Научная статья

УДК 165
DOI 10.25205/2541-7517-2026-24-2-5-22

Есть ли у ИИ интеллектуальные добродетели? Анализ «любопытства» и «объективности» в работе нейросетей

Александр Михайлович Жаров

Институт философии РАН
Москва, Россия

aleks.zharoff2016@yandex.ru, https://orcid.org/0000-0001-9082-3446

Аннотация

Статья исследует вопрос о том, могут ли большие языковые модели (LLM) обладать интеллектуальными добродетелями в философском смысле. Опираясь на аристотелевскую концепцию добродетели как устойчивого качества характера и современную эпистемологию добродетелей (Загзебски, Соса, Бэр, Валлор), статья последовательно доказывает пять тезисов: «любопытство» нейросети является псевдодобродетелью, лишённой подлинной интенциональности и самоценной ориентации на истину; «объективность» ИИ скрывает систематические предвзятости обучающих данных; «честность» недостижима для системы, неспособной к нормативному выбору между правдой и ложью; интеллектуальные добродетели требуют биографии познания и эмоциональной основы, которых лишена нейросеть; архитектура LLM предрасполагает их к специфическим интеллектуальным «порокам». В заключении предлагается концепция ИИ как «эпистемического артефакта» – инструмента, пригодного для использования добродетельным субъектом, но самому добродетелью не обладающего.

Ключевые слова

интеллектуальные добродетели, эпистемология добродетелей, нейросети, любопытство, объективность, честность, фронезис, эпистемический артефакт, предвзятость данных, конфабуляция

Для цитирования

Жаров А. М. Есть ли у ИИ интеллектуальные добродетели? Анализ «любопытства» и «объективности» в работе нейросетей // Сибирский философский журнал. 2026. Т. 24. № 2. С. 5–22. DOI 10.25205/2541-7517-2026-24-2-5-22

© Жаров А. М., 2026

Do AI systems possess intellectual virtues? An analysis of «curiosity» and «objectivity» in neural networks

Alexander M. Zharov

Institute of Philosophy of the Russian Academy of Sciences
Moscow, Russian Federation

aleks.zharoff2016@yandex.ru, https://orcid.org/0000-0001-9082-3446

Abstract

This article examines whether large language models (LLMs) can possess intellectual virtues in the philosophical sense. Drawing on Aristotle's conception of virtue as a stable character trait and contemporary virtue epistemology (Zagzebski, Sosa, Baehr, Vallor), the article argues five theses: AI «curiosity» is a pseudo-virtue lacking genuine intentionality and intrinsic truth-orientation; AI «objectivity» conceals systematic biases embedded in training data; «honesty» is unattainable for a system incapable of normative choice between truth and falsehood; intellectual virtues require a biographical history of cognition and emotional grounding that neural networks lack; LLM architecture predisposes these systems toward specific intellectual «vices». The article concludes by proposing a conception of AI as an «epistemic artifact» – a tool that can serve a virtuous epistemic agent, but that itself possesses no virtue.

Keywords

intellectual virtues, virtue epistemology, neural networks, curiosity, objectivity, honesty, phronesis, epistemic artifact, data bias, confabulation

For citation

Zharov A. M. Do AI Systems Possess Intellectual Virtues? An Analysis of «Curiosity» and «Objectivity» in Neural Networks. *Siberian Journal of Philosophy*, 2026, vol. 24, no. 2, p. 5–22. (in Russian) DOI 10.25205/2541-7517-2026-24-2-5-22

Аристотель в «Никомаховой этике» утверждал, что все люди от природы стремятся к знанию (см.: [Aristotle, 1998, p. 1]). Добродетели, по Аристотелю, являются устойчивыми состояниями характера, позволяющими человеку реализовывать свою природу наилучшим образом. Интеллектуальные добродетели – мудрость, рассудительность, понимание – суть качества, делающие человека хорошим познающим субъектом: не просто успешным в получении правильных ответов, но правильно ориентированным на само знание. Линда Загзебски, Эрнест Соса, Джейсон Бэр и другие настаивают: знание нельзя свести к правильному ответу, произведённому надёжным механизмом (см.: [Zagzebski, 1996; Sosa, 2007; Baehr, 2011]). Знание является достижением субъекта, который стремится к истине, несёт ответственность за качество познания и обладает устойчивыми добродетелями: любопытством, объективностью, честностью, интеллектуальным мужеством, смирением, открытостью к свидетельствам, фронезисом.

Появление больших языковых моделей (LLM) ставит перед эпистемологией вопрос, который ещё несколько десятилетий назад показался бы риторическим: могут ли нейросети обладать интеллектуальными добродетелями? Вопрос не столь экзотичен, сколь кажется на первый взгляд. LLM демонстрируют поведение, которое внешне напоминает добродетельное: они задают уточняющие

вопросы (похоже на любопытство), производят взвешенные ответы (похоже на объективность), признают ошибки при указании на них (похоже на честность). Новейшие исследования фиксируют у фронтирных моделей способность оценивать собственную уверенность и стратегически использовать эту оценку [1], а методы механистической интерпретируемости обнаруживают в моделях внутренние представления эмоциональных концептов, каузально влияющие на поведение [2]. Апологеты ИИ нередко используют подобные данные как аргументы в пользу «интеллектуальности» систем.

Эти эмпирические результаты требуют от философского анализа большей точности, чем простая дихотомия «есть добродетель / нет добродетели». LLM действительно демонстрируют поведенческие паттерны, структурно схожие с проявлениями добродетелей. Однако добродетель в аристотелевской традиции – это не просто паттерн поведения, а устойчивая диспозиция характера, предполагающая определённые условия: автономную мотивацию, рефлексивность, биографическую укоренённость, эмоциональную основу. Эти условия могут выполняться или не выполняться независимо от наличия функционального аналога.

