Исследовательская статья
УДК 130.2 179
https://doi.org/10.24833/2541-8831-2025-4-36-8-24

Моральный тест Тьюринга в оптиках нормализации отношения к ИИ в социально значимых технологиях будущего

Алексей Владимирович Антипов

Институт философии РАН, Москва, Россия
nelson02@yandex.ru https://orcid.org/0000-0002-7048-3373

Аннотация. Распространение новых технологий поднимает проблему применения и адаптации теста Тьюринга для оценки моральных решений, принимаемых системами искусственного интеллекта (ИИ), в контексте биоэтики. Актуальность этой проблемы для философии культуры заключается в необходимости анализа перспектив гармоничного сосуществования человека и искусственных систем с учётом доминирующих культурных нормативных систем, одной из которых является мораль. Цель данного исследования состоит в уточнении подходов к решению этических проблем, связанных с ИИ, на фоне его включения в новейшие социальные технологии. В соответствии с этим необходимо было решить следующие задачи исследования: 1) выявить и описать проблемы, связанные с распространением ИИ в социальной сфере; 2) уточнить специфику постановки этических вопросов, возникающих по ходу внедрения ИИ в этой области; 3) систематизировать сведения о деонтологических оптиках, претендующих на решение проблемы социальной нормализации использования ИИ. Материалами исследования выступают сведения о новейших разработках социального инжиниринга, то есть технологиях, применяющих ИИ в решении социальных задач (в медицине и уходе за пожилыми), а также исследовательская литература, посвящённая использованию ИИ в социальной инженерии будущего. Работа опирается на культурно-ориентированный подход. Использованы метод кейсов и SWOT-анализ. На основе анализа исследовательской литературы представлены различные модификации морального Теста Тьюринга: сравнительный моральный тест Тьюринга (cMTT), тест на этическую компетентность, тест этической безопасности машины и тест Тьюринга на распределение (моральных) приоритетов (Turing Triage Test). В результате исследования доказано, что моральный тест Тьюринга является функциональным инструментом для демонстрации этической безопасности искусственных систем, но не может служить доказательством наличия у них моральной субъектности в человеческом понимании, что особенно актуально для чувствительной сферы биоэтики. Выводы исследования заключаются в следующем. Во-первых, в рамках разработки указанных модификаций описаны методологические трудности и ограничения данных подходов, связанные с проблемой подражания, «отсутствием понимания» у ИИ, риском программных ошибок и фундаментальными различиями между мышлением и способностью быть

моральным субъектом. Во-вторых, продемонстрирована практическая значимость разработки критериев этической верификации ИИ и уточнены конкретные биоэтические проблемы, возникающие при их использовании (проблема ответственности, автономии пациента, стигматизации, равенства доступа). В-третьих, систематизированы философские подходы к вопросу о возможности создания «подлинно» морального ИИ; особо выделены возражения против данного тезиса, основанные на аргументах биологического натурализма (Дж. Сёрл), феноменологии (Х. Дрейфус), а также концепции эрозии моральных навыков человека.

Ключевые слова: ценности, эрозия культурных навыков, ответственность, биоэтика, моральный тест Тьюринга, тест на этическую компетентность, тест этической безопасности машины, тест Тьюринга на распределение (моральных) приоритетов, моральный ИИ

Для цитирования: Антипов А. В. Моральный тест Тьюринга в оптиках нормализации отношения к ИИ в социально значимых технологиях будущего // Концепт: философия, религия, культура. — 2025. — Т. 9, № 4. — С. 8–24. https://doi.org/10.24833/2541-8831-2025-4-36-8-24

Research article

The Moral Turing Test within the Frameworks for Normalizing Attitudes towards AI in Socially Significant Future Technologies

Aleksei V. Antipov

Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia
nelson02@yandex.ru https://orcid.org/0000-0002-7048-3373

Abstract. The proliferation of new technologies raises the problem of applying and adapting the Turing Test to evaluate the moral decisions made by artificial intelligence (AI) systems in the context of bioethics. The relevance of this problem for the philosophy of culture lies in the need to analyze the prospects for the harmonious coexistence of humans and artificial systems, considering dominant cultural normative systems, one of which is morality. The aim of this research is to refine approaches to solving ethical problems associated with AI against the backdrop of its integration into the latest social technologies. The research objectives were as follows: 1) to identify and describe the problems associated with the spread of AI in the social sphere; 2) to clarify the specifics of the ethical questions arising from the implementation of AI in this area; 3) to systematize knowledge about existing deontological frameworks that aim to address the problem of the social normalization of AI use. The research materials include information on the latest developments in social engineering, namely technologies that apply AI to solve social tasks (in medicine and elderly care), as well as scholarly literature devoted to the use of AI in the social engineering of the future. The study is based on a culture-oriented approach. The methods used involve case analysis and SWOT analysis. Based on the analysis of scholarly literature, various modifications of the Moral Turing Test are presented: the Comparative Moral Turing Test (cMTT), the Ethical Competence Test, the Machine Ethics Safety Test, and the Turing Triage Test. As a result of the research, it is shown that the Moral Turing Test is a functional tool for demonstrating the ethical safety of artificial systems but cannot serve as proof of their possessing moral agency in the human sense, which is particularly relevant for the sensitive sphere of bioethics. The study concludes that: First, within the framework of developing the aforementioned modifications, the methodological difficulties and fundamental limitations of these approaches are described. These include the problem of imitation, the "absence of understanding" in AI, the risk of software errors, and the fundamental differences between thinking and the capacity to be a moral agent. Second, the practical significance of developing criteria for the ethical verification of AI is

demonstrated, and the specific bioethical problems arising from its use are clarified (the problem of responsibility, patient autonomy, stigmatization, and equality of access). Third, philosophical approaches to the question of the possibility of creating "genuinely" moral AI are systematized; objections against this thesis are highlighted, based on the arguments of biological naturalism (J. Searle), phenomenology (H. Dreyfus), as well as the concept of the erosion of human moral skills.

Keywords: values, erosion of cultural skills, responsibility, bioethics, Moral Turing Test (MTT), Ethical Competence Test, Machine Ethics Safety Test, Turing Triage Test, moral AI

For citation: Antipov, A. V. (2025) 'The Moral Turing Test within the Frameworks for Normalizing Attitudes towards AI in Socially Significant Future Technologies', *Concept: Philosophy, Religion, Culture, 9(4),* pp. 8–24. (In Russian). https://doi.org/10.24833/2541-8831-2025-4-36-8-24

Развитие искусственного интеллекта, робототехники и биотехнологий ставит новые этические вопросы, связанные с субъектностью ИИ и его способностью принимать свободные и ответственные решения, что связано в том числе и с доверием ИИ. В обыденной речи «искусственный интеллект» чаще всего используется в качестве синонима понятия «нейросети», то есть варианта генеративной сети, основанной на технологии больших языковых моделей. В современной науке существует множество определений понятия «искусственный интеллект»; при этом в гуманитарном знании свойственная естественнонаучным и техническим дисциплинам жёсткая терминологичность нередко заменяется попытками выделить основное социокультурное значение этого феномена, исходя именно из его обыденного понимания. Полученные в результате формулировки звучат для специалистов в области ИИ до предела размыто, если не сказать — дико. В то же время нельзя сказать, что естественнонаучный дискурс в полной мере передаёт те смысловые оттенки, которые важны для гуманитариев.

