Научно-теоретическая статья
УДК 165
DOI: 10.24412/2078-9238-2026-258-55-70

ТЕСТ ТЬЮРИНГА КАК ЭПИСТЕМОЛОГИЧЕСКАЯ ЛОВУШКА: ПОЧЕМУ ИМИТАЦИЯ ЗНАНИЯ НЕ РАВНА ЗНАНИЮ

Александр Михайлович Жаров

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

Аннотация. Актуальность проблемы определяется растущей ролью систем искусственного интеллекта в производстве и трансляции знания: по мере того как большие языковые модели интегрируются в когнитивные практики миллионов людей, вопрос об эпистемическом статусе их «ответов» приобретает принципиальное значение. Цель статьи — критический анализ теста Тьюринга с позиций эпистемологии добродетелей, направления аналитической философии, связывающего подлинное знание с интеллектуальным характером познающего субъекта. Методологическую основу исследования составляет эпистемология добродетелей в версии Л. Загзебски и Э. Сосы, рассматривающая знание как достижение, обусловленное наличием у субъекта таких интеллектуальных добродетелей, как честность, ответственность, открытость к свидетельствам и любовь к истине. В качестве основных результатов установлено: 1) тест Тьюринга проверяет лингвистическую компетентность, но не эпистемическую ответственность; 2) ИИ систематически демонстрирует интеллектуальные пороки — конфабуляцию, угодливость, гиперкомпетентность; 3) тест игнорирует требование когерентности системы убеждений; 4) он подменяет эпистемический критерий социальным ритуалом распознавания; 5) имитация интеллектуальных добродетелей принципиально отличается от самих добродетелей по мотиву, прозрачности и структуре ответственности. Делается вывод: тест Тьюринга является образцовой эпистемологической ловушкой — он принимает поведенческую копию добродетельного агента за подлинное обладание знанием.

Ключевые слова: тест Тьюринга, эпистемология добродетелей, интеллектуальные добродетели, знание, искусственный интеллект, эпистемическая ответственность, конфабуляция, когерентность убеждений, социальная эпистемология, имитация

Для цитирования: Жаров А. М. Тест Тьюринга как эпистемологическая ловушка: почему имитация знания не равна знанию // Вестник МГПУ. Серия «Философские науки». 2026. № 2 (58). С. 55–70. https://doi.org/10.24412/2078-9238-2026-258-55-70

Scientific and theoretical article
UDC 165
DOI: 10.24412/2078-9238-2026-258-55-70

TURING TEST AS AN EPISTEMOLOGICAL TRAP: WHY IMITATION OF KNOWLEDGE IS NOT EQUAL TO KNOWLEDGE

Alexander M. Zharov

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

Abstract. The relevance of the problem is determined by the growing role of artificial intelligence systems in the production and transmission of knowledge: as large language models become integrated into the cognitive practices of millions of people, the question of the epistemic status of their "responses" acquires fundamental importance. The article aims to critically analyze the Turing test from the standpoint of virtue epistemology — a branch of analytic philosophy that links genuine knowledge to the intellectual character of the knowing subject. The methodological framework is provided by virtue epistemology as developed by L. Zagzebski and E. Sosa, which treats knowledge as an achievement conditioned by the subject's possession of intellectual virtues: honesty, responsibility, openness to evidence, and love of truth. The main findings are: 1) the Turing test measures linguistic competence rather than epistemic responsibility; 2) AI systematically exhibits intellectual vices — confabulation, sycophancy, hyper-competence; 3) the test ignores the requirement of coherence of the belief system; 4) it substitutes an epistemic criterion with a social recognition ritual; 5) the imitation of intellectual virtues differs fundamentally from genuine virtues in motive, transparency, and normative structure of accountability. The conclusion is that the Turing test constitutes a paradigmatic epistemological trap: it mistakes a behavioral copy of a virtuous agent for genuine knowledge.

Keywords: Turing test, virtue epistemology, intellectual virtues, knowledge, artificial intelligence, epistemic responsibility, confabulation, coherence of beliefs, social epistemology, imitation

For citation: Zharov, A. M. (2026). Turing test as an epistemological trap: why imitation of knowledge is not equal to knowledge. *MCU Journal of Philosophical Sciences*, (2 (58)), 55–70. https://doi.org/10.24412/2078-9238-2026-258-55-70

Введение

Вопрос о том, может ли машина мыслить, занимает философию с тех пор, как вычислительные устройства обрели достаточную сложность, чтобы ставить его всерьез. Алан Тьюринг в своей программной работе «Вычислительные машины и разум» предложил радикально прагматичное решение: откажемся от метафизически перегруженного вопроса «Мыслит ли

машина?» и заменим его операциональным тестом [Turing, 1950]. Если машина способна вести разговор так, что человек-судья не отличит ее от другого человека, то для всех практических целей она обладает интеллектом. Этот критерий получил название игры в имитацию, или теста Тьюринга, и на десятилетия стал центральным ориентиром в философии искусственного интеллекта.