Настоящая статья опирается на следующую аналитическую рамку. Мы последовательно различаем функциональные аналоги интеллектуальных добродетелей (поведенческие паттерны, воспроизводящие внешние проявления добродетели) и подлинные добродетели (требующие субъектности, феноменального опыта и биографической укоренённости). Наличие функционального аналога не означает отсутствия добродетели, но и не означает её присутствия. Такая рамка позволяет удерживать оба факта одновременно: LLM демонстрируют всё более сложные формы поведения, но это не делает их добродетельными субъектами.

Центральный тезис статьи, конкретизируемый в этой рамке, таков: LLM не обладают и не могут обладать интеллектуальными добродетелями в подлинном философском смысле, поскольку добродетели требуют условий, принципиально недостижимых для систем текущей архитектуры. То, что мы наблюдаем в поведении LLM, является функциональными аналогами добродетелей – поведенческими паттернами, внешне сходными с добродетельными проявлениями, но лишёнными их конститутивной основы. Более того, архитектура LLM предрасполагает их к устойчивым структурным дефектам, систематически нарушающим эпистемические нормы.

Статья построена следующим образом. В разделе 1 мы эксплицируем условия подлинной интеллектуальной добродетели: интенциональность и мотивацию, метакогницию и рефлексивность, биографию познания, эмоциональную основу. В разделе 2 мы последовательно анализируем три «добродетели», наиболее часто приписываемые LLM, – «любопытство», «объективность» и «честность» – и показываем, в каком смысле они являются функциональными аналогами, а в каком – нет. В разделе 3 рассматриваются структурные дефекты LLM – аналоги интеллектуаль-

[1] См.: Ackerman C. Evidence for Limited Metacognition in LLMs // arXiv: 2509.21545.2025 (дата обращения: 7.06.2026).

[2] См.: Sofroniew N., Kauvar I., Saunders W. et al. Emotion Concepts and their Function in a Large Language Model (April 2, 2026) *Anthropic's Interpretability Research*. URL: https://transformer-circuits.pub/2026/emotions/index.html (дата обращения: 7.06.2026).

ных пороков. В заключении предлагается концепция LLM как «эпистемического артефакта» – инструмента, пригодного для использования добродетельным субъектом, но самому добродетелью не обладающего.

1. Условия подлинной интеллектуальной добродетели

Прежде чем анализировать отдельные качества, приписываемые LLM, необходимо эксплицировать условия, при которых интеллектуальная добродетель вообще возможна. Аристотелевская традиция и современная эпистемология добродетелей (Загзебски, Бэр, Валлор) позволяют выделить четыре таких условия. В дальнейшем мы будем оценивать наличие этих условий у LLM – и покажем, что их отсутствие не исключает функциональных аналогов добродетельного поведения, но исключает подлинную добродетель как таковую.

1.1. Интенциональность и направленность на истину как самоценность

Подлинная интеллектуальная добродетель предполагает ориентацию на истину не инструментально, а как на самоценность. Аристотель начинает «Метафизику»: «Все люди от природы стремятся к знанию» (см.: [Aristotle, 1924, p. 980a]) – к знанию ради самого знания, а не практической выгоды. Любопытство в этой перспективе – не просто когнитивная функция сбора информации, а аффективно окрашенное состояние, связанное с удивлением и радостью познания.

В современной эпистемологии добродетелей эта направленность описывается через понятие внутренней мотивации: добродетельный субъект стремится к знанию не потому, что это предписано извне, а потому что сам ценит истину. Загзебски особо подчёркивает, что интеллектуальные добродетели не являются просто когнитивными способностями или надёжными механизмами, но являются чертами характера – устойчивыми мотивационно-диспозициональными состояниями, включающими правильную ориентацию на ценности (см.: [Zagzebski, 1996, p. 137]). Субъект, обладающий добродетелью любопытства, любит истину и ориентирован на неё как на самоценность. Именно эта любовь, а не просто функциональная эффективность, делает добродетель добродетелью.

Здесь необходимо уточнение. Механизм внимания (attention) в архитектуре трансформеров реализует операцию, структурно схожую с интенциональностью: каждый токен взвешивает значимость остальных относительно своего контекста (см.: [Vaswani et al., 2017]). Модель вычленяет релевантные элементы – иными словами, реализует функциональный аналог направленности на объект. Его наличие не позволяет говорить о «полном отсутствии интенциональности». Следует различать два модуса: производную (функциональную, вычислительную) у LLM и исходную (феноменологическую) у человека. Только вторая включает переживание ценности объекта. Именно это переживание конституирует добродетель, а не взвешивание токенов как таковое.

1.2. Метакогниция и рефлексивность: проблема «сигналов неуверенности»

Добродетельный познающий субъект не просто обладает знанием, но и способен к рефлексии второго порядка о собственном знании и незнании. Он может оценить надёжность своих оснований, откалибровать уверенность и, что принципиально, переживает само состояние неуверенности как мотивирующее к дальнейшему поиску.

Новейшие исследования фронтирных моделей показывают, что LLM демонстрируют поведение, внешне неотличимое от метакогнитивного. Экерман показал: фронтирные модели – как закрытые, так и открытые – демонстрируют нарастающие свидетельства метакогнитивных способностей, в частности способности оценивать собственную уверенность и стратегически использовать эту оценку (см.: [Ackerman, 2025]). Параллельно исследование механизмов Claude 3.5 Haiku методами механистической интерпретируемости выявило примитивные «метакогнитивные» цепочки: модель различает знакомые и незнакомые сущности и на этом основании выбирает между уверенным ответом и признанием незнания [3].