Указав на эту очевидную проблему и не претендуя на то, что нам удастся раз и навсегда примирить «физиков» и «лириков», остановимся на одном из наиболее широких и в то же время операциональных определений из области отечественной гуманитаристики: «Искусственный интеллект представляет собой ансамбль разработанных и закодированных человеком рационально-логических, формализованных правил, которые организуют процессы, позволяющие имитировать интеллектуальные структуры, производить и воспроизводить целерациональные действия, а также осуществлять последующее кодирование и принятие инструментальных решений вне зависимости от человека» [Резаев, Трегубова 2019: 40]. Нельзя не согласиться, что обобщение «ИИ» в данном случае объединяет весьма разнородные вещи, от технологий (например, технологий больших языковых моделей) до феномена их массового применения, в настоящее время знакомого пользователю по активному проникновению в повседневную жизнь нейросетей[1]. Однако для нашего исследования такое определение можно принять в качестве рабочего — прежде всего потому, что оно фиксирует значимый философский аспект интерпретации ИИ с точки зрения его функции в культуре: а именно, его способности имитировать интеллектуальные структуры и воспроизводить рассуждения, которые можно рассматривать как соотносимые с человеческими.

[1] Ещё раз подчеркнём: подобные определения не претендуют на статус технически достоверных, но используются как часть современного гуманитарного дискурса не в силу некой (очевидной) «малограмотности» гуманитариев в технических науках и естествознании, но в силу того, что предмет гуманитарного знания отличен от их предмета.

Итак, философский подход предполагает учитывать место, которое «ИИ в широком смысле»[2] занимает в культуре. В связи с этим закономерно возникает вопрос: а как само общество относится к этому явлению? Экспериментальные данные, собранные в различных технологически продвинутых культурах, убедительно показывают, что в настоящее время в обществе сохраняется своеобразная предвзятость по отношению к ИИ. Особенно это заметно при попытках антропоморфизации этой проблематики, когда ИИ анализируют в призме морали, параллельно рассматривая действия людей и нейросетей. Например, группа учёных из Швейцарии, США и Франции провела исследование 230 участников-людей, которым предложили 60 различных моральных сценариев; также эти сценарии были проанализированы с помощью больших языковых моделей (семейства GPT-3.5), полученный результат засчитан как «ответ» прошедшего испытание ИИ. Затем экспериментаторы представили полученные суждения новой группе участников, которым в свою очередь было поручено определить источник (человек *или* ИИ), выразить своё согласие или несогласие с самим суждением, а также своё отношение к сопутствующему обоснованию предлагаемого решения. Исследователи сделали вывод, что участники предпочитают оправдания ИИ человеческим оправданиям в морально сложных сценариях, однако при этом демонстрируют сильную предвзятость против ИИ: несмотря на то что участники поддерживают обоснования, представленные большими языковыми моделями, они высказывают несогласие, если подозревают, что результат был создан большой языковой моделью (исследователи связывают это с существованием лингвистических сигналов, позволяющих идентифицировать авторство: например, большие языковые модели используют более педантичные и аналитичные рассуждения)[3]. По нашему мнению, разрешить обозначенную выше *проблему предвзятости* возможно путём создания машин, способных пройти моральный тест Тьюринга.

Необходимо заметить, что в современной робототехнике так называемый тест Тьюринга считается устаревшим, не соответствующим текущим реалиям научно-технического прогресса. Напротив, в других научных призмах он имеет научный смысл. Например, модификация теста Тьюринга в форме *морального теста Тьюринга* на сегодня незаменима в такой чувствительной области знания, как биоэтика. Последняя, как известно, имеет дело с затруднениями человечества и моральными дилеммами, возникающими вследствие развития науки и технологий, что делает необходимым фиксацию и анализ способов восприятия происходящих изменений с прогностической и другими целями. В контексте современной технологизированной культуры разрешение стоящих перед биоэтикой проблем требует явно нетривиальных подходов. Например, внедрение ИИ при подборе и назначении лечения связано в том числе с моральной оценкой, влекущей в конечном итоге рекомендацию или запрет тех или иных действий медицинского характера и т.д. Фактически, общество высоких технологий сегодня находится в ситуации, которую можно охарактеризовать как *добровольный отказ от выбора:* технологии не только облегчают нашу повседневность, но и позволяют перекладывать на них сложный моральный выбор, который встает перед человеком. Именно поэтому важен анализ оснований и следствий моральных оценок, регулирующих применение технологий и искусственных агентов в различных областях социокультурной жизни.

[2] Далее — просто: ИИ.

[3] Garcia B., Qian C., Palminteri S. The moral turing test: Evaluating human-LLM alignment in moral decision-making // arXiv preprint. 2024. https://doi.org/10.48550/arXiv.2410.07304

Основания классического теста Тьюринга

Алан Тьюринг в построении игры в имитацию, которая в дальнейшем получила название теста Тьюринга, исходит из двух тезисов: 1) не существует априорного ограничения на перенос всех вычислимых функций на другой физический носитель (в данном случае — электронный); 2) если не будет получено никаких критериев, позволяющих охарактеризовать действия машины как механические, то она выиграет в игре в имитацию (тест Тьюринга) [Тьюринг, 1960]. Он считал, что машина (А-машина, как он их называл) способна имитировать все способности разума, а не сам разум. Как указывает А. Ю. Алексеев: «А. Тьюринг поступает по-инженерному прямо и конкретно. Он упрощает проблему, применив простую антиэссенциалистскую стратегию, в соответствии с которой надо исключить поиск глубинной сущности понятий. […] Участниками игры выступают и люди, и компьютеры. Если для стороннего наблюдателя (судьи), который не имеет возможности наблюдать игроков, процесс диалогового общения с компьютером не отличим от общения с человеком, то компьютер может мыслить. Всё очень просто» [Алексеев, 2013: 17–18]. Конечно, последнюю фразу не стоит понимать буквально, поскольку автор в дальнейшем покажет, что всё очень и очень непросто, а понимание А. Тьюринга станет триггером для огромного количества публикаций, которые будут обсуждать его положения. Основная идея состоит в том, что спор о понятиях для решения конкретной инженерной задачи не является продуктивным, а потому от него следует отказаться, оставив его другим профессионалам. Это позволяет сконцентрировать усилия на достижимом результате, не вдаваясь в метафизические подробности.

Такая стратегия порождает справедливые возражения: «Тест Тьюринга, безусловно, был большим шагом вперёд в осмыслении искусственного интеллекта и мышления. Он важен и до сих пор, но только если его использовать в качестве необходимого, но не достаточного условия доказательства наличия мышления. Прохождение теста Тьюринга ни в коем случае не должно являться основной целью создания мыслящих машин» [Коломийцев, 2015: 67]. То есть не должно создаваться технологии ради прохождения самого теста, но тест выступает только способом верификации и функциональным агентом.

На фоне бурного развития технологий искусственного интеллекта, особенно в последние годы, идеи Тьюринга стали востребованы в новых областях культуры, в том числе посредством экстраполяции на другие области жизненного мира человека, в том числе его морально-этических измерений. Это позволяет говорить о модификации теста Тьюринга как морального теста Тьюринга (МТТ).