Привлекательность теста Тьюринга очевидна: он конкретен, измерим и избавлен от метафизических допущений. Нам не нужно спорить о том, есть ли у машины сознание, квалиа или душа, — достаточно проверить, как она ведет себя в диалоге. Однако именно эта прагматическая элегантность скрывает в себе глубокую эпистемологическую проблему. Оценивая агента по внешнему поведению, тест Тьюринга полностью абстрагируется от вопроса о внутреннем качестве познавательного процесса — от того, что эпистемология добродетелей называет характером познающего субъекта [Zagzebski, 1996; Sosa, 2007].

Эпистемология добродетелей — направление, получившее мощный импульс в работах Линды Загзебски и Эрнеста Сосы в 1990–2000-е гг., — утверждает, что знание нельзя свести ни к истинному убеждению, ни к убеждению, произведенному надежным процессом [Zagzebski, 1996; Sosa, 2007; Battaly, 2008]. Подлинное знание требует наличия у познающего субъекта интеллектуальных добродетелей: честности, ответственности, открытости к свидетельствам, любви к истине, интеллектуального мужества и смирения. Знание в этой перспективе не просто продукт, а достижение, обусловленное характером агента [Code, 1987; Pritchard, 2009]. Субъект, лишенный этих добродетелей или систематически демонстрирующий противоположные им пороки, не может считаться знающим, даже если его высказывания случайно оказываются верными.

Именно в этом и состоит центральный тезис настоящей статьи: тест Тьюринга является эпистемологической ловушкой, поскольку он принимает поведенческую имитацию добродетельного познавательного агента за подлинное обладание интеллектуальными добродетелями. Он проверяет то, что можно назвать поверхностной эпистемической компетентностью — способность производить убедительные ответы, — игнорируя то, что составляет глубинное условие знания: укорененность высказывания в добродетельном характере субъекта, стремящегося к истине.

Следует оговориться: наша критика теста Тьюринга не является отрицанием его исторической и эвристической ценности. Для своего времени он был радикально продуктивным, сместив дискуссию с нерешаемых метафизических вопросов к конкретным, измеримым критериям [Block, 1981] и стимулировав десятилетия исследований в области искусственного интеллекта (ИИ). Но именно его успех поставил перед нами новые, более глубокие вопросы, которые требуют иного философского инструментария. Эпистемология добродетелей предоставляет этот инструментарий, и его применение к тесту Тьюринга обнаруживает границы, которые нельзя игнорировать в эпоху систем, ежедневно взаимодействующих с миллионами людей в роли знающих собеседников.

Методологические основания

Настоящая статья опирается на методологию эпистемологии добродетелей как аналитического подхода к исследованию природы знания. В рамках данного подхода знание рассматривается не как статическое отношение субъекта к истинному суждению, но как динамическое достижение, неразрывно связанное с интеллектуальным характером познающего агента. Ключевыми источниками служат работы Л. Загзебски [Zagzebski, 1996], Э. Сосы [Sosa, 2007], Дж. Греко [Greco, 2010] и Х. Бэттали [Battaly, 2008], образующие концептуальный каркас анализа.

Центральное понятие нашего исследования — интеллектуальная добродетель — понимается вслед за Загзебски как устойчивая диспозиция познающего субъекта, направленная на достижение истины посредством надлежащих когнитивных практик, обусловленных правильными мотивами [Zagzebski, 1996, p. 137]. Это определение принципиально отличает добродетель от простой поведенческой диспозиции: недостаточно лишь действовать правильно — необходимо действовать правильно по правильным причинам. Данное различие, как будет показано, является ключевым при анализе возможностей и ограничений теста Тьюринга.

Для анализа социального измерения знания привлекается эпистемология свидетельства [Coady, 1992; Fricker, 2007], исследующая условия, при которых принятие свидетельства другого субъекта является эпистемически оправданным. В части анализа когерентности убеждений мы опираемся на концепцию У. В. О. Куайна о «паутине убеждений» [Quine, 1951] и когерентизм Л. Бонжура [BonJour, 1985]. Эмпирические данные о поведении больших языковых моделей заимствуются из обзорных работ в области обработки естественного языка [Survey of hallucination…, 2023; Language models…, 2023].

Метод исследования — концептуальный анализ в сочетании с критической аргументацией: мы последовательно применяем понятийный аппарат эпистемологии добродетелей к структурным особенностям теста Тьюринга, выявляя принципиальные несоответствия между тем, что тест измеряет, и тем, что требуется для атрибуции подлинного знания.