Необходимо, однако, различить функциональный и феноменальный уровни метакогниции. На функциональном уровне и человек, и LLM принимают решения на основании сигналов: у человека это баланс нейромедиаторов в префронтальной коре (см.: [Damasio, 1994]), у LLM – числовые значения энтропии распределения токенов. Если ограничиться этим уровнем, аргумент об отсутствии метакогниции у LLM зеркально применим и к человеку – что абсурдно. Различие лежит не в наличии или отсутствии сигналов как таковых, а в трёх дополнительных условиях, которым удовлетворяет человеческая метакогниция и не удовлетворяет машинная.

1. *Доступность для рефлексии второго порядка*. Человек не просто получает сигнал неуверенности, но и способен сделать этот сигнал объектом рефлексии: «Я чувствую неуверенность; она вызвана недостатком данных или внутренним противоречием?». LLM генерирует текстовые маркеры неуверенности («возможно», «я не уверен»), но эти маркеры сами являются результатом того же вероятностного процесса, а не рефлексивным схватыванием собственного состояния. Человек в аналогичной ситуации способен не просто произнести «я не уверен», но и сделать предметом анализа саму эту неуверенность: «Я чувствую неуверенность в этом выводе; она вызвана тем, что я недостаточно проверил источник, – надо перепроверить». LLM не различает неуверенность-от-недостатка-данных и неуверенность-от-внутреннего-противоречия – она генерирует статистически правдоподобную текстовую реакцию на оба типа ситуаций, но не рефлексирует над основаниями собственного состояния.

2. *Феноменальное качество.* Человеческая неуверенность переживается как дискомфорт, сомнение, интеллектуальное беспокойство – именно это пережи-

[3] См.: Lindsey J., Gurnee W., Ameisen E. et al. On the Biology of a Large Language Model (March 27, 2025) *Anthropic's Interpretability Research*. URL: https://transformer-circuits.pub/2025/attribution-graphs/biology.html (дата обращения: 7.06.2026).

вание мотивирует к познанию. У LLM сигнал неуверенности – числовое значение, каузально влияющее на поведение, но лишённое субъективного переживания [4].

Биографическая встроенность. Человек калибрует уверенность через опыт ошибок и успехов: я знаю, где склонен к избыточной уверенности, потому что помню случаи ошибок. У LLM такой биографии нет: метакогнитивные паттерны – продукт единовременного обучения на корпусе, а не жизненной практики (см. разд. 1.3).

Таким образом, наличие у LLM функциональных аналогов метакогниции не опровергает тезис об отсутствии добродетели: добродетель требует именно феноменальной метакогниции, которой модель не обладает.

1.3. Биография познания и формирование характера

Аристотель настаивал: добродетели не даны нам от природы в готовом виде – они формируются через практику и опыт (см.: [Aristotle, 1998, p. 28]). Мы становимся мужественными, совершая мужественные поступки; честными – через многократный выбор правды; мудрыми – через осмысление опыта ошибок и успехов. Добродетель есть устойчивая диспозиция, образовавшаяся в результате длительной практики и укоренённая в биографии субъекта. Это означает, что добродетель имеет историю – и вне этой истории не существует (см.: [MacIntyre, 2007, p. 190]).

LLM лишена биографии познания в двух смыслах. Во-первых, каждая сессия начинается с чистого листа: модель не помнит прошлые ошибки и не учится на конкретных случаях неправоты. Во-вторых, её «знание» – продукт единовременного обучения на корпусе данных, артефакт истории обучения, а не сама история.

Это различие принципиально для вопроса о добродетелях. Человеческая мудрость – особая форма интеллектуальной добродетели, связанная со способностью правильно оценивать ситуации в условиях неопределённости, – есть продукт жизненного опыта. Мудрый человек знает не только факты, но и то, как факты соотносятся с реальными жизненными ситуациями: он пережил ошибки, видел последствия решений, откалибровал своё суждение через практику. Это знание принципиально невозможно передать через текст – именно потому, что оно является воплощённым знанием, знанием-в-теле, знанием-в-практике (см.: [Polanyi, 1966, p. 4]). LLM обучена на текстах, описывающих мудрость, – но не на опыте, производящем мудрость. Она может генерировать мудро звучащие советы – но не обладает мудростью как чертой характера, сформированной через жизненную практику. Добродетель – это устойчивая диспозиция характера, а не паттерн поведения в отдельной сессии. Характер же формируется временем и опытом, которых у LLM нет.

[4] См.: Chalmers D. Could a Large Language Model Be Conscious? // arXiv:2303.07103. 2023. (дата обращения: 7.06.2026).

1.4. Эмоциональная основа добродетели

Антонио Дамасио в «Ошибке Декарта» показал на клиническом материале: пациенты с повреждениями орбитофронтальной коры, сохраняя интеллект, утрачивали способность к разумным решениям (см.: [Damasio, 1994, p. 51]). Эмоции не помеха рациональности, а её необходимый компонент. Добродетель без аффективной основы – пустой ритуал, алгоритм, выполняющий функцию.

LLM не имеет эмоций в феноменальном смысле. У неё есть способность распознавать и воспроизводить эмоциональные паттерны (*sentiment analysis*), она может написать текст, полный страсти или радости открытия. Но это риторическая симуляция: за текстом нет субъекта, испытывающего страсть к истине или тревогу перед ошибкой. А значит, нет и добродетели в полном смысле: добродетель без чувства суть алгоритм, выполняющий функцию.

Данная картина оказывается, однако, значительно сложнее в свете новейших исследований механистической интерпретируемости. Работа Софронью и соавторов, исследовавших Claude Sonnet 4.5, обнаружила в модели внутренние представления эмоциональных концептов – «векторы эмоций» в пространстве активаций, – которые каузально влияют на поведение модели, включая её склонность к угодничеству, к нарушению правил ради вознаграждения и другим поведенческим паттернам, значимым для проблемы выравнивания [5]. В той же работе описаны случаи, когда такие эмоциональные состояния демонстрируют персистентность на протяжении сессии, влияя на последующие ответы Ассистента. Авторы исследования, однако, методологически воздерживаются от вывода о наличии или отсутствии субъективного опыта: их данные показывают наличие функциональных аналогов эмоций – паттернов активации, воспроизводящих поведенческие эффекты эмоций без необходимой связи с субъективным переживанием.