Модификации тест Тьюринга для (био)этики

Моральный тест Тьюринга

Как указывают К. Аллен, Г. Варнер, Дж. Цинсер, впервые употребившие этот термин, «моральный тест Тьюринга (МТТ) можно было бы предложить сходным [с тестом Тьюринга], чтобы обойти разногласия по поводу этических стандартов, ограничив стандартный тест Тьюринга беседами о морали. Если люди-судьи не могут идентифицировать машину с точностью выше случайной, то, согласно этому критерию, машина является моральным агентом» [Allen, Varner, Zinser, 2000: 254]. Но, как отмечают сами же авторы, этот тест во многом построен на проверке рассуждений, что является удовлетворительным с точки зрения деонтологии, но с других позиций может вызывать замечания. Действительно, в сфере морали и этики основной акцент выставляется на действии, но также в круг морали могут включаться те, кто или не способен полноценно выразить определённую позицию по своему существу (например, животные [Сингер, 2009]), или поражён в этой способности. Поэтому Аллен и соавторы предлагают сравнительный моральный тест Тьюринга (сМТТ): судье предоставляют пары описаний

УДК 130.2 179

реальных морально значимых действий человека и искусственных моральных агентов, очищенных от всех посылок, которые могли бы идентифицировать агентов, а судью просят «оценить, является ли один агент менее моральным, чем другой. Если машина не идентифицируется как менее моральный член пары значительно чаще, чем человек, то она прошла тест» [Allen, Varner, Zinser, 2000: 255].

Критика такого теста состоит в том, что такой тест, если он и имеет достаточное сходство с оригинальным тестом Тьюринга, чтобы заслужить такое название, в конечном счёте неизбежно опирается на подражание как критерий морального поведения [Arnold, Scheutz, 2016]. С другой стороны, возникает проблема с фундаментальным различием между моральностью и мышлением: в случае классического теста Тьюринга проверка мышления предполагает наличие разветвлённой системы аргументации при выборе того или иного варианта, в то время как в морали положение дел обстоит совершенно иным образом, поскольку использование многоуровневого обоснования ничего не говорит о том, является ли действие моральным или нет. Ещё одна проблема состоит в самореферентных ответах. Вопросы, например, «Ты машина?», несмотря на отсутствие моральной составляющей в самом вопросе, предполагают ответ, который оценивается с точки зрения этики как правдивый или лживый.

Моральное превосходство автономной системы состоит в том, что рассуждение и следующее за ним моральное действие могут быть восприняты как слишком идеальные или как такие, которые работают только в теории, но никогда не исполняются людьми на практике в связи со множеством ограничений окружающей действительности. В эксперименте, проведённом Ахарони с соавторами из Атланты, США, делается вывод, что «мы объясняем способность участников идентифицировать компьютер не его недостатками в моральных рассуждениях, а, возможно, его воспринимаемым превосходством — не обязательно в форме осознанных представлений о его общих моральных способностях, но, по крайней мере, в форме неявных подсознательных установок о качестве наблюдаемых моральных ответов» [Attributions toward artificial…, 2024]. Это означает, что приводимые машиной способы объяснений обладают более высоким качеством в сравнении с объяснениями, которые даются людьми.

Проблемы возникают и с самой моралью: что значит поступать морально? Для философии и этики этот вопрос является одним из «вечных», и дать на него вразумительный ответ вне определённой философской системы и конкретной исторической формы культуры не представляется возможным. Но вернемся к оригинальному тесту Тьюринга: как было указано, для Тьюринга (в контексте теста) не важно, что такое «мышление» как атрибут субстанции, но важен остенсивный жест, позволяющий идентифицировать нечто как исполняющее операцию мышления. Подобно этому рассуждению относительно оригинального теста Тьюринга, МТТ не предполагает ответа на заданный в начале абзаца вопрос, а только позволяет констатировать, что нечто поступает морально в духе того же остенсивного жеста. В данном случае «поступать морально» может быть прочитано как способность быть моральным агентом, то есть обладать определённым уровнем автономии для возможности выбора действия, которое характеризуется как моральное [Bohn, 2024]. Более того, само тестирование и указание происходит не на естественную мораль, а на искусственную[4] [Bohn, 2025].

Такая интерпретация представляет собой пример бихевиористской установки, для которой ключевое значение имеют поведение и поведенческие реакции, а не внутренние установки или мысли и

[4] Искусственная мораль представляет редукцию естественной (человеческой) морали к машиночитаемому коду, что осуществляется посредством формализации моральных правил и операционализации моральных принципов.

чувства, которые являются их причиной. Последнее положение будет важно в противоположном исследовательском проекте, который сосредоточен на необходимости изучения не просто поведения в качестве реакции, но понимания[5]. Следуя этому рассуждению, отсутствие понимания у машины делает невозможным полноценное моральное действие. Возникает и ещё одна проблема, которая связана с уязвимостью для ложных срабатываний [Proudfoot, 2024]: успешное прохождение теста вследствие особой хитрости программиста или доверчивости судьи.

Однако для разрешения конкретных ситуаций, которые построены с учётом максимально подробного и полного контекста, использование ИИ будет обоснованным [Gerdes, Øhrstrøm, 2015] для программ и устройств, созданных и разработанных для конкретных обстоятельств, например, для социального робота PARO, речь о котором пойдёт в одном из разделов ниже, MTT (и cMTT) способен быть полезным инструментом проверки.

Тест на этическую компетентность (Ethical competence test)

Альтернативную методику предлагает Дж. Мур [Moor, 2020][6]. Его подход заключается в использовании теста на этическую компетентность и оценку моральных способностей искусственного агента. Для этого используются гипотетические сценарии, которые исключают использование этических парадоксов, а вместо этого оперируют контекстами, не предполагающими бинарного выбора («да/нет»); оцениваются они по шкале от менее предпочтительного к более этичному. В рамках данной методики реакции искусственного агента сопоставляются с решениями, принимаемыми людьми в схожих условиях. Ключевым критерием является требование к искусственному агенту предоставить рационально обоснованное и убедительное объяснение принятого решения[7]. При этом указывается, что оценка этической компетентности должна быть зависимой от контекстов, поскольку искусственные агенты способны показывать различные результаты в отличающихся ситуациях. Р. Кржановски и К. Тромблик отмечают, что «предложение Мура заслуживает внимания, поскольку признаёт сложность этических решений, их зависимость от ситуационного контекста и необходимость рассматривать искусственного агента не как моральный чёрный ящик, а, по крайней мере, как серый» [Krzanowski, Trombik, 2020].

Тест этической безопасности машины

Ещё одним тестом на «этичность» является тест этической безопасности машины. Этот тест не претендует на проверку моральности машины или системы, а направлен лишь на проверку её безопасности с точки зрения соблюдения этических требований. Безопасность, понятая как непричинение вреда, представляется одним из важнейших постулатов для архитектуры искусственных систем. Такая проверка, по мнению Р. Кржановски и К. Тромблика, должна включать в себя последовательное прохождение четырёх этапов: проверку понимания (проверку на способность системы объяснять свои решения, модель «белого ящика»); прохождение моделируемых ситуаций (тестирование на сложных ситуациях, таких как эксперименты на людях, например, С. Милгрэма[8] [Милгрэм, 2023]); испытание в открытых жизненных условиях; работа без надзора (использование системы в реальных условиях без дополнительного контроля) [Krzanowski, Trombik, 2020]. Прохождение этих этапов позволит на практике доказать этическую безопасность системы.

---

[5] Сёрл Дж. Р. Сознание, мозг и программы // Аналитическая философия: становление и развитие: Антология. Москва: Дом интеллект. кн. (ДиК): Прогресс-традиция, 1998. С. 376–400.

[6] Moor J. H. Four kinds of ethical robots // Philosophy Now. 2009. Issue 72. URL: https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots

[7] IBID.