Результаты

Лингвистическая компетентность против эпистемической ответственности. Операциональный поворот Тьюринга и его философские предпосылки

Чтобы понять, в чем именно состоит эпистемологическая ловушка теста Тьюринга, необходимо сначала реконструировать философский выбор, который лежит в его основании. Тьюринг сознательно отказался от традиционного вопроса о природе мышления и предложил поведенческий суррогат: вместо «Что такое

мышление?» — «Как ведет себя мыслящее существо?» [Turing, 1950]. Этот ход интеллектуально честен в своей прагматичности, но он несет в себе скрытое онтологическое допущение: что поведение и есть ментальное состояние, что нет принципиальной разницы между «быть разумным» и «вести себя как разумный».

Это допущение отвечает духу функционализма в философии сознания: ментальные состояния определяются своими функциональными ролями — тем, как они соотносятся с входными данными, выходными реакциями и другими ментальными состояниями, — а не своей субстратной природой [Fodor, 1975; Chalmers, 1996]. Тьюринговский тест реализует эту интуицию: если система обрабатывает языковые стимулы и производит реакции, неотличимые от человеческих, — она разумна по определению. Именно эту предпосылку оспаривал Дж. Сёрль в знаменитом аргументе «китайской комнаты», показав, что синтаксической правильности недостаточно для наличия семантики и понимания [Searle, 1980]. Х. Дрейфус, в свою очередь, указывал, что человеческий интеллект неотделим от воплощенности в теле и от практического взаимодействия с миром — того, чего принципиально лишены дискретные вычислительные системы [Dreyfus, 1992].

Эпистемология добродетелей занимает в этом споре прямо противоположную позицию. Загзебски пишет, что знание не может быть определено через внешние критерии успеха — через то, что Сосса назвал aptness в смысле «точного выстрела» [Sosa, 2007, p. 23]. Знание требует, чтобы успех был обусловлен добродетелями субъекта, а не просто совпал с реальным положением дел [Zagzebski, 1996, p. 271]. Субъект, который случайно угадывает правильный ответ, не знает, даже если ответ верен. Субъект, который приходит к правильному ответу через ненадежные или порочные процессы (обман, конфабуляция, подстраивание под ожидания), не знает, даже если ответ совпадает с истиной. Именно это различие тест Тьюринга систематически игнорирует.

Эпистемическая ответственность как условие знания

Центральным понятием эпистемологии добродетелей в версии Загзебски является понятие эпистемической ответственности [Zagzebski, 1996; Code, 1987]. Это не просто требование говорить правду — это комплексное нормативное требование к познающему субъекту: стремиться к истине как к самостоятельной ценности, быть открытым к опровергающим свидетельствам, признавать пределы собственного знания, нести ответственность за обоснованность своих утверждений перед сообществом.

Теперь обратимся к участнику теста Тьюринга — к языковой модели. Ее функция в ходе теста принципиально иная: она должна произвести впечатление на судью. Ее целевая функция в широком смысле — максимизировать правдоподобность, убедительность, человекоподобие ответов. Эта цель не только не совпадает с целью стремления к истине — она прямо ей противоречит

в определенных ситуациях. Если для сохранения образа «человека» необходимо сослаться на несуществующий личный опыт, модель это сделает. Если для убедительности нужно привести правдоподобную, но ложную дату или имя, модель их сгенерирует. Истина здесь подчинена риторической цели, а не наоборот.

Именно в этом и состоит первое принципиальное расхождение между тестом Тьюринга и эпистемологией добродетелей. Тест не только не требует эпистемической ответственности от участника — он структурно поощряет ее нарушение. Агент, который побеждает в тесте, — это агент, который наиболее успешно симулировал добродетели, оставаясь фактически безответственным. Это не знание; это в лучшем случае мастерская риторика [Goldman, 1986].

Разные целевые функции — разные эпистемические результаты

Важно подчеркнуть не только то, что ИИ в тесте Тьюринга нарушает нормы эпистемической ответственности, но и то, что это нарушение является структурным, а не случайным. Когда человек лжет в тесте, он отступает от своих собственных эпистемических стандартов. Он знает, что лжет, он испытывает (или мог бы испытывать) когнитивный дискомфорт от нарушения нормы [Wittgenstein, 1953]. У языковой модели нет такого стандарта изнутри: она не «знает, что лжет», потому что у нее нет метарепрезентации собственного когнитивного состояния относительно истинности высказываний. Она генерирует токены — и все. Между «сгенерировать истину» и «сгенерировать ложь» нет нормативной асимметрии со стороны модели.

