УДК 004.8

136. ТЕСТ ТЬЮРИНГА: КОНЦЕПЦИЯ, ИСТОРИЯ СОЗДАНИЯ, АНАЛОГИ И КЛЮЧЕВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Яровская И.А., студент гр.578103, Лукашевич А.Э., студент гр.376741
Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Петрович Ю.Ю. – ст. преподаватель каф. ЭИ, магистр

Аннотация. Работа посвящена анализу теста Тьюринга как фундаментального критерия оценки искусственного интеллекта. Рассматривается история создания концепции «игры в имитацию», её эволюция от первых программ ELIZA и PARRY до современных языковых моделей GPT-4 и Gemini. Особое внимание уделяется критике подхода, включая мысленный эксперимент «китайская комната» Джона Сёрла, и анализу современных альтернатив: схеме Винограда и тесту Ады Лавлейс. В статье исследуются этические аспекты имитации человеческого поведения и перспективы перехода к комплексным методам оценки когнитивных способностей машин.

Ключевые слова. Искусственный интеллект, тест Тьюринга, игра в имитацию, машинное мышление, обработка естественного языка, большие языковые модели, когнитивные способности, имитация человека, глубокое обучение, критерии интеллекта.

Вопрос о способности машин мыслить долгое время оставался абстрактной философской проблемой. В 1950 году Алан Тьюринг предложил превратить его в практическую задачу, сформулировав критерий интеллекта, основанный на способности машины имитировать человека в диалоге [1]. Этот тест определил развитие искусственного интеллекта на десятилетия вперед.

Актуальность изучения теста Тьюринга обусловлена стремительным развитием систем искусственного интеллекта. Современные большие языковые модели — GPT-4, Claude, Gemini — демонстрируют способности к генерации текста и поддержанию диалога, которые ставят под вопрос традиционные критерии оценки машинного интеллекта. В 2014 году программа Eugene Goostman убедила 33% судей в своей «человечности», а современные чатботы регулярно вводят пользователей в заблуждение относительно своей природы. Эти достижения требуют переосмысления подходов Тьюринга и анализа того, действительно ли прохождение теста свидетельствует о наличии интеллекта или лишь о совершенстве имитации.

Тьюринг предполагал, что способность вести осмысленный диалог является достаточным критерием интеллекта, однако критики указывают на ограниченность этого подхода: тест не оценивает понимание физического мира, способность к обучению, творчество, здравый смысл. Более того, современные системы могут проходить тест, используя статистические методы и сопоставление шаблонов, не обладая подлинным пониманием.

В октябре 1950 года в философском журнале Mind была опубликована статья Тьюринга «Computing Machinery and Intelligence» («Вычислительные машины и разум»). В этой работе ученый поставил вопрос: «Могут ли машины мыслить?» и предложил заменить его более операциональным: «Могут ли машины делать то, что можем делать мы (как мыслящие существа)?»

Тьюринг осознавал, что прямой ответ на вопрос о способности машин мыслить требует определения самого понятия «мышление», что представляет собой философскую проблему. Вместо этого он предложил практический тест, основанный на способности машины имитировать человеческое поведение в диалоге.

Тьюринг описал процедуру, которую назвал «игрой в имитацию». Первоначально эта игра включала трех участников: мужчину (A), женщину (B) и судью (C), который должен был определить пол собеседников, задавая им вопросы в письменной форме. Задача мужчины — ввести судью в заблуждение, а женщины — помочь судье сделать правильный выбор [1].

Затем Тьюринг предложил заменить одного из участников машиной и поставил вопрос: «Что произойдет, если в этой игре роль A будет играть машина?» Если судья не сможет надежно отличить машину от человека, это будет свидетельствовать о наличии у машины интеллекта, сравнимого с человеческим.

Тьюринг предсказал, что к 2000 году машины смогут обманывать среднего судью в течение пяти минут примерно в 30% случаев [1]. Это предсказание стимулировало исследователей к созданию программ, способных вести осмысленный диалог с человеком.

В современной интерпретации тест Тьюринга проводится следующим образом: человек-судья взаимодействует через текстовый интерфейс с двумя собеседниками — человеком и компьютерной программой. Судья не знает, кто из них является машиной, и его задача — определить это, задавая любые вопросы. Если после серии вопросов судья не может с уверенностью идентифицировать машину, или если его выбор не превышает случайного угадывания, машина считается прошедшей тест.

Важно отметить, что тест проводится в текстовом формате, что исключает необходимость в физическом воплощении машины и позволяет сосредоточиться на интеллектуальных способностях системы.

Одной из первых программ, продемонстрировавших возможность имитации человеческого диалога, стала ELIZA, созданная Джозефом Вейценбаумом в Массачусетском технологическом институте в 1966 году. В основе программы лежала имитация работы психотерапевта, использующего клиент-центрированный подход Карла Роджерса.

ELIZA использовала простой алгоритм сопоставления шаблонов: она искала в высказываниях пользователя ключевые слова и формировала ответы на основе заранее заданных правил. Например, если пользователь упоминал слово «мать», программа могла ответить: «Расскажите мне больше о вашей семье».