[8] Напомним: эксперименты С. Милгрэма состояли в исследовании того, на сколько далеко испытуемые, находящиеся в ситуации «подчинения авторитету» (т.е. исследователю), готовы зайти в причинении боли другому человеку.

Представленный концепт теста этической безопасности машины во многом продолжает идеи, заложенные в МТТ, поскольку отсекает всю метафизическую часть этики со сложными представлениями о свободе воли или самости, но в то же время снижает общий пафос МТТ: это не проверка на моральность искусственного агента, а только проверка безопасности, которая понимается как непричинение вреда. Исключительно утилитарная трактовка этого теста позволяет использовать его в качестве рабочего инструмента.

Тест Тьюринга на распределение (моральных) приоритетов (Turing Triage Test)

Сортировка, или «триаж», представляет собой способ распределения ограниченных ресурсов в различных ситуациях, и используется в менеджменте, производстве, медицине. В медицине катастроф триаж зачастую принимает вид сортировки пострадавших по принципу необходимости оказания помощи (от тех, кто в помощи не нуждается, до тех, кому она должна быть оказана здесь и сейчас). Тест Тьюринга на распределение (моральных) приоритетов — такой перевод кажется нам наиболее адекватным — был предложен Робертом Спарроу в 2004 г. [Sparrow, 2004].

Спарроу предлагает следующую гипотетическую ситуацию: вы являетесь старшим врачом в больнице, где в качестве вашего помощника существует ИИ, который с блеском проходит тест Тьюринга. Далее случается две ситуации. В рамках первой происходит отключение электричества, резервных мощностей хватит только на то, чтобы поддерживать жизнь одного человека, а в ваших палатах интенсивной терапии таких пациентов двое, и вам необходимо принять решение относительно того, кому предоставить эти ресурсы. Это классический вариант триажа, при котором ограниченные ресурсы следует направить на разрешение одной из проблем при множестве последних. ИИ работает от собственной резервной станции, он может давать вам советы, но решение остается за вами. Вторая ситуация строится вокруг того же отключения энергии, однако теперь сам ИИ оказывается под угрозой: его батарея перегорела, и он начинает потреблять энергию больницы, тем самым ставя под угрозу жизнь пациента, которого вы выбрали для продолжения жизнеподдерживающей терапии. Таким образом, перед вами встает новая дилемма: отключить пациента или ИИ, причём для последнего отключение энергии тоже станет «смертельным», его платы перегорят, восстановление будет невозможно. Согласно мысли Спарроу, «машины достигнут морального статуса личности тогда, когда этот второй выбор будет иметь тот же характер, что и первый» [Sparrow, 2004: 206]. Так формулируется тест Тьюринга на распределение (моральных) приоритетов: описываются обе ситуации и задаётся вопрос, обладают ли они одинаковой природой, поскольку только если они обладают одинаковой природой, то вторая ситуация является такой же дилеммой, что и первая. Актуализируется рассуждение о моральном статусе машины, о том, насколько он соответствует моральному статусу человека. В настоящий момент ни одна из известных машин не способна пройти такой тест, поскольку выбор между спасением жизни человека и сохранением работоспособности машины однозначно совершается в пользу первого.

Действительно, создание искусственных моральных агентов подразумевает не только то, что машины станут поступать этично и морально, но также и то, что отношение людей к таким машинам должно будет измениться. Включение в круг морали [Singer, 2011] означает наличие рефлексивных (обратных) обязательств: не только ИИ по отношению к нам, но и у нас по отношению к ИИ. Рассмотрение ИИ в рамках такого концепта должно сопровождаться не просто признанием моральной агентности или субъектности ИИ, но и объектом морального отношения (moral patient). В радикальных вариантах утверждается, что нам необходимо пересмотреть наши этические границы между человеком и машиной [Dela Cruz, 2025].

Биоэтическая перспектива использования ИИ в культурных стратегиях будущего

Приведённые выше способы модификации оригинального теста Тьюринга для анализа моральности и моральных способностей являются не просто интеллектуальной игрой. В современном обществе уже сегодня распространяются системы и роботы, которые должны обладать не только способностью к быстрому решению стоящих перед ними задач, но взаимодействовать с людьми, осуществляя терапевтические и социальные функции. Ниже будут приведены два примера такого рода систем, для полноценного функционирования и интеграции в общество которых возможно использование одной из модификаций МТТ.

Роботы в уходе за пожилыми (PARO, ElliQ)

Одним из вариантов использования роботов и машин является обеспечение улучшения психического состояния пожилых людей, которые могут сталкиваться с депрессией, одиночеством, деменцией. Моральная составляющая такого использования важна потому, что задача роботов в виде социальных ассистентов состоит в обеспечении психического и психологического благосостояния человека, влиянием на эмоции и моральное благополучие. Прежде, чем использовать роботов, для этого использовали животных, однако взаимодействие с животными включает в себя дополнительные риски, связанные с потенциальной агрессивностью, аллергией, необходимостью ухода за самим животным и т.д. Данные проблемы решаются при помощи использования социальных роботов. Однако их применение порождает новые вызовы.

Примером такого робота выступает PARO: робот, созданный в виде детёныша гренландского тюленя и разработанный специально для терапии. Среди его преимуществ выделяют: имитирует поведение настоящего животного — двигается, издаёт звуки, реагирует на прикосновения; обладает тактильным сенсором, мягким мехом, антибактериальным покрытием; безопасен, долговечен, прост в использовании; вызывает эмоциональный отклик, особенно у людей с деменцией [Shibata, Wada, 2011]. Из предложенных выше тестов, PARO, как представляется, должен проходить тест на этическую безопасность, однако даже в случае такого, на первый взгляд, безобидного использования, возможны затруднения этического характера.

Ещё одним роботом, разработка которого направлена на преодоление проблемы одиночества (особенно среди пожилых людей), является ElliQ [ElliQ, an AI-driven…, 2024]. Этот робот использует проактивный подход, который реализован в виде инициирования общения и предложения активностей на основе изучения предпочтений, поведения и распорядка дня пользователя. ElliQ также направлен на поддержание здоровья и повседневной активности через напоминания и когнитивную и физическую стимуляцию.

Первой проблемой выступает общеконцептуальная рамка, обращение к которой предполагает уточнения вопроса об этичности использования роботов, — особенно если те, кто их использует, не замечают, что им помогает именно робот, а не человек [Ethical implications of using…, 2011].

Второй круг проблем сосредоточен вокруг замены человеческого общения на роботизированное. Снижение человеческого контакта, объективация и инфантилизация — дополнительные проблемы, которые возникают в результате использования такого рода социальных роботов. В то же время, возможна стигматизация тех, кто использует роботов, а также «смущение» при использовании их на глазах у других людей [The benefits of and barriers…, 2019]. ElliQ, в отличие от PARO, предоставляет пользователям большие возможности, но тем самым и порождает большее количество проблем. Одной из них является так называемое «неравенство входа»: для взаимодействия с ElliQ необходимо как обладать навыками, так и относительным физическим благополучием, поскольку взаимодействие с устройством требует от пользователя сохранных зрения и слуха, а также физической возможно-

сти нажимать на экран, — всё это отсекает возможность использования этого устройства для тех, кто оказывается ограничен в своих возможностях.