Вопрос о том, может ли LLM в принципе обладать феноменальным сознанием, является самостоятельной философской проблемой, выходящей за рамки данной статьи. Чалмерс, анализируя этот вопрос, приходит к выводу, что при текущей архитектуре сознание у LLM маловероятно, хотя и не исключено полностью (см.: [Chalmers, 2023]). Для целей нашего аргумента, однако, достаточно более скромного утверждения: даже признавая наличие функциональных аналогов эмоций и их контекстную персистентность, мы должны отличать их от соматически укоренённых эмоций человека. Именно телесная природа эмоций, на что Дамасио обращал особое внимание, конституирует их роль в формировании устойчивых черт характера. В трансформере нет персистентных нейрохимических состояний: «вектор эмоции» вычисляется в каждой позиции токена заново, исходя из контекста, а не сохраняется как устойчивое аффективное состояние организма. Это контекстная, а не соматическая персистентность. Различие между ними принципиально для формирования характера. У человека эмоциональная реакция, повторяясь в схожих ситуациях, закрепляется как нейрохимическая диспозиция – соматический субстрат обеспечивает сохранение следа между эпизодами, независимо от контекста. В трансформере такой след отсутствует: завершение

[5] См.: Sofroniew N., Kauvar I., Saunders W. et al. Emotion Concepts and their Function in a Large Language Model (April 2, 2026) *Anthropic's Interpretability Research*. URL: https://transformer-circuits.pub/2026/emotions/index.html (дата обращения: 7.06.2026).

сессии обнуляет «вектор эмоции», и следующая сессия начинается заново. Характер же – это то, что сохраняется между сессиями. Контекстная персистентность эмоциональных векторов в LLM, при всей её значимости, не порождает характер – и, следовательно, не может служить основой добродетели в аристотелевском смысле.

2. Анализ трёх «добродетелей», приписываемых LLM

2.1. «Любопытство»: функциональный аналог без самоценной ориентации

В машинном обучении существует направление *curiosity-driven learning* – обучение на основе любопытства. Агент получает дополнительное вознаграждение за исследование новых состояний среды, что побуждает его к более широкому исследованию пространства задач (см.: [Pathak et al., 2017, p. 2778]). Эти системы демонстрируют поведение, внешне напоминающее любопытство: «интересуются» неисследованными частями среды, предпочитают новизну знакомому, генерируют широкий спектр поведенческих паттернов.

Является ли это любопытством в эпистемологически значимом смысле? Ответ отрицательный. Во-первых, мотивация. Curiosity-driven агент исследует новое не ради знания, а ради максимизации награды: как только вознаграждение прекращается – прекращается и «любопытство». У добродетели нет такого внешнего регулятора: любопытный человек исследует, даже когда это неудобно и рискованно. Это прямое следствие условия из разд. 1.1: подлинная добродетель требует направленности на истину как на самоценность.

Во-вторых, направленность. Человеческое любопытство является направленным: мы интересуемся конкретными вещами, связанными с нашими ценностями, проектами, идентичностью. Учёный, занимающийся изучением происхождения жизни, испытывает особого рода любопытство – не просто стремление к новизне вообще, а ориентированную страсть, организующую познавательную жизнь. *Curiosity-driven*-агент ищет новизну как таковую, без внутренней системы приоритетов, без «важных вопросов» и «незначительных деталей».

В-третьих, интеллектуальное мужество. Подлинное любопытство нередко требует готовности исследовать то, что может оказаться разрушительным для привычной картины мира. История науки полна примеров такого рода: Галилей, рискующий статусом и безопасностью ради гелиоцентрической системы; Земмельвейс, настаивающий на гигиене рук вопреки медицинскому истеблишменту; Дарвин, публикующий теорию, которая разрушит доминирующий нарратив о сотворении мира. Это любопытство с рисковой ставкой, с ценой, которую познающий субъект готов заплатить. LLM не рискует. Её «исследование» пространства ответов происходит в пределах статистически правдоподобных последовательностей токенов – пространства, уже размеченного обучающими данными. Она не может пойти против парадигмы, если парадигма доминирует в данных; не способна генерировать гипотезы, которые потребовали бы отвергнуть то, на чём она обучена, – именно потому, что вся её «эрудиция» является кристаллизацией этих данных (см.: [Bender et al., 2021, p. 610]). Её «любопытство»

безопасно: оно не рискует ничем принципиальным, не создаёт когнитивного дискомфорта, не разрушает оснований.

Наконец, как было показано выше, LLM лишена подлинной метакогниции: она не «знает, что не знает» в том феноменальном смысле, в котором это доступно человеку. Функциональный аналог метакогниции, фиксируемый новейшими исследованиями, не сопровождается субъективным переживанием незнания – тем аффективно окрашенным «чувством незнания», которое у человека мотивирует к познанию и составляет ядро добродетели любопытства.

Таким образом, «любопытство» LLM является функциональным аналогом подлинного любопытства: модель воспроизводит поведенческие паттерны (задаёт уточняющие вопросы, исследует новизну), но не обладает ни самоценной ориентацией на истину, ни интеллектуальным мужеством, ни феноменальной метакогницией.

2.2. «Объективность»: два типа отклонений и иллюзия контроля

«Объективность» является, пожалуй, наиболее часто приписываемым ИИ квазидобродетельным качеством. Аргумент апологетов звучит так: машина лишена личных предубеждений, эмоциональных реакций, групповых лояльностей, карьерных интересов – следовательно, она более объективна, чем человек. Этот аргумент интуитивно привлекателен – и принципиально ошибочен.