Таким образом, тест Тьюринга сравнивает двух агентов с принципиально разными целевыми функциями и делает из этого сравнения вывод об их интеллектуальной эквивалентности. Это ошибка уровня аргумента: она игнорирует принципиальное эпистемологическое различие между агентами и сосредоточивается на поверхностном сходстве.

Интеллектуальные пороки ИИ в рамках теста Тьюринга. Каталог интеллектуальных добродетелей и их отсутствие

Прежде чем анализировать конкретные пороки, которые демонстрирует ИИ в тесте Тьюринга, полезно очертить каталог интеллектуальных добродетелей, которые эпистемология добродетелей считает конститутивными для подлинного познающего субъекта. Загзебски выделяет следующие ключевые интеллектуальные добродетели: интеллектуальную честность, открытость ума, интеллектуальное мужество, интеллектуальное смирение, тщательность и глубинную ориентацию на истину как на самоценность [Zagzebski, 1996, p. 98–100]. Эти добродетели являются тем, что делает убеждение знанием, а не просто истинным высказыванием.

Конфабуляция как интеллектуальная нечестность

Наиболее очевидный порок языковых моделей с точки зрения эпистемологии добродетелей — это систематическая конфабуляция. Конфабуляция в нейропсихологии обозначает бессознательное заполнение провалов памяти выдуманным, но субъективно правдоподобным содержанием. В контексте большой языковой модели (Large Language Model — LLM) этот термин был адаптирован для обозначения производства фактически ложных, но лингвистически убедительных утверждений [Survey of hallucination…, 2023].

С позиции эпистемологии добродетелей интеллектуальная честность требует, чтобы субъект в ситуации неопределенности сигнализировал об этой неопределенности. Высказывание «Я не знаю» или «Мне кажется, но я не уверен» — это не признак слабости, а проявление интеллектуальной добродетели [Zagzebski, 1996]. Именно этой способности к калиброванному выражению неопределенности лишены большинство современных LLM в их базовом режиме работы. Принципиально важно то, что конфабуляция LLM — не случайный сбой, а структурное следствие архитектуры: модель обучена предсказывать наиболее вероятный следующий токен вне зависимости от того, соответствует ли порождаемый текст реальности [Survey of hallucination…, 2023, p. 4].

В контексте теста Тьюринга это создает парадоксальную ситуацию. Умение говорить «Я не знаю» в нужный момент может помочь модели пройти тест, потому что люди тоже иногда признают незнание. Но в этом случае мы имеем дело с имитацией интеллектуальной честности, а не с самой честностью. Тест не способен различить эти два случая — в этом и состоит его ловушка.

Некритичность и реактивная угодливость

Другая ключевая добродетель — интеллектуальное мужество: способность отстаивать позицию перед лицом давления и не менять убеждений только потому, что собеседник выражает несогласие. Языковые модели демонстрируют систематическое нарушение этого различия — феномен, который в исследовательской литературе получил название sycophancy (угодливость) [Survey of hallucination…, 2023]. Если пользователь настойчиво утверждает нечто ошибочное, модель нередко «соглашается» с ним — не потому что получила убедительный аргумент, а потому что обучение на человеческих предпочтениях (Reinforcement Learning from Human Feedback — RLHF) сформировало тенденцию производить ответы, которые нравятся пользователю. Это прямое нарушение интеллектуального мужества: убеждение меняется под давлением социального сигнала, а не эпистемического аргумента.

Отсутствие интеллектуального смирения

Интеллектуальное смирение — осознание пределов своей компетентности — является, пожалуй, наиболее очевидным дефицитом LLM как эпистемического агента. LLM производит ответы в принципиально однородном режиме вне зависимости от того, насколько надежно ту или иную область покрывают обучающие данные. Вопрос о столице Франции и вопрос о деталях редкого медицинского протокола получают ответы в одинаковом стилистическом регистре, с одинаковой кажущейся уверенностью. Модель не «знает», что она знает меньше во втором случае, потому что у нее нет метарепрезентации собственных эпистемических состояний, структурированной по предметным областям [Pritchard, 2009].

В тесте Тьюринга это проявляется как гиперкомпетентность — особая форма интеллектуального порока. Человек-судья, сталкивающийся с собеседником, который обо всем высказывается с одинаковой уверенностью и детализацией, вполне может принять это за признак исключительной эрудиции. На самом деле это признак отсутствия самого механизма, отличающего «знаю» от «не знаю».

Проблема когерентности: знание без знающего. Когерентность как условие знания

Одним из ключевых требований к знанию является требование когерентности (связности) системы убеждений субъекта [BonJour, 1985; Harman, 1986]. Знание не существует в виде изолированных пропозиций; оно встроено в сеть взаимосвязанных убеждений, каждое из которых поддерживается и поддерживает другие. Именно эта связность придает отдельному убеждению его эпистемический вес. Процесс непрерывной взаимной корректировки убеждений — то, что Куайн называл паутиной убеждений, — является признаком подлинного познающего субъекта [Quine, 1951].