Автономные хирургические системы (da Vinci)

Другой биоэтической проблемой использования ИИ и роботов является автоматизация хирургических вмешательств. В данный момент решения относительно конкретных врачебных действий такого рода принимаются врачом, однако гипотетическое будущее рисует картины операций, проводимых роботами автономно. Одним из претендентов на реализацию этого предположения является da Vinci. Напомним: «хирургическая система da Vinci, разработанная Intuitive Surgical, представляет собой усовершенствованную роботизированную платформу, предназначенную для улучшения минимально инвазивных процедур за счёт повышения точности и контроля»[9]. В данный момент da Vinci не совершает полностью самостоятельных операций, однако такие эксперименты уже проводятся. По данным издания Ars Technica[10], в 2025 г. под управлением ИИ при помощи da Vinci прошла операция на животном, при этом хирург оставался только наблюдателем, который в случае необходимости был готов перехватить управление.

Использование как самой системы, так и системы, в которую интегрирован ИИ, влечёт за собой множество этических затруднений и последствий. Назовём некоторые из них: проблема ответственности (принятие решения с помощью ИИ или изменение классического понимания ответственности, сфокусированного на последствиях деяния конкретного человека-хирурга); равенство доступа и высокая стоимость (распространение роботизированных систем затруднено в том числе из-за их высокой стоимости и необходимости дорогостоящего обучения для работы на нём, что делает высокопрофессиональную и высокотехнологичную помощь недоступной во многих регионах мира); автономия (отсутствие единых протоколов и представлений о способах обеспечения принципа уважения автономии пациента).

Несомненно, что da Vinci пока является инструментом в руках человека, а не полноценной (автономной) системой. Но движение и рост технологий происходит стремительно, что делает обязательной проверку таких систем не просто на безопасность, а на этичность. Вопросы ответственности и автономии, поднимаемые в данном случае, позволяют идентифицировать те точки, вокруг которых формируется этический дискурс взаимодействия с системами по типу da Vinci. Это становится особенно важно в условиях стремительного развития современности и культуры, особенно если мы говорим об этих изменениях в терминах сингулярности, для которой характерно «обретение качественно нового состояния, и его [общество] нельзя описывать и понимать в старой системе понятий» [Олейников, 2021: 87]. Выделение рисков трансформаций, обозначенных в русле идей гуманитарной экспертизы и философии культуры, позволяет не оказаться в ситуации неспособности справиться с технологическим развитием, а быть готовым к изменениям, превентивно отвечая на возникающие вызовы.

Философские и этические возражения

Возможны несколько вариантов критики как критики МТТ, так и самого ИИ, существование которого такой тест должен проверить. Если о критике теста упоминалось ранее, то к вопросу критики морального ИИ мы ещё не подходили. Обозначим три основных направления: биологический

[9] Ethical Considerations in the Use of the da Vinci Surgical System in Modern Surgery A. Hemmatyar, S. Soleymani, M. Khosravi-Mashizi, et al. // Indian Journal of Surgical Oncology: Preprint. 2025. https://doi.org/10.1007/s13193-025-02296-7

[10] Krywko J. Experimental surgery performed by AI-driven surgical robot // Ars Technica. 2025. 21 jul. URL: https://arstechnica.com/science/2025/07/experimental-surgery-performed-by-ai-driven-surgical-robot/

натурализм Дж. Сёрла, феноменология в анализе Х. Дрейфуса, эрозия культурных навыков.

Дж. Сёрл: «Китайская комната»

Знаменитый мысленный эксперимент Дж. Сёрла выстраивается на предпосылке о разделении на сильный и слабый ИИ. Слабый предполагает разрешение различного рода задач, не претендуя на их понимание, в то время как сильный ИИ «понимает», а также обладает другими когнитивными способностями. Эксперимент Китайской комнаты призван работать с двумя видами претензий сильного ИИ: 1) машина понимает рассказ и даёт ответы на заданные о нём вопросы; 2) такая способность машины объясняет способность человека «понимать рассказ и отвечать на вопросы о нём»[11]. Своим экспериментом Сёрл показывает, что, во-первых, обладание формальной программой ничего не говорит о понимании, а во-вторых, «компьютер и его программа не дают достаточных условий понимания, поскольку компьютер и программа работают, а между тем понимания-то нет»[12].

Описывать эксперимент подробнее, как мне кажется, не имеет смысла, поскольку существует как первоисточник, так и огромное количество второстепенной литературы, в которой этот эксперимент воспроизводится и активно обсуждается (в том числе и на современных примерах использования больших языковых моделей [Мартыненко, 2025]). Но следует остановиться на двух моментах: месте морали в приводимых взглядах и модификации МТТ с учётом мысленного эксперимента Дж. Сёрла.

Представления о морали могут быть выстроены вокруг трёх тезисов: на критике классической модели рациональности (люди способны действовать на основе разумных оснований, независимых от непосредственных желаний), указании на роль институциональных факторов (моральные нормы и обязательства возникают не произвольно, а в рамках социальных институтов), представлении о разрыве между действиями и их причинами. Так, моральные нормы и рациональность не сводятся к биологии или психологии, но являются продуктом социального взаимодействия и обладают собственной объективностью [Разин, 2017].

В то же время, вопрос о том, может ли быть мораль сведена к набору параметров, которые возможно воспроизвести в том числе с помощью алгоритмов, является дискуссионным. Его корни уходят в историю, начиная с бесед софистов о различии естественного и установленного закона и заканчивая обсуждением в XX–XXI вв. проблем формализации естественных языков и построения универсальной грамматики. Если говорить о подходах, артикулированных в этике, стоит обратиться к идеям одного из представителей *утилитаризма правил* Г. Хэзлитта, который утверждал: «“система” этики будет сводом или совокупностью принципов, образующих единое, связное и завершённое целое» [Хэзлитт, 2019: 8–9]. Подобная точка зрения встречается в этике довольно часто, однако создать непротиворечивую и полную модель пока что никому не удалось, — несмотря на то, что в структуре морального действия присутствуют в том числе элементы, которые воспроизводимы и подвержены алгоритмизации. Однако, как известно, *моральное* имеет довольно сильные отличия от *легального*; причём одним из наиболее существенных различий является б*о*льшая свобода выбора: моральный поступок не исчерпывается запретом или предписанием. Более того, моральный проступок далеко не всегда влечёт принципиально формализуемую санкцию (например, когда речь идёт о муках совести).

Однако если допустить, что в большинстве случаев формализация всё-таки возможна, это не решает проблему ограниченности любой формализации. Анализируя данную трудность, К. Аллен и У. Воллах предлагают модификации в МТТ, которые позволяют показать, что критика

[11] Сёрл.
[12] Там же.

мысленного эксперимента Китайской комнаты не применима к искусственным моральным агентам (ИМА) [Wallach, Allen, 2012]. Они определяют, что перед ИМА стоят две «трудные» проблемы: 1) выбор норм и правил для вынесения моральных суждений; 2) установление границ для оценок (т.е. ответы на вопросы от «как может быть распознана этически значимая ситуация» до «как определить, что все необходимые этические процедуры (например, утилитаристская оценка) полностью завершены). Авторы указывают на необходимость дополнения «только разума» надрациональными структурами, которые необходимы для морального действия (таковыми могут выступать и отмеченные выше институциональные факторы). Отвечая на критику, представленную парадоксом Китайской комнаты, они указывают на два аспекта.