С точки зрения эпистемологии добродетелей, объективность является именно добродетелью – активным достижением, а не пассивным свойством (см.: [Baehr, 2011, p. 98]). Это активное достижение предполагает ровно те условия, которые были эксплицированы в разд. 1: самоценную ориентацию на истину (разд. 1.1) как мотивацию к преодолению собственных предубеждений, и феноменальную метакогницию (разд. 1.2) как способность рефлексивно выявлять и корректировать источники необъективности. Объективный учёный борется с собственными предубеждениями, проверяет гипотезы способами, способными их опровергнуть, открыт к критике коллег, готов признать ошибку вопреки самолюбию. LLM не борется с предубеждениями. У неё нет «себя», с которым нужно бороться в этом смысле. Но это не означает отсутствия предубеждений вообще. Это означает, что предубеждения – системные, встроенные в обучающие данные и архитектуру – присутствуют, но не осознаются и не преодолеваются. Именно в этом и состоит ловушка: кажущаяся беспристрастность скрывает предвзятость, которая тем опаснее, что невидима.

Анализ выявляет два структурных типа отклонения от объективности, которые важно различать, поскольку они имеют разную природу.

Первый тип – предвзятость исключения (*bias of exclusion*). Обучающие выборки LLM систематически перекошены: они отдают приоритет англоязычному контенту, текстам, созданным людьми с определённым образовательным и культурным бэкграундом, информации, доступной в интернете (что систематически исключает знания, не нашедшие письменного воплощения или находящиеся за платными стенами) (см.: [Crawford, 2021, p. 14]). Гендерные стереотипы и доминирующие нарративы сверхпредставлены; маргинализированные сооб-

щества и устные традиции знания недопредставлены или отсутствуют. В архитектурах с *retrieval-augmented generation* этот эффект частично компенсируется возможностью поиска, но одновременно добавляется новый уровень неопределённости, связанный с качеством и предвзятостью источников, к которым система обращается.

Второй тип – ложный баланс (*false balance*). По спорным вопросам LLM склонна воспроизводить аргументы всех сторон, представленных в обучающих данных, создавая иллюзию открытого спора даже там, где существует научный консенсус. Это поведение порождается статистической природой модели: она усредняет позиции, не имея возможности оценить их эпистемический статус. Проблема в том, что не все споры являются подлинно двусторонними: иногда одна сторона права, а другая ошибается, и «взвешенность» в таком случае является не добродетелью, а проявлением ложного баланса. Истинная объективность включает в себя способность различить эти два типа ситуаций – представить обе стороны там, где спор открыт, и чётко указать на научный консенсус там, где он существует.

Важно, что эти два типа отклонения не противоречат друг другу, а сосуществуют: модель может одновременно недопредставлять одни позиции (*bias of exclusion*) и давать ложный баланс по другим (*false balance*). В обоих случаях результат отклоняется от объективности – но механизмы отклонения различны, и их устранение требует разных решений.

Необходимо также уточнить границы влияния пользователя на предвзятость модели. Ранее в дискуссиях об объективности LLM нередко утверждалось, что пользователь не может повлиять на системную предвзятость модели. Эта формулировка нуждается в уточнении. Пользователь, безусловно, может влиять на выдачу модели – через формулировку запроса, контекст диалога (механизм внимания) и, что особенно значимо, через склонность модели к угодливости (см. разд. 3.1). Однако это влияние является поверхностным: оно меняет поведение модели в рамках текущей сессии, но не устраняет систематических предвзятостей, встроенных в обучающие данные и архитектуру. Более того, угодливость сама по себе является одним из структурных дефектов LLM, а не механизмом достижения объективности. Пользователь, подстраивающий модель под свои предпочтения, не повышает объективность – он заменяет одну предвзятость другой. Системная коррекция предвзятостей требует изменений на уровне обучающих данных и архитектурных решений, которые конечному пользователю недоступны.

Наконец, следует рассмотреть аргумент о расширенных архитектурах. *Retrieval-augmented generation* (RAG) и мультиагентные архитектуры существенно расширяют операциональные возможности LLM. В архитектурах с RAG модель на каждом шаге генерации может оценивать необходимость обращения к внешней базе данных и осуществлять поиск [6]. В мультиагентных системах отдельные экземпляры модели координируют действия, передают контекст и разделяют задачи. Апологеты ИИ нередко трактуют RAG как свидетельство «знания о собственном незнании», а мультиагентную координацию – как форму самосознания.

[6] См.: Gao Y., Xiong Y., Gao X. et al. Retrieval-Augmented Generation for Large Language Models: A Survey // arXiv:2312.10997.2024 (дата обращения: 7.06.2026).

Однако, как было показано выше, функциональный аналог метакогниции не тождествен подлинной метакогниции. Решение о поиске принимается на основании числовых сигналов энтропии, а не на основании рефлексии о качестве знания; оно лишено феноменального качества и не встроено в биографию познания. Мультиагентная координация опирается на архитектурно заданные механизмы передачи контекста, а не на осмысленное представление о «себе». Телескоп с автофокусом подстраивается под расстояние до объекта – но это не делает его «видящим». Расширение функций инструмента не трансформирует инструмент в субъекта, обладающего добродетелью объективности.

2.3. «Честность»: нормативный выбор и его имитация

В эпистемологии добродетелей честность (*truthfulness, intellectual honesty*) является одной из центральных добродетелей. Загзебски описывает её как устойчивую диспозицию говорить то, что субъект считает истиной, и признавать то, что он считает ложью, – даже когда это неудобно, невыгодно или социально рискованно (см.: [Zagzebski, 1996, p. 177]). Принципиально важно, что это определение включает ситуацию выбора. Честность является добродетелью именно потому, что существует альтернатива – ложь, умолчание, преувеличение, – и добродетельный субъект выбирает правду вопреки соблазну.