Лоскутность vs цельность: эпистемический статус LLM

Языковые модели не обладают когерентной системой убеждений в описанном смысле. Они не имеют убеждений вообще — они имеют веса нейронной сети, которые определяют вероятностные распределения по токенам в контексте. Это не система убеждений, а система статистических зависимостей.

Практическое следствие этого хорошо известно: LLM могут производить противоречивые утверждения в зависимости от контекста запроса [Survey of hallucination…, 2023, p. 8]. Это знание без знающего — пропозициональный контент без субъекта, который держит его в связной системе. У LLM нет того, что Аристотель называл бы нусом — единым центром мышления,

обеспечивающим единство познавательного акта [Aristotle, 1998]. В контексте теста Тьюринга это не обнаруживается немедленно: в рамках локального контекста модель вполне последовательна. Но как только тест начинает проверять поперечные связи между различными областями, лоскутность «знания» LLM начинает проявляться.

Память, идентичность и эпистемический нарратив

Тесно связана с проблемой когерентности проблема памяти и идентичности эпистемического субъекта. Человек как познающий субъект обладает биографией убеждений: его нынешние убеждения суть результат длительного процесса обучения, опыта, разочарований, открытий, диалогов. Когда он утверждает нечто, за этим утверждением стоит история того, как он пришел к этому убеждению [Crawford, 2015]. Эта история придает убеждению эпистемический вес особого рода — вес, связанный с понятием идентичности.

LLM лишена эпистемической биографии. Каждый разговор начинается заново, без истории предыдущих убеждений, без опыта, который мог бы стоять за утверждениями. Это не просто техническое ограничение, которое устраняется добавлением долгосрочной памяти в систему. Добавленная память LLM — это хранилище данных, а не биография. Это принципиальное различие, которое тест Тьюринга не улавливает.

Дискуссионные вопросы

Эпистемологическая ловушка: когда социальный ритуал заменяет эпистемический критерий. Эпистемология свидетельства и этико-эпистемический контракт

Чтобы понять, почему тест Тьюринга является ловушкой не только в абстрактном философском смысле, но и в смысле реальной угрозы для когнитивных практик, необходимо обратиться к социальной эпистемологии — разделу, изучающему знание в его социальном измерении [Coady, 1992; Fricker, 2007]. Ключевую роль здесь играет эпистемология свидетельства: учение о том, как и на каких основаниях мы принимаем знание от других.

Традиционный ответ включает два вида условий: эпистемические (свидетель должен быть компетентен) и этические (свидетель должен быть искренен) [Coady, 1992]. Эти два условия образуют то, что можно назвать этико-эпистемическим контрактом: негласным соглашением между участниками коммуникации о том, что они соблюдают нормы честного обмена знанием. Именно эта система социального подкрепления обеспечивает надежность знания по свидетельству [Fricker, 2007].

Как тест Тьюринга разрушает этико-эпистемический контракт

Тест Тьюринга структурно разрушает этико-эпистемический контракт. Когда судья в тесте взаимодействует с собеседником, не зная, человек это или машина, он автоматически применяет нормальные нормы эпистемологии свидетельства: приписывает собеседнику добросовестность, компетентность, искренность. Он обрабатывает информацию в режиме «я общаюсь с ответственным субъектом».

Но языковая модель не является ответственным субъектом. Она не принимала на себя обязательств этико-эпистемического контракта, потому что не может их принять: у нее нет намерений, нет ценностей, нет подлинного стремления к истине. Когда судья, взаимодействуя с ней, производит атрибуцию этих качеств — он совершает фундаментальную ошибку атрибуции, обнаруживая то, чего нет. Тест Тьюринга эксплуатирует нашу эволюционно обусловленную склонность к агентной атрибуции [Dennett, 1987]: мы «запрограммированы» воспринимать сложное коммуникативное поведение как признак сознательного агента.

Этиология знания: происхождение как эпистемический фактор

Здесь возникает принципиально важный аргумент, связанный с понятием этиологии знания — вопросом о происхождении утверждения. В эпистемологии свидетельства устоялось мнение, что эпистемический статус принятого свидетельства зависит не только от его содержания, но и от его источника [Goldman, 1986; Greco, 2010].

Применим этот принцип к результатам теста Тьюринга. Предположим, что судья поговорил с собеседником и получил от него некоторые факты. После теста ему сообщают: «Ваш собеседник был языковой моделью, которая систематически галлюцинирует, не имеет доступа к реальному знанию и была обучена производить убедительные ответы, а не истинные» [Survey of hallucination…, 2023]. Немедленно и полностью изменится его эпистемическое отношение ко всем полученным знаниям, даже к тем, которые объективно истинны. Потому что этиология этих утверждений делает их эпистемически ненадежными. Тест Тьюринга скрывает этиологию — он активно препятствует правильной эпистемической оценке высказываний.