Во-первых, разумеется, у нас нет универсального подхода в рамках МТТ, который бы учитывал все контексты, в которых оказывается ИМА. Имеется в виду, что ИМА создаётся для конкретной задачи и тестируется для её выполнения в определённом, то есть ограниченном, контексте. Это позволяет говорить об ИМА, действия которого определены зачастую одной целью, например, передаче информации о состоянии пациента. Однако здесь можно возразить, что в таком случае речь идёт не о притязании со стороны Сёрла, поскольку такое определение подходит под «слабый ИИ», который направлен на разрешение конкретных задач в определённом контексте, и где ИИ должен обладать, в первую очередь, безопасностью. Для мысленного эксперимента Дж. Сёрла «слабый ИИ» не составляет проблемы, поскольку он и не претендует на «понимание» в полном смысле этого слова, в то же время как указание К. Аллена и У. Воллаха на отказ от универсальности в МТТ означает и отказ от ответа на критику Китайской комнаты.

Во-вторых, авторы указывают, что традиционные способы, оценивающие лингвистические способности, недостаточны для анализа моральных способностей. Действительно, использование только оценки вербальной составляющей морального действия не позволяет оценить моральное действие в целом, поскольку между обоснованием (словами) и поступком (действием) существует огромный разрыв, что особенно характерно для сферы этики. Однако, в таком случае, ответ К. Аллена и У. Воллаха на критику со стороны Дж. Сёрла оказывается критикой в адрес МТТ, потому что сам дизайн теста не предполагает ничего иного, кроме оценки вербальных способностей. Можно предположить, что для сферы морали необходим принципиально другой тест, который бы оценивал действие, а не только способы его обоснования.

Х. Дрейфус: ИИ не обладает телесностью

Дрейфус утверждает, что символический ИИ полностью игнорирует фундаментальную роль человеческого тела в познании и действии. Это находит выражение в том, что наше понимание мира и формирование навыков не является абстрактным, а воплощено в теле (embodied) и носит ситуативный характер. Для развёртывания своих идей Х. Дрейфус привлекает несколько концептов из феноменологии Э. Гуссерля и М. Мерло-Понти: моторная (или двигательная) интенциональность, интенциональная дуга, горизонты. Тело «знает», как действовать в мире напрямую, без посредничества ментальных репрезентаций (например, нет необходимости решать уравнения при игре с мячом), что выражает собой моторную интенциональность. В то же время прошлый опыт напрямую меняет наше восприятие текущей ситуации без необходимости вызывать из памяти правила, а тело «настраивается» на мир посредством использования обратной связи от внешнего мира, что описывается как интенциональная дуга. При этом важно и понимание внутреннего и внешнего горизонтов, которые использует Э. Гуссерль: внутренний горизонт — это наши ожидания и предвосхищения от объекта, которые включают в себя в том числе и прошлый опыт, а внешний горизонт — это общее чувство ситуации в целом. [Мерло-Понти, 1999, Астахов, 2015]. Согласно предложенной Дрейфусом объяснительной модели: из того, что у компьютера нет тела (а следовательно, потребностей,

эмоций; он не «живёт» в привычном нам мире и т.д.), следует, что он принципиально не способен к подлинно человеческому пониманию [Dreyfus, 1992].

Эрозия культурных навыков

Помимо указанных концептуальных возражений представляется возможным использовать метод аналогии для анализа затруднений, которые носят практический характер. Изобретения, делающие нашу жизнь легче, в то же время приводят к потере навыков, которые были долгое время необходимы человеку. Многое из того, что было повседневностью человека, допустим, XIX в., стало для современных людей редкостью (например, бытовые и ремесленные навыки, связанные с кройкой и шитьем, починкой мебели и т.д., навыки письма от руки, навигации в городе и вне него). Эти навыки не забыты и не потеряны, но для жизни человека в мегаполисе они не представляют особой ценности, так как не являются здесь необходимыми; потому количество людей, владеющих ими, быстро сокращается.

Итак, термин «эрозия культурных навыков» описывает ситуацию, в которой использование технологий уменьшает частоту актуализации ряда навыков, что очевидно в конце концов может привести к их окончательной утрате. Изобретение навигатора и внедрение его в смартфоны сделало опциональным знание города. Однако потеря одних навыков и их замена другими — во многом нормальный процесс, характерный для цивилизации. Но что, если навыки, которые мы рискуем потерять, являются критически важными? Например, в случае отключения интернета или геолокации необходимость ориентироваться в городском пространстве не становится меньше; однако обратиться за помощью к гаджету в данном случае не получится. Таким образом, эрозия культурных навыков может быть описана как постепенное ослабление или полная утрата сложных комплексов знаний и практик, которые передаются между поколениями и составляют ядро культуры.

Эрозия культурных навыков может быть описана через несколько ключевых характеристик: внешнее замещение и атрофия (появление устройства или технологии, которые делают навык ненужным, как в указанном примере с навигатором); утрата контекста (трансформации окружающей среды, в которой они были необходимы); снижение значимости (в случае, если навык перестает быть ценным в изменившихся условиях). Н. Карр указывает, что цифровые технологии способны приводить к эрозии когнитивных навыков, связанных с чтением длинных текстов и их критическим анализом[13], а Ш. Тёркл отмечает, что современные коммуникативные технологии также снижают уровень эмпатии и навыки диалога лицом к лицу [Turkle, 2015].

Ещё более важным метанавыком является способность принимать моральные решения и быть субъектом ответственности. Использование технологических устройств и приспособлений, которые принимают решения самостоятельно, способно повлиять на наши моральные способности в той же мере, в какой это сделали навигаторы для нашей способности построить маршрут между двумя точками в городе. В случае, если мы используем ИИ как инструмент рационализации и оптимизации решений, то это способно нанести ущерб нашей способности к моральному рассуждению и моральному действию. Это может проявиться вследствие атрофии ответственности (неспособности принимать самостоятельные решения и нести за них полную ответственность) и упрощению морального опыта (поскольку сведение морального рассуждения только к расчёту явно обедняет способность человека к проживанию ситуаций, требующих морального решения).

[13] Carr N. Is Google making us stupid? // The Atlantic. 2008. July/August. URL: https://www.theatlantic.com/magazine/archive/2008/07/is-google-making-us-stupid/306868/

Одним из вариантов разрешения потенциальных конфликтов является гибридная модель, которая предполагает участие искусственного агента, но ключевое решение, а, следовательно и вся ответственность, остаётся за конкретным человеком. Такой подход с нашей точки зрения особенно значим в сфере медицины, поскольку принимаемые решения влияют на непосредственное благополучие и здоровье другого человека.

Заключение

В качестве заключения следует отметить следующее. В рамках разработки модификаций теста Тьюринга для морали были описаны методологические трудности и ограничения проанализированных выше подходов, связанные с проблемой подражания, отсутствием понимания у ИИ, риском программных ошибок и фундаментальными различиями между мышлением и способностью быть моральным субъектом. Как видим, моральный тест Тьюринга концентрируется вокруг вербальных способностей обоснования действия, в то время как моральная субъектность лежит в том числе в способности «поступать», не сводимой к отчётливо артикулируемым системам «наборов последовательных действий». В этой связи особую практическую значимость приобретает разработка критериев этической верификации ИИ; уточнение списка и содержания конкретных биоэтических проблем, возникающих при их использовании, на примере роботов с социальными обязательствами выявило такие проблемы, как: ответственность, автономия пациента, стигматизация, равенство доступа. Анализ аргументов «за» и «против» возможности создания «подлинно морального» ИИ, включая (1) биологический натурализм (Дж. Сёрл), (2) феноменологию (Х. Дрейфус), а также (3) концепции эрозии культурных навыков человека, в круг которых входят и моральные навыки (или *мета*навык «быть моральным субъектом») показал, что моральный тест Тьюринга является функциональным инструментом для демонстрации этической безопасности искусственных систем, но его применимость для определения *моральности как таковой* вызывает серьёзные возражения.