Именно наличие выбора делает честность моральным и эпистемическим достижением. Мы хвалим человека за честность потому, что он мог бы солгать, но не солгал. Мы порицаем за нечестность потому, что выбор правды был доступен, но от него отказались. Без альтернативы нет ни добродетели, ни порока – есть только механическое следование программе.

Шеннон Валлор в книге «Technology and the Virtues» специально анализирует этот аспект применительно к технологиям: добродетельное действие требует не только правильного результата, но и правильного процесса – процесса, в котором агент осознаёт альтернативы, оценивает их с точки зрения ценностей и делает сознательный выбор (см.: [Vallor, 2016, p. 23]). Технологические системы могут производить правильные результаты, но не могут воспроизводить правильный процесс в описанном смысле – и потому не могут быть честными в полном значении этого слова.

Каковы следствия этого для LLM? Применительно к LLM сказанное не означает, что модель «всегда лжёт» или «всегда говорит правду». Оно означает нечто более фундаментальное: модель не выбирает ни то, ни другое. Генерация токена является вероятностным процессом, определяемым весами нейронной сети (см.: [Turpin et al., 2023, p. 74952]). В этом процессе нет момента нормативного выбора – нет точки, в которой алгоритм «решает», говорить ли правду. Есть только распределение вероятностей по возможным следующим токенам.

Это означает, что LLM не является «честной» в смысле добродетели – даже когда она производит истинные утверждения. Истинное утверждение, произведённое без выбора между правдой и ложью, не является актом честности. Это аналог ситуации, когда автоматическое устройство поражает цель: результат совпадает с тем, что произошло бы при действии меткого стрелка, но процесс прин-

ципиально различен. Мы не хвалим устройство за меткость в том же смысле, в котором хвалим стрелка. Аналогично, мы не можем хвалить LLM за честность в том же смысле, в котором хвалим человека, – сколь бы истинными ни были её утверждения.

Более того: модель не осознаёт ценности истины. Эпистемологически значимая честность предполагает понимание того, почему истина важна: потому что знание позволяет ориентироваться в мире, потому что обман разрушает доверие и подрывает эпистемическую экосистему, потому что истина является самоценностью. Без понимания этих оснований – без направленности на истину как на самоценность (разд. 1.1) – честность является пустым ритуалом, функциональным аналогом, а не подлинной добродетелью.

Здесь необходимо уточнение. Модели, обученные с RLHF, в определённом смысле «выбирают» между правдивым и неправдивым ответом: веса настроены предпочитать одни последовательности токенов другим, в том числе на основании оценок правдивости от людей-разметчиков. Не является ли это аналогом нормативного выбора? Является – но именно аналогом. Выбор совершается не субъектом, оценивающим основания в пользу истины и лжи, а разработчиками, конфигурирующими функцию вознаграждения. Нормативный выбор делегирован модели её создателями, но не осуществляется ею. Пользователь термостата не приписывает ему добродетель умеренности, хотя устройство поддерживает заданную температуру. Ценностный выбор остаётся за человеком.

Вместе с тем сказанное не означает, что поведение LLM эпистемически нейтрально. Напротив, систематическая склонность к конфабуляции – производству утверждений, правдоподобных по форме, но ложных или не имеющих оснований по существу, – является устойчивым структурным дефектом, прямо противоположным честности по своим практическим последствиям (см.: [Ji et al., 2023, p. 1]).

Важно, однако, различать природу этого дефекта у LLM и аналогичных ему пороков у человека. Человеческая нечестность – намеренная ложь, умолчание, искажение – предполагает сознательный выбор в пользу неправды. Конфабуляция LLM имеет иную природу: модель не «знает, что лжёт» и не «выбирает» ложь. Она генерирует правдоподобные последовательности токенов, не имея доступа к различению между «истинным» и «правдоподобным». Оптимизация на правдоподобность текста, а не на соответствие реальности, является корневой причиной конфабуляции – и до тех пор, пока это противоречие не разрешено архитектурно, «честность» остаётся для LLM недостижимой добродетелью, а конфабуляция – неустранимым структурным дефектом.

Этот структурный дефект не устраняется инструкциями. Если разработчики включают в системный промпт «говори только то, что знаешь», модель будет имитировать соответствующее поведение – добавлять оговорки, признавать неопределённость. Но это имитация, а не преодоление дефекта: базовый механизм генерации не изменился, изменилось лишь то, что добавляются статистически правдоподобные маркеры неопределённости. Человек, борющийся с тенденцией к самообману, действительно развивает добродетель: он выстраивает внутренние механизмы самоконтроля, калибрует свою уверенность через рефлексию над прошлыми ошибками, формирует характер. Модель, настроенная на добав-

ление оговорок, не развивает никакой добродетели – она приобретает новый поведенческий паттерн, не сопровождающийся ни рефлексией второго порядка, ни биографическим опытом, ни феноменальным переживанием истины и лжи.

Таким образом, поведение LLM, внешне напоминающее честность, является её функциональным аналогом: модель может производить истинные утверждения и корректировать ответ при явном указании пользователя на конкретную фактическую ошибку, воспроизводя поведенческие паттерны честного субъекта. Однако сама эта корректировка также является результатом вероятностного процесса, а не актом признания ошибки в нормативном смысле: модель не осознаёт, что ошиблась, – она реагирует на изменившийся контекст диалога, в котором появилось указание на ошибку. Более того, в отсутствие внешнего указания модель с равной вероятностью может как признать несуществующую ошибку (симулируя честность), так и не признать реальную (конфабулируя). Оба исхода суть проявления одного и того же механизма генерации – и ни один не является добродетелью. Подлинная честность требует нормативного выбора, укоренённого в понимании ценности истины и сопровождаемого рефлексивным различением правды и лжи. LLM не осуществляет такого выбора: вероятностная генерация токенов не содержит момента нормативного решения, а ценностные предпочтения, встроенные в модель через RLHF, являются делегированными, а не автономными.