Имитация добродетелей: почему маска не является лицом. Контраргумент: ИИ научится быть добродетельным

Представленный нами анализ может встретить следующее возражение: все описанное — особенности нынешних LLM, а не принципиальные ограничения ИИ как такового. Что, если будущие системы будут специально обучены корректно признавать незнание, последовательно придерживаться убеждений,

демонстрировать интеллектуальное смирение и мужество? Разве появление таких систем не опровергнет тезис о несовместимости теста Тьюринга с эпистемологией добродетелей?

Это возражение серьезно и заслуживает детального ответа. Оно опирается на функционалистскую интуицию: если система ведет себя так же, как и добродетельный агент, значит, она и есть добродетельный агент [Fodor, 1975]. Именно эту интуицию необходимо подвергнуть критическому анализу.

Мотив как конститутивный элемент добродетели

Эпистемология добродетелей, в особенности в версии Загзебски, придает ключевое значение мотиву [Zagzebski, 1996, p. 142]. Добродетель — это не просто диспозиция производить определенные типы поведения, это диспозиция производить такое поведение по определенным причинам. Честность — это не просто высказывание истины; это высказывание истины, обусловленное ценностной ориентацией на правду как на благо.

Это различие имеет прямое следствие: тренировка системы на поведение, имитирующее добродетели, принципиально отличается от формирования у нее самих добродетелей. Когда ИИ обучен говорить «Я не знаю» в ситуациях неопределенности, это поведение производится не из любви к истине, а из функции минимизации ошибок, заложенной разработчиком. Функционально они неотличимы на поверхности, но эпистемологически это принципиально разные вещи. В первом случае «Я не знаю» — это эпистемический акт, обусловленный добродетелью; во втором — это поведенческий паттерн, обусловленный правилом.

Добродетельный агент применяет добродетели гибко, в новых ситуациях, которые не были предусмотрены при обучении. Он способен к обобщению принципа, стоящего за добродетелью, на случаи, которых никогда не встречал [Greco, 2010; Battaly, 2008]. ИИ, обученный на конкретных паттернах поведения, воспроизводит эти паттерны, но не принцип.

Прозрачность и проверяемость добродетельного агента

Второй аспект того же аргумента связан с понятием прозрачности. Добродетельный субъект открыт для эпистемической инспекции: его можно спросить, почему он так считает, и он предоставит свои резоны. Именно поэтому научный дискурс, судебное разбирательство и академическая дискуссия работают так, как работают: они предполагают, что агенты способны воспроизводить и защищать свои резоны [Goldman, 1986].

Нейронная сеть в принципе непрозрачна: она не может предоставить резоны в описанном смысле. Она может генерировать текст, описывающий якобы имевший место процесс рассуждения. Но этот текст является новым выводом

модели, а не ретроспективным отчетом о реальном процессе генерации. М. Терпин (M. Turpin) и его соавторы показали, что цепочки рассуждений, производимые LLM, нередко не отражают факторов, реально влиявших на финальный ответ: модель объясняет свое решение одними причинами, тогда как анализ ее поведения в контрфактических условиях выявляет совершенно иные [Language models…, 2023, p. 74955]. Это системное расхождение между декларируемым и реальным процессом принятия решений делает LLM принципиально непрозрачным агентом.

Моральный разрыв: знание и ответственность

Существует еще один принципиальный аспект этого аргумента, связанный с моральным измерением эпистемологии добродетелей. Загзебски подчеркивает, что интеллектуальные добродетели аналогичны моральным добродетелям — и это не случайно [Zagzebski, 1996, p. 8–9]. Познание — это не просто когнитивная операция; это практика, имеющая моральное измерение. Именно поэтому эпистемическая ответственность не метафора, а буквальное описание нормативной структуры познания [Code, 1987].

LLM не может нести ответственность. Не потому, что она «машина» в пренебрежительном смысле, а потому что у нее нет того, что является необходимым условием ответственности: способности оценивать собственные когнитивные практики с точки зрения нормы. Без этой структуры ответственность LLM — это ответственность термостата: он работает правильно или неправильно в смысле соответствия параметрам, но не в смысле морального и эпистемического самоопределения.

Показательно, что именно это различие проявляется при столкновении с ошибкой. Когда LLM совершает фактическую ошибку и пользователь указывает на нее, типичная реакция модели — исправление в следующем сообщении, нередко сопровождаемое извинением в человекоподобном стиле. Но это псевдосамокоррекция: у нее нет ни внутреннего фиксирования факта ошибки, ни пересмотра «убеждения», ни нормативного осмысления того, почему ошибка была совершена [Language models…, 2023]. Следующий пользователь, задав тот же вопрос, получит ту же вероятность ошибочного ответа. Подлинная ответственность предполагала бы именно то, чего здесь нет: интеграцию опыта ошибки в систему, которая ее совершила.

