лишен разума. Однако Кант говорит не о сознании, а о логической форме суждения о том, как представления связываются в утверждение.

Современные философы продолжают эту линию. Так, В. Барасси в своей работе 2024 г. показывает, что ошибки ИИ – это не артефакты, связанные с помехами или недостатком данных, а следствие структурного дефекта в производстве знания: модели учатся на социальных данных, в которых уже заложены искажения, иерархии и умолчания. Галлюцинация – это не «фантазия», а проекция этих скрытых структур в форму утверждения. «Объединяя философские подходы к теории ошибки с антропологическими перспективами, я утверждаю, что теория ошибки необходима, поскольку она проливает свет на то, что сбои в наших системах проистекают из ошибочных процессов производства знания, неправильных характеристик и дефектных когнитивных связей» [Barassi, 2024]. Это заставляет пересмотреть и классическое определение знания как обоснованного истинного убеждения [Šekrst, 2024]. Галлюцинации ИИ нарушают условие как истинности, так и обоснованности: модель может быть уверена в ложном, не имея механизма критической оценки собственных выводов. При этом важно не смешивать галлюцинацию с ложью: ложь предполагает намерение ввести в заблуждение; галлюцинация – это искренняя ошибка, совершаемая без злого умысла, но с полной уверенностью. Это именно эпистемическая, а не этическая проблема.

Переход от понимания галлюцинаций как отсутствия к пониманию их как ошибки позволяет задать новые вопросы: какие критерии обоснования нарушаются в ИИ-суждении? Можем ли мы говорить о «неправильном соединении» данных, даже если у системы нет

интенциональности? И главный вопрос: если галлюцинация функционирует в общественном пространстве как знание, то какую ответственность за ее последствия мы готовы взять на себя как разработчики, пользователи и те, кто доверяет ее выводам?

Признание галлюцинации формой ошибки, а не индифферентности, смещает фокус с «нехватки данных» на устройство когнитивных операций. Это важно по трем причинам. Во-первых, открывается возможность говорить об «эпистемической ответственности» применительно к ИИ. Если система лишь «не знает», она за пределами нормативной оценки. Но если она «ошибается», выдавая ложное за истинное, то речь идет о дефекте в самой структуре ее «познавательного» процесса. Ответственность лежит не только на алгоритме, но и на разработчиках, регуляторах, пользователях, доверяющих выводам. Это вопрос не этики в узком смысле слова, а пересмотра наших эпистемических практик: мы обязаны выстраивать практики верификации, критики, контекстуализации, обеспечения максимально возможной прозрачности. Разработчики не могут прятаться за метафорой «просто инструмента», когда инструмент производит систематические заблуждения.

Во-вторых, ошибка предполагает наличие структуры, которая работает неверно. Это позволяет задавать философски значимые вопросы о том, как именно модель соотносит категории, строит причинно-следственные связи, какие принципы лежат в основе ее «суждений». Подобный подход сближает анализ ИИ с традиционной эпистемологией, где ошибка всегда была маркером для изучения устройства разума. Даже если этот разум искусственный и лишен рефлексии, его дефекты могут пролить свет на скрытые предположения, заложенные в данных и архитектуре модели.

В-третьих, возникают социальные последствия. Признавая, что ИИ способен не просто «молчать», а формировать убеждения, влиять на решения, искажать коллективную картину мира, мы признаем, что перед нами не техническая неполадка, а социальный риск. Это обязывает разрабатывать не просто более точные модели, но также социальные и технологические механизмы, способные противостоять распространению машинных заблуждений. Критиковать альтернативные трактовки, например, метафору «стохастического попугая», важно не ради полемики, а потому что они уводят от сути. ИИ не «понимает» в человеческом смысле. Но это не делает его безвредным. Его ложные утверждения функционируют как знание в социальном пространстве, и именно это делает их эпистемически значимыми.

В фундаментальной эпистемологии мы редко сталкиваемся с простым отсутствием знания, гораздо важнее уметь распознавать активные когнитивные ошибки, логически оформленные, но не соответствующие действительности. Обратимся к двум популярным философским концепциям – «теории моральной ошибки» Дж. Л. Маки, которая давно вышла за рамки морали, и концепции интенциональности Д. Деннета.

В основе моральной теории ошибки (1977 г.) лежит тезис Дж. Л. Маки о том, что ложные суждения – это не просто отсутствие знания, а активное производство утверждений с претензией на истину, когда самой онтологической опоры нет. Все моральные утверждения он считает ложными или ошибочными, поскольку объективных моральных фактов не существует, хотя моральный дискурс предполагает их существование [Mackie, 1977, pp. 30-38]. В применении к ИИ это значит: машина «галлюцинирует» не из-за незнания, а потому что создает формально правильное, но неверное суждение, действуя в логике эпистемической ошибки.