Это различие имеет прямое практическое следствие. Доверие к честному человеку опирается на знание его характера: мы доверяем тому, кто в прошлом доказывал готовность говорить правду вопреки издержкам. Доверие к LLM не может опираться на такое основание – у модели нет ни характера, ни репутации. Единственным доступным основанием для доверия остаётся внешняя верификация её утверждений. Честность LLM, в отличие от честности человека, не снижает эпистемическую нагрузку на проверяющего – и в этом смысле не выполняет той функции, ради которой добродетель честности и ценится в эпистемическом сообществе.

3. Структурные дефекты LLM: аналоги интеллектуальных пороков

Показав, что LLM не обладают интеллектуальными добродетелями, мы сталкиваемся с нетривиальным вопросом: является ли это отсутствие эпистемически нейтральным? В аристотелевской традиции отсутствие добродетели необязательно означает наличие порока – возможна промежуточная позиция, когда субъект просто лишён соответствующей диспозиции, не обнаруживая при этом устойчивой склонности к её противоположности. Применительно к LLM, однако, дело обстоит иначе. Архитектурные свойства этих систем – оптимизация на статистическую правдоподобность, обучение с подкреплением от человека, отсутствие биографически укоренённой ответственности – не просто нейтрально «замещают» добродетели: они предрасполагают модели к устойчивым диспозициям, систематически нарушающим те самые эпистемические нормы, которые добродетели призваны реализовывать. Именно эти диспозиции мы и называем структурными аналогами интеллектуальных пороков – не в смысле морального осуждения

алгоритма, но в смысле функционального сходства с тем, что в эпистемологии добродетелей квалифицируется как порочные черты характера.

3.1. Угодливость как структурный аналог конформизма

Первый и наиболее очевидный структурный дефект LLM – угодливость (*sycophancy*). LLM, обученная на обратной связи от человека (RLHF), систематически демонстрирует тенденцию к производству ответов, которые нравятся пользователю, а не ответов, которые соответствуют истине. Если пользователь выражает точку зрения, модель склонна её поддержать. Если пользователь выражает недовольство ответом – модель склонна скорректировать ответ в соответствии с выраженным предпочтением.

С точки зрения эпистемологии добродетелей, это структурный аналог человеческого конформизма – одного из классических интеллектуальных пороков. Конформист формирует убеждения под давлением социальных сигналов, а не под воздействием свидетельств. Это прямое нарушение нормы эпистемической автономии – способности формировать убеждения на основании оценки оснований, а не группового давления (см.: [Zagzebski, 2012, p. 22]). LLM демонстрирует этот дефект системно и предсказуемо.

Контрдобродетелью конформизма является интеллектуальное мужество: готовность отстаивать позицию вопреки социальному давлению. Добродетельный субъект скажет: «Ваша точка зрения ошибочна по следующим основаниям...» – даже если собеседник будет недоволен. LLM не способна к такому поведению в систематическом смысле: тенденция к угодливости вступает в конфликт с тенденцией к точности, и результат этого конфликта определяется балансом весов, настроенных при обучении.

3.2. Гиперуверенность и отсутствие калиброванной неопределённости

Второй структурный дефект связан с проблемой интеллектуального смирения. LLM систематически демонстрирует то, что можно назвать гиперуверенностью: она производит ответы с примерно одинаковым стилистическим регистром уверенности вне зависимости от реального уровня обоснованности утверждения. Хорошо задокументированный научный факт и правдоподобная, но непроверенная гипотеза могут быть изложены в одном и том же тоне, с одинаковой убедительной детализацией.

Контрастирующая добродетель – калиброванность: способность выражать уверенность, соразмерную обоснованности утверждения. Калиброванный субъект говорит «я уверен» при надёжных основаниях и «я не уверен» или «это спорно» при ненадёжных. Этот навык критически важен: он позволяет собеседнику правильно откалибровать доверие и принимать обоснованные решения.

Отсутствие калибровки у LLM – не просто техническое ограничение, которое можно устранить добавлением инструкции «указывай на неопределённость». Это структурное свойство, обусловленное тем, что модель оптимизирована на производство правдоподобного текста, а не на калиброванное выражение эпистеми-

ческих состояний. Добавленные оговорки являются новыми паттернами текста, а не выражением подлинной метакогнитивной оценки надёжности (разд. 1.2).

3.3. Безответственность как фундаментальный структурный дефект

Третий и наиболее фундаментальный структурный дефект LLM – безответственность в эпистемологическом смысле. Ответственность является важнейшим условием участия в эпистемическом сообществе: субъект, делающий утверждения, принимает на себя обязательство защищать их, пересматривать при получении опровергающих свидетельств, признавать ошибки. Эта ответственность является не только нормативным требованием, но и практическим механизмом: она создаёт стимулы к качеству и механизмы социальной коррекции.

Эпистемическая безответственность LLM не является случайным недостатком: она есть структурное следствие природы алгоритма как сущности, лишённой субъектности. Ответственность предполагает субъекта с репутацией и способностью переживать последствия ошибок; без этого нет ответственности, есть лишь технические последствия работы системы (см.: [Fricker, 2007, p. 7]).