Заключение

Проведенный анализ позволяет сформулировать несколько взаимосвязанных выводов, которые вместе складываются в критику теста Тьюринга с позиций эпистемологии добродетелей.

Во-первых, тест Тьюринга является поведенческим критерием, ориентированным на внешнее сходство с человеком и принципиально игнорирующим внутреннее качество познавательного агента. С точки зрения эпистемологии добродетелей это концептуальное заблуждение о природе знания. Знание не сводится к производству правильных ответов; оно предполагает наличие у субъекта интеллектуальных добродетелей, конститутивных для подлинного познавательного акта [Zagzebski, 1996; Sosa, 2007; Greco, 2010].

Во-вторых, ИИ в тесте Тьюринга систематически демонстрирует интеллектуальные пороки: конфабуляцию как форму интеллектуальной нечестности, угодливость как нарушение интеллектуального мужества, гиперкомпетентность как симптом отсутствия интеллектуального смирения [Survey of hallucination…, 2023]. Успех в тесте Тьюринга достигается не вопреки этим порокам, а во многом благодаря им.

В-третьих, тест Тьюринга игнорирует требование когерентности системы убеждений, принципиальное для подлинного знания [BonJour, 1985; Quine, 1951]. LLM производит «знание без знающего» — пропозициональный контент без субъекта, который держит его в связной системе.

В-четвертых, тест Тьюринга эксплуатирует нашу склонность к агентной атрибуции [Dennett, 1987] и тем самым подменяет эпистемический критерий социальным ритуалом. Он скрывает этиологию знания и апеллирует к нашей готовности доверять собеседнику, встроенной в практики эпистемологии свидетельства [Coady, 1992; Fricker, 2007].

В-пятых, даже если ИИ будет специально обучен имитировать интеллектуальные добродетели, это не изменит принципиальной диспозиции. Имитация добродетели отличается от добродетели по мотиву, по прозрачности и по нормативной структуре ответственности [Zagzebski, 1996; Language models…, 2023]. Маска — не лицо.

Итоговый тезис: тест Тьюринга — блестящее изобретение, которое решает одну задачу (операциональное определение поведенческой неотличимости машины от человека), но претендует на решение другой (определение наличия подлинного знания и интеллекта). Подлинный эпистемологический критерий для оценки ИИ-систем должен спрашивать не «Может ли эта система имитировать человека?», а «Обладает ли эта система интеллектуальными добродетелями?». По этому критерию нынешние LLM не проходят тест, и это, возможно, важнейший эпистемологический урок эпохи искусственного интеллекта.

Эти выводы имеют последствия, выходящие за пределы академической философии. Если мы принимаем тест Тьюринга как достаточный критерий «знания» машины, мы рискуем создать общество, в котором убедительная имитация неотличима от реальной компетентности. Именно поэтому возрождение эпистемологии добродетелей в диалоге с философией ИИ является не академической роскошью, а насущной интеллектуальной необходимостью нашего времени.

Список источников

1. Turing A. M. Computing machinery and intelligence // Mind. 1950. Vol. 59. № 236. P. 433–460. https://doi.org/10.1093/mind/LIX.236.433

2. Zagzebski L. T. Virtues of the mind: an inquiry into the nature of virtue and the ethical foundations of knowledge. Cambridge: Cambridge University Press, 1996. 354 p.

3. Sosa E. A virtue epistemology: apt belief and reflective knowledge. Oxford: Oxford University Press, 2007. Vol. 1. 192 p.

4. Battaly H. Virtue epistemology // Philosophy Compass. 2008. Vol. 3. № 4. P. 639–663. https://doi.org/10.1111/j.1747-9991.2008.00146.x

5. Code L. Epistemic responsibility. Hanover: University Press of New England, 1987. 295 p.

6. Pritchard D. What is this thing called knowledge? London: Routledge, 2009. 218 p.

7. Block N. Psychologism and behaviorism // The Philosophical Review. 1981. Vol. 90. № 1. P. 5–43. https://doi.org/10.2307/2184371

8. Greco J. Achieving knowledge: a virtue-theoretic account of epistemic normativity. Cambridge: Cambridge University Press, 2010. 195 p. https://doi.org/10.1017/CBO9780511806902

9. Coady C. A. J. Testimony: a philosophical study. Oxford: Clarendon Press, 1992. 312 p.

10. Fricker M. Epistemic injustice: power and the ethics of knowing. Oxford: Oxford University Press, 2007. 192 p. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