Несмотря на примитивность алгоритма, ELIZA произвела сильное впечатление на пользователей. Многие люди, включая секретаря Вейценбаума, приписывали программе понимание и эмпатию, хотя она лишь манипулировала текстовыми шаблонами. Данный феномен, обозначенный как «эффект ELIZA», выявил свойственную человеку склонность наделять компьютерные системы антропоморфными качествами.

Сам Вейценбаум был обеспокоен тем, что люди воспринимают ELIZA как нечто большее, чем простую программу, и впоследствии стал критиком искусственного интеллекта, предупреждая об опасности чрезмерного доверия к машинам.

В 1972 году психиатр Кеннет Колби из Стэнфордского университета создал программу PARRY, которая моделировала поведение человека с параноидальной шизофренией. В отличие от ELIZA, PARRY обладала более сложной внутренней моделью эмоциональных состояний и убеждений.

Для оценки реалистичности PARRY был проведен эксперимент: группе опытных психиатров предложили проанализировать транскрипты разговоров и определить, какие из них велись с реальными пациентами, а какие — с программой. Психиатры не смогли надежно различить PARRY и реальных пациентов, что стало важным достижением в области искусственного интеллекта.

Интересно, что в 1973 году был организован «диалог» между ELIZA и PARRY через сеть ARPANET (предшественник интернета), что стало одним из первых примеров сетевого взаимодействия программ искусственного интеллекта.

В 1990 году американский изобретатель Хью Лёбнер учредил ежегодный конкурс, в котором программы-чатботы соревнуются в прохождении теста Тьюринга. Конкурс предусматривает денежный приз в размере 100 000 долларов США для первой программы, которая пройдет полноценный тест Тьюринга, включая способность обрабатывать аудио и видео.

Хотя ни одна программа не выиграла главный приз, конкурс Лёбнера стимулировал развитие технологий обработки естественного языка и привлек внимание к проблемам создания диалоговых систем. Критики конкурса, включая Марвина Минского, утверждали, что он поощряет создание программ, ориентированных на обман, а не на подлинное понимание.

В июне 2014 года на мероприятии в Королевском обществе в Лондоне, посвященном 60-летию со дня смерти Алана Тьюринга, программа-чатбот Eugene Goostman убедила 33% судей в том, что она является человеком [3]. Программа имитировала 13-летнего мальчика из Украины, что, по мнению создателей, объясняло возможные ошибки в грамматике и ограниченность знаний.

Это событие вызвало широкую дискуссию в научном сообществе. Многие исследователи, включая Гэри Маркуса из Нью-Йоркского университета, критиковали заявления о прохождении теста, указывая на то, что программа использовала уловки и что критерий в 30% не является общепринятым стандартом. Кроме того, выбор образа подростка-иностранца был воспринят как попытка снизить ожидания судей.

Тем не менее, Eugene Goostman продемонстрировал прогресс в области создания диалоговых систем и показал, что современные технологии приближаются к порогу, установленному Тьюрингом.

Развитие глубокого обучения и архитектур трансформеров привело к созданию больших языковых моделей, таких как GPT (Generative Pre-trained Transformer), BERT, и других систем,

способных генерировать связный и контекстуально релевантный текст. Эти модели обучаются на огромных массивах текстовых данных и демонстрируют впечатляющие способности в понимании и генерации естественного языка.

Исследования показывают, что современные языковые модели могут успешно выполнять многие задачи, связанные с обработкой текста, включая ответы на вопросы, перевод, резюмирование и даже написание программного кода. Однако вопрос о том, обладают ли эти системы подлинным пониманием или лишь имитируют его, остается предметом дебатов.

Одна из наиболее влиятельных критик теста Тьюринга была представлена философом Джоном Сёрлом в 1980 году в виде мысленного эксперимента «китайская комната» [2]. Сёрл предложил представить человека, не знающего китайского языка, который находится в закрытой комнате с набором правил на английском языке для манипуляции китайскими символами.

Когда в комнату передают вопросы на китайском языке, человек использует правила для формирования ответов, которые выглядят осмысленными для носителей китайского языка. С точки зрения внешнего наблюдателя, система (комната с человеком) понимает китайский язык и проходит тест Тьюринга. Однако человек внутри комнаты не понимает ни слова по-китайски — он лишь манипулирует символами согласно формальным правилам.

Сёрл использовал этот аргумент, чтобы показать, что синтаксическая обработка символов (которую выполняют компьютеры) не эквивалентна пониманию смысла. Следовательно, прохождение теста Тьюринга не доказывает наличие сознания или подлинного понимания у машины.

Этот аргумент вызвал обширную дискуссию в философии сознания и искусственного интеллекта, породив множество ответов и контраргументов, включая «системный ответ» (понимание присуще всей системе, а не отдельному человеку) и «робот-ответ» (система, взаимодействующая с физическим миром, может обладать пониманием).

Критики указывают, что тест Тьюринга оценивает способность машины имитировать человека, а не наличие интеллекта как такового.