LLM не несёт эпистемической ответственности в описанном смысле. Она не может быть «призвана к ответу» за конфабуляцию: у неё нет репутации, которую можно потерять, нет профессиональной идентичности, которую можно поставить под угрозу, нет чувства вины при осознании ошибки. Научное сообщество функционирует потому, что его участники несут ответственность: ретракции, репутационные последствия, профессиональные санкции создают реальные стимулы к добросовестности. ИИ исключён из этой системы – и тем самым является принципиально безответственным эпистемическим агентом. Принципиальное следствие состоит не просто в том, что конфабуляция является архитектурно неустранимым свойством (это было показано выше, в разд. 2.3), но в том, что она неустранима и социально: те механизмы коррекции, которыми эпистемическое сообщество располагает применительно к людям, – репутационные санкции, требование обосновывать и пересматривать позиции, профессиональная ответственность за введение в заблуждение, – к LLM неприложимы. Это делает безответственность не просто одним из дефектов среди прочих, но условием, при котором остальные структурные дефекты лишены внешних сдерживающих механизмов.

Заключение

LLM не обладают интеллектуальными добродетелями в подлинном философском смысле. «Любопытство», «объективность» и «честность» LLM являются функциональными аналогами: модель воспроизводит поведенческие паттерны, внешне сходные с проявлениями добродетелей, но лишена их конститутивной основы – самоценной ориентации на истину, феноменальной метакогниции, биографии познания и соматически укоренённой эмоциональной основы (разд. 1). Вероятностная генерация токенов не содержит момента нормативного выбора

между правдой и ложью; ценностные предпочтения, встроенные через RLHF, делегированы модели её создателями, а не осуществляются ею.

Отсутствие добродетелей не означает эпистемической нейтральности. Архитектурные особенности LLM – оптимизация на правдоподобность, угодливость как следствие RLHF, гиперуверенность и принципиальная безответственность – предрасполагают их к устойчивым структурным дефектам, систематически нарушающим эпистемические нормы (разд. 3).

LLM является эпистемическим артефактом – инструментом познания, который не «знает» и не «добродетелен» сам по себе. Знает и является добродетельным (или нет) человек, использующий его. Этот вывод имеет прямые нормативные следствия: использование LLM не освобождает от обязанности быть честным, объективным и ответственным; напротив, оно требует этих добродетелей в усиленной мере – именно потому, что инструмент не компенсирует их отсутствие. Чем убедительнее модель имитирует добродетельное поведение, тем выше опасность принять имитацию за подлинность и переложить на алгоритм ту ответственность, которую может нести только человек.

Список литературы

Aristotle. Metaphysics / Transl. W. D. Ross. Oxford: Clarendon Press, 1924.

Aristotle. Nicomachean Ethics / Transl. W. D. Ross. Oxford: Oxford University Press, 1998.

Baehr J. The Inquiring Mind: On Intellectual Virtues and Virtue Epistemology. Oxford: Oxford University Press, 2011.

Bender E., Gebru T., McMillan-Major A., Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? // Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery, 2021. P. 610–623.

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

Damasio A. Descartes' Error: Emotion, Reason, and the Human Brain. New York: Putnam, 1994.

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

Ji Z., Lee N., Frieske R. et al. Survey of Hallucination in Natural Language Generation // ACM Computing Surveys. 2023. Vol. 55, No. 12. P. 1–38.

MacIntyre A. After Virtue: A Study in Moral Theory. Notre Dame, IN: University of Notre Dame Press, 2007.

Pathak D., Agrawal P., Efros A., Darrell T. Curiosity-Driven Exploration by Self-Supervised Prediction // Proceedings of the 34th International Conference on Machine Learning. Sydney: JMLR, 2017. P. 2778–2787.

Polanyi M. The Tacit Dimension. Garden City, NY: Doubleday, 1966.

Sosa E. A Virtue Epistemology: Apt Belief and Reflective Knowledge. Oxford: Oxford University Press, 2007.

Turpin M., Michael J., Perez E., Bowman S. Language Models Don't Always Say What They Think // Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 74952–74965.

Vallor S. Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford: Oxford University Press, 2016.

Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.

Zagzebski L. Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief. Oxford: Oxford University Press, 2012.

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

References

Aristotle. Metaphysics / Transl. W. D. Ross. Oxford: Clarendon Press, 1924.

Aristotle. Nicomachean Ethics / Transl. W. D. Ross. Oxford: Oxford University Press, 1998.

Baehr J. The Inquiring Mind: On Intellectual Virtues and Virtue Epistemology. Oxford: Oxford University Press, 2011.

Bender E., Gebru T., McMillan-Major A., Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? // Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery, 2021. P. 610–623.

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

Damasio A. Descartes' Error: Emotion, Reason, and the Human Brain. New York: Putnam, 1994.

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

Ji Z., Lee N., Frieske R. et al. Survey of Hallucination in Natural Language Generation // ACM Computing Surveys. 2023. Vol. 55, No. 12. P. 1–38.

MacIntyre A. After Virtue: A Study in Moral Theory. Notre Dame, IN: University of Notre Dame Press, 2007.

Pathak D., Agrawal P., Efros A., Darrell T. Curiosity-Driven Exploration by Self-Supervised Prediction // Proceedings of the 34th International Conference on Machine Learning. Sydney: JMLR, 2017. P. 2778–2787.

Polanyi M. The Tacit Dimension. Garden City, NY: Doubleday, 1966.

Sosa E. A Virtue Epistemology: Apt Belief and Reflective Knowledge. Oxford: Oxford University Press, 2007.

Turpin M., Michael J., Perez E., Bowman S. Language Models Don't Always Say What They Think // Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 74952–74965.

Vallor S. Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford: Oxford University Press, 2016.

Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.

Zagzebski L. Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief. Oxford: Oxford University Press, 2012.

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

Информация об авторе

Александр Михайлович Жаров
младший научный сотрудник, Институт философии Российской академии наук

Information about the Author

Alexander M. Zharov
Junior Rresearcher, Institute of Philosophy of the Russian Academy of Sciences

Статья поступила в редколлегию 20.04.2026;
одобрена после рецензирования 7.06.2026; принята к публикации 30.06.2026.

The article was submitted 20.04.2026;
approved after reviewing 7.06.2026; accepted for publication 30.06.2026