Д. Дэннет в своей книге «Интенциональная установка» [Dennett, 1987] утверждает, что мы объясняем сложные системы, приписывая им «убеждения» и «цели», когда это делает их поведение удобным для прогнозирования. Главное здесь не внутреннее устройство системы, а эффективность объяснительной стратегии: система, по его мнению, является агентом по отношению к наблюдателю тогда и только тогда, когда наилучшая модель этой системы, доступная наблюдателю, представляет ее как обладающую «целями» и «убеждениями». А по-настоящему обладать убеждением – это и значит быть такой системой, поведение которой надежным образом поддается предсказанию посредством интенциональной стратегии [Dennett, 1987, ch. 2].

Для искусственного интеллекта это означает: даже если у него нет сознания, мы практически и теоретически относимся к нему как к агенту, способному «заблуждаться», т. е. совершать ошибки, которые имеют социально и эпистемически значимые последствия. Когда ИИ генерирует утверждение, оно становится «действием» в системе социальных смыслов. Если модель регулярно совершает определенные ошибки, то для нас рационально ожидать (и объяснять) их как «ошибки агента», а не просто технический сбой. Таким образом, ошибки ИИ в интенциональной перспективе воспринимаются не только как программные сбои, но как ложные «убеждения», такие как фальшивое приписывание свойств или фактов. Такой подход делает наше отношение к ИИ более пристрастным: мы оцениваем, подвергаем критике и даже налагаем санкции за его ошибки в публичном пространстве так же, как было бы с человеческим агентом, именно потому, что интенциональный подход обеспечивает наилучшее объяснение (и управление) поведением системы.

В современной эпистемологии искусственного интеллекта ошибка должна рассматриваться не как отсутствие знания, а как структурированный акт, создающий ложное убеждение, функционирующее как знание в социальном пространстве. Роль ИИ определяется тем успехом, с которым мы можем прогнозировать его поведение через приписывание «суждения» и «цели», а не его внутренним состоянием. Метафора «стохастического попугая» оказывается поверхностной: за ней стоит сложная эпистемическая динамика, где ответственность за ошибку распределена между алгоритмом, данными и обществом, которое назначает ИИ на место познавательного агента.

Список литературы / References

Кант, И. (1964). *Сочинения: в 6 т. Т. 3: Критика чистого разума*. М.: Изд-во «Мысль».
Kant, I. (1964). *Collected Works. In 6 vols. Vol. 3. Critique of Pure Reason*. Moscow.

Barassi, V. (2024). Toward a Theory of AI Errors: Making Sense of Hallucinations, Catastrophic Failures, and the Fallacy of Generative AI. [Online]. *Harvard Data Science Review, (Special Issue 5)*. Available at: https://www.researchgate.net/publication/383732733_Toward_a_Theory_of_AI_Errors_Making_Sense_of_Hallucinations_Catastrophic_Failures_and_the_Fallacy_of_Generative_AI (Accessed: 10.10.2025).

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), 3-10 March 2021. Virtual Event. Canada. ACM, New York, NY, USA. Pp. 610-623. DOI: 10.1145/3442188.3445922.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge. MA: MIT Press.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. London. Penguin.

Matilal, B. K. (1986). *Perception. An Essay on Classical Indian Theories of Knowledge*. Oxford. Clarendon Press.

Šekrst, K. (2024). Chinese Chat Room: AI Hallucinations, Epistemology and Cognition. *Studies in Logic, Grammar and Rhetoric*. University of Białystok. Vol. 69. No. 1. Pp. 365-381. DOI: 10.2478/slgr-2024-0029.

Herrmann, D. A., Levinstein, B. A. (2025). Standards for Belief Representations in LLMs. *Philosophy and Machine Learning*. Vol. 35. No. 5. DOI: 10.48550/arXiv.2405.21030.

Wachter, S., Mittelstadt, B., Russell, C. (2024). Epistemic Challenges of Large Language Models. *AI & Society*. Vol. 39. No. 1. Pp. 55-72.

Сведения об авторе / Information about the author

Шевченко Александр Анатольевич – доктор философских наук, ведущий научный сотрудник Института философии и права Сибирского отделения Российской академии наук, г. Новосибирск, ул. Николаева, 8, e-mail: shev@philosophy.nsc.ru, https://orcid.org/0000-0002-8563-5464.

Статья поступила в редакцию: 15.10.2025
После доработки: 10.11.2025
Принята к публикации: 17.11.2025

Shevchenko Aleksandr – Doctor of Philosophical Sciences, Leading Researcher of the Institute of Philosophy and Law of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Nikolaeva Str., 8, e-mail: shev@philosophy.nsc.ru, https://orcid.org/0000-0002-8563-5464.

The paper was submitted: 15.10.2025
Received after reworking: 10.11.2025
Accepted for publication: 17.11.2025