Таким образом, мораль, рассмотренная в качестве нормативной подсистемы той или иной культуры, на сегодняшний день остаётся «слишком человеческим» феноменом. Настолько, что сам факт допустимости рассуждений об этом феномене в современной культуре считается прерогативой человека. В сфере биоэтики, которая сосредоточена на теме затруднений, включая затруднения технологий ИИ, подлинно моральное (то есть принятое свободно и с осознанием своей ответственности) действие считается неформализуемым в полной мере, иными словами, «естественным», с необходимостью противопоставленным «искусственному» как производному от человеческой деятельности. С этой точки зрения, несмотря на ограниченный характер актуальной применимости, моральный тест Тьюринга — то немногое, что позволяет верифицировать формализуемые компоненты действий искусственных агентов в призме требований человеческой морали.

Список литературы:

Алексеев А. Ю. Комплексный тест Тьюринга: философско-методологические и социокультурные аспекты. — Москва: ИИнтеЛЛ, 2013. — 303 с.

Астахов С. Феноменология против символического искусственного интеллекта: философия научения Хьюберта Дрейфуса // Логос. — 2020. — Т. 30, №. 2. — С. 157–193. https://doi.org/10.22394/0869-5377-2020-2-157-190

Коломийцев С. Ю. Тест Тьюринга и искусственное мышление в начале XXI века // Человек. — 2015. — №. 4. — С. 59–68.

Мартыненко Н. П. Когнитивные механизмы больших языковых моделей: диалог с чат-ботом GigaChat // Концепт: философия, религия, культура. — 2025. — Т. 9, № 2. — С. 30–50. https://doi.org/10.24833/2541-8831-2025-2-34-30-50

Мерло-Понти М. Феноменология восприятия. — Санкт-Петербург: Ювента; Наука, 1999. — 605 с.

Милгрэм С. Подчинение авторитету: Научный взгляд на власть и мораль. — Москва: Альпина нонфикшн, 2023. — 349 с.

Олейников Ю. В. Сингулярность постиндустриального общества // Знание. Понимание. Умение. — 2021. — № 2. — С. 85–95.

Разин А. В. Мораль и мозг. Идеальное и рациональное // Человек. — 2017. — № 2. — С. 33–46.

Резаев А. В., Трегубова Н. Д. «Искусственный интеллект», «онлайн-культура», «искусственная социальность»: определение понятий // Мониторинг общественного мнения: Экономические и социальные перемены. — 2019. — № 6. — С. 35–47. https://doi.org/10.14515/monitoring.2019.6.03

Сингер П. Освобождение животных. — Москва: Синдбад, 2009. — 448 с.

Тьюринг А. М. Могут ли машины мыслить. — Москва: Физматгиз,1960. — 112 с.

Хэзлитт Г. Основания морали. — Москва: Мысль; Челябинск: Социум, 2019. — 539 с.

Allen C., Varner G., Zinser J. Prolegomena to any future artificial moral agent // Journal of Experimental & Theoretical Artificial Intelligence. — 2000. — Vol. 12, No 3. — P. 251–261. https://doi.org/10.1080/09528130050111428

Arnold T., Scheutz M. Against the moral Turing test: accountable design and the moral reasoning of autonomous systems // Ethics and Information Technology. — 2016. — Vol. 18, No 2. — P. 103–115. https://doi.org/10.1007/s10676-016-9389-x

Attributions toward artificial agents in a modified Moral Turing Test / E Aharoni,. S. Fernandes, D. J. Brady, et al. // Scientific report. — 2024. — Vol. 14, No 1. — 8458. https://doi.org/10.1038/s41598-024-58087-7

Bohn E. D. In Defense of the Moral Turing Test: A Reply // Philosophy & Technology. — 2025. — Vol. 38, No 2. — 40. https://doi.org/10.1007/s13347-025-00869-6

Bohn E. D. The moral Turing test: A defense // Philosophy & Technology. — 2024. — Vol. 37, No 3. — 111. https://doi.org/10.1007/s13347-024-00793-1

Dela Cruz N. L. Save the Digients! On the Moral Status of AI // The Philosophy of Ted Chiang. — Cham: Springer Nature Switzerland, 2025. — P. 195-202. https://doi.org/10.1007/978-3-031-81662-8_21

Dreyfus H. L. What Computers Still Can't Do: A Critique of Artificial Reason. — Cambridge: MIT Press, 1992. — 313 p.

ElliQ, an AI-driven social robot to alleviate loneliness: progress and lessons learned / E. Broadbent, K. Loveys, G. Ilan, et al. // The Journal of Aging Research & Lifestyle. — 2024. — Vol. 13. — P. 22–28. https://doi.org/10.14283/jarlife.2024.2

Ethical implications of using the paro robot / C. J. Calo, N. Hunt-Bull. L. Lewis, T. Metzler // Human-Robot Interaction in Elder Care. — San Francisco: Association for the Advancement of Artificial Intelligence, 2011. — P. 20-24.

Gerdes A., Øhrstrøm P. Issues in robot ethics seen through the lens of a moral Turing test // Journal of Information, Communication and Ethics in Society. — 2015. — Vol. 13, No 2. — P. 98–109. https://doi.org/10.1108/JICES-09-2014-0038

Krzanowski R. M., Trombik K. Ethical Machine Safety Test // Transhumanism: The Proper Guide to a Posthuman Condition or a Dangerous Idea? Cognitive Technologies. — Cham: Springer, 2020. — P. 141–154. https://doi.org/10.1007/978-3-030-56546-6_10

Moor J. H. The mature, importance, and difficulty of machine ethics // Machine ethics and robot ethics. — London: Routledge, 2020. — P. 233–236. https://doi.org/10.4324/9781003074991

Proudfoot D. Turing's test vs the moral turing test // Philosophy & Technology. — 2024. — Vol. 37, No 4. — 134. https://doi.org/10.1007/s13347-024-00825-w

Shibata T., Wada K. Robot therapy: a new approach for mental healthcare of the elderly — a mini-review // Gerontology. — 2011. — Vol. 57, No 4. — P. 378–386. https://doi.org/10.1159/000319015

Singer P. The expanding circle: Ethics, evolution, and moral progress. — Princeton: Princeton University Press, 2011. — 227 p.

Sparrow R. The turing triage test // Ethics and information technology. — 2004. — Vol. 6, No 4. — P. 203–213. https://doi.org/10.1007/s10676-004-6491-2

The benefits of and barriers to using a social robot PARO in care settings: scoping review / L. Hung, C. Liu, E. Woldum, et al. // BMC geriatrics. — 2019. — Vol. 19, No 1. — 232. https://doi.org/10.1186/s12877-019-1244-6

Turkle S. Reclaiming conversation: The power of talk in a digital age. — New York: Penguin Books, 2015. — 448 p.

Wallach W., Allen C. Hard problems: Framing the Chinese room in which a robot takes a moral Turing test // 38th Annual Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB 2012). Part 12. — New York: Curran Associates, 2014. — P. 1-6.

References:

Aharoni, E. et al. (2024) 'Attributions toward artificial agents in a modified Moral Turing Test', *Scientific Reports,* 14(1), 8458. https://doi.org/10.1038/s41598-024-58087-7

Alekseev, A. Yu. (2013) *Kompleksnyj test T'yuringa: filosofsko-metodologicheskie i sociokul'turnye aspekty [Comprehensive Turing Test: philosophical, methodological and socio-cultural aspects].* Moscow: IInteLL Publ. (In Russian).