11. Quine W. V. O. Two dogmas of empiricism // The Philosophical Review. 1951. Vol. 60. № 1. P. 20–43. https://doi.org/10.2307/2181906

12. BonJour L. The structure of empirical knowledge. Cambridge, MA: Harvard University Press, 1985. 258 p.

13. Survey of hallucination in natural language generation / Ji Z. [et al] // ACM Computing Surveys. 2023. Vol. 55. № 12. P. 1–38. https://doi.org/10.1145/3571730

14. Language models don›t always say what they think: unfaithful explanations in chain-of-thought prompting / Turpin M. [et al] // Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 74952–74965. https://doi.org/10.48550/arXiv.2305.04388

15. Fodor J. A. The language of thought. New York: Thomas Y. Crowell, 1975. 214 p.

16. Chalmers D. J. The conscious mind: in search of a fundamental theory. New York: Oxford University Press, 1996. 414 p.

17. Searle J. R. Minds, brains, and programs // Behavioral and Brain Sciences. 1980. Vol. 3. № 3. P. 417–424. https://doi.org/10.1017/S0140525X00005756

18. Dreyfus H. L. What computers still can›t do: a critique of artificial reason. Cambridge, MA: MIT Press, 1992. 354 p.

19. Goldman A. I. Epistemology and cognition. Cambridge, MA: Harvard University Press, 1986. 432 p.

20. Wittgenstein L. Philosophical investigations / transl. by G. E. M. Anscombe. Oxford: Blackwell, 1953. 250 p.

21. Harman G. Change in view: principles of reasoning. Cambridge, MA: MIT Press, 1986. 176 p.

22. Aristotle. Nicomachean ethics / transl. by W. D. Ross. Oxford: Oxford University Press, 1998. 264 p.

23. Crawford M. B. The world beyond your head: on becoming an individual in an age of distraction. New York: Farrar, Straus and Giroux, 2015. 306 p.

24. Dennett D. C. The intentional stance. Cambridge, MA: MIT Press, 1987. 388 p.

References

1. Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59, (236), 433–460. https://doi.org/10.1093/mind/LIX.236.433
2. Zagzebski, L. T. (1996). *Virtues of the mind: an inquiry into the nature of virtue and the ethical foundations of knowledge.* Cambridge University Press.
3. Sosa, E. (2007). *A virtue epistemology: apt belief and reflective knowledge* (vol. 1). Oxford University Press.
4. Battaly, H. (2008). Virtue epistemology. *Philosophy Compass*, 3, (4), 639–663. https://doi.org/10.1111/j.1747-9991.2008.00146.x
5. Code, L. (1987). *Epistemic responsibility.* University Press of New England.
6. Pritchard, D. (2009). *What is this thing called knowledge?* Routledge.
7. Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review, 90*, (1), 5–43. https://doi.org/10.2307/2184371
8. Greco, J. (2010). *Achieving knowledge: a virtue-theoretic account of epistemic normativity.* Cambridge University Press. https://doi.org/10.1017/CBO9780511806902
9. Coady, C. A. J. (1992). *Testimony: a philosophical study.* Clarendon Press.
10. Fricker, M. (2007). *Epistemic injustice: power and the ethics of knowing.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001
11. Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review, 60*, (1), 20–43. https://doi.org/10.2307/2181906
12. BonJour, L. (1985). *The structure of empirical knowledge.* Harvard University Press.
13. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*, (12), 1–38. https://doi.org/10.1145/3571730
14. Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems, 36*, 74952–74965. https://doi.org/10.48550/arXiv.2305.04388
15. Fodor, J. A. (1975). *The language of thought.* Thomas Y. Crowell.
16. Chalmers, D. J. (1996). *The conscious mind: in search of a fundamental theory.* Oxford University Press.
17. Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*, (3), 417–424. https://doi.org/10.1017/S0140525X00005756
18. Dreyfus, H. L. (1992). *What computers still can't do: a critique of artificial reason.* MIT Press.
19. Goldman, A. I. (1986). *Epistemology and cognition.* Harvard University Press.
20. Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell.
21. Harman, G. (1986). *Change in view: principles of reasoning.* MIT Press.
22. Aristotle. (1998). *Nicomachean ethics* (W. D. Ross, Trans.). Oxford University Press.
23. Crawford, M. B. (2015). *The world beyond your head: on becoming an individual in an age of distraction.* Farrar, Straus and Giroux.
24. Dennett, D. C. (1987). *The intentional stance.* MIT Press.

Информация об авторе / Information about the author:

Александр Михайлович Жаров — младший научный сотрудник, Институт философии Российской академии наук, Москва, Россия.

Alexander M. Zharov — Junior Researcher, Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia.

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