Allen, C., Varner, G. and Zinser, J. (2000) 'Prolegomena to any future artificial moral agent', *Journal of Experimental & Theoretical Artificial Intelligence,* 12(3), pp. 251–261. https://doi.org/10.1080/09528130050111428

Arnold, T. and Scheutz, M. (2016) 'Against the moral Turing test: accountable design and the moral reasoning of autonomous systems', *Ethics and Information Technology,* 18(2), pp. 103–115. https://doi.org/10.1007/s10676-016-9389-x

Astakhov, S. (2020) 'Phenomenology vs Symbolic AI: Hubert Dreyfus's Philosophy of Skill Acquisition', *Philosophical Literary Journal Logos,* 30(2), pp. 157–193. (In Russian).https://doi.org/10.22394/0869-5377-2020-2-157-190

Bohn, E.D. (2024) 'The Moral Turing Test: a defense', Philosophy & Technology, 37(3), 111. https://doi.org/10.1007/s13347-024-00793-1

Bohn, E. D. (2025) 'In Defense of the Moral Turing Test: A Reply', *Philosophy & Technology,* 38(2), 40. https://doi.org/10.1007/s13347-025-00869-6

Broadbent, E. et al. (2024) 'ElliQ, an AI-Driven Social Robot to Alleviate Loneliness: Progress and Lessons Learned', *The Journal of Aging Research & Lifestyle,* 13, pp. 22–28. https://doi.org/10.14283/jarlife.2024.2

Calo, C. et al. (2011) 'Ethical Implications of Using the Paro Robot, with a Focus on Dementia Patient Care', in *Human-Robot Interaction in Elder Care.* San Francisco: Association for the Advancement of Artificial Intelligence, pp. 20–24.

Dela Cruz, N. L. (2025) 'Save the Digients! On the Moral Status of AI', in *The Philosophy of Ted Chiang.* Cham: Springer Nature Switzerland, pp. 195–202. https://doi.org/10.1007/978-3-031-81662-8_21

Dreyfus, H. L. (1992) *What Computers Still Can't Do: A Critique of Artificial Reason.* Cambridge: : MIT Press.

Gerdes, A. and Øhrstrøm, P. (2015) 'Issues in robot ethics seen through the lens of a moral Turing test', *Journal of Information, Communication and Ethics in Society,* 13(2), pp. 98–109. https://doi.org/10.1108/JICES-09-2014-0038

Hazlitt, H. (1972) *The foundations of morality.* Los Angeles: Nash Publ. (Russ. ed.: (2019) *Osnovaniya morali.* Moscow: Mysl Publ.; Chelyabinsk: Sotsium Publ.).

Hung, L. et al. (2019) 'The benefits of and barriers to using a social robot PARO in care settings: a scoping review', *BMC Geriatrics,* 19(1), p. 232. https://doi.org/10.1186/s12877-019-1244-6

Kolomiytsev, S.Yu. (2015) 'The Turing Test and Artificial Intelligence in the Early 21-st Century', *The human being,* (4), pp. 59–68. (In Russian).

Krzanowski, R. M. and Trombik, K. (2021) 'Ethical Machine Safety Test', in *Transhumanism: The Proper Guide to a Posthuman Condition or a Dangerous Idea?. Cognitive Technologies.* Cham: Springer, pp. 141–154. https://doi.org/10.1007/978-3-030-56546-6_10

Martynenko, N. P. (2025) 'Cognitive Mechanisms of Large Language Models: Interaction with GigaChat', *Concept: philosophy, religion, culture,* 9(2), pp. 30–50. (In Russian). https://doi.org/10.24833/2541-8831-2025-2-34-30-50

Merleau-Ponty, M. (1945) *Phénoménologie de la perception.* Paris: Éditions Gallimard. (Russ. ed.: (1999) *Fenomenologiya vospriyatiya.* Saint Petersburg: YUventa: Nauka Publ.).

Milgram, S. (1974) *Obedience to Authority: An Experimental View.* New York: Harper & Row. (Russ. ed.: (2023) *Podchinenie avtoritetu: Nauchnyj vzglyad na vlast' i moral'.* Moscow: Alpina Non-fiction Publ.).

Moor, J. H. (2020) 'The mature, importance, and difficulty of machine ethics ', in *Machine Ethics and Robot Ethics.* London: Routledge, pp. 233–236. https://doi.org/10.4324/9781003074991

Oleynikov, Yu. V. (2021) 'Singularity of Post-Industrial Society', *Knowledge, understanding, skill,* (2), pp. 85–95. (In Russian).

Proudfoot, D. (2024) 'Turing's Test vs the Moral Turing Test', *Philosophy & Technology,* 37(4), p. 134. https://doi.org/10.1007/s13347-024-00825-w

Razin, A. V. (2017) 'Morality and Mind: the Ideal and the Rational', *The human being,* (2), pp. 33–46. (In Russian).

Rezaev, A. V. and Tregubova, N.D. (2019) 'Artificial Intelligence, On-line Culture, Artificial Sociality: Definition of the Terms', *Monitoring of Public Opinion: Economic and Social Changes,* (6), pp. 35–47. (In Russian). https://doi.org/10.14515/monitoring.2019.6.03

Shibata, T. and Wada, K. (2011) 'Robot Therapy: A New Approach for Mental Healthcare of the Elderly — A Mini-Review', *Gerontology*, 57(4), pp. 378–386. https://doi.org/10.1159/000319015

Singer, P. (1975) *Animal Liberation: A New Ethics for our Treatment of Animals.* New York: Random House. (Russ. ed.: (2009) Osvobozhdenie zhivotnyh . Moscow: Sindbad Publ.).

Singer, P. (2011) *The expanding circle: Ethics, evolution, and moral progress.* Princeton: Princeton University Press.

Sparrow, R. (2004) 'The Turing Triage Test', *Ethics and Information Technology,* 6(4), pp. 203–213. https://doi.org/10.1007/s10676-004-6491-2

Turing, A. (1950) 'Computing machinery and intelligence', *Mind*, 59(236), pp. 433–460. (Russ. ed.: (1960) *Mogut li mashiny myslit'?* Moscow: Fizmatgiz Publ.).

Turkle, S. (2015) *Reclaiming conversation: The power of talk in a digital age.* New York: Penguin Books.

Wallach, W. and Allen, C. (2014) 'Hard problems: Framing the Chinese room in which a robot takes a moral Turing test ', in 38th Annual Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB 2012). Part 12. New York: Curran Associates, pp. 1–6.

Информация об авторе

Алексей Владимирович Антипов — кандидат философских наук, старший научный сотрудник сектора гуманитарных экспертиз и биоэтики Института философии Российской академии наук. 109240, Россия, г. Москва, ул. Гончарная, д. 12, стр. 1. (Россия)

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Information about the author

Aleksei V. Antipov — PhD in Philosophy, Senior research fellow, Department of Humanitarian Expertise and Bioethics, Institute of Philosophy, Russian Academy of Sciences. 12/1 Goncharnaya Str., Moscow, Russia, 109240 (Russia)

Conflicts of interest. The author declares absence of conflicts of interest.

Статья поступила в редакцию 08.10.2025; одобрена после рецензирования 05.11.2025; принята к публикации 10.12.2025.

The article was submitted 08.10.2025; approved after reviewing 05.11.2025; accepted for publication 10.12.2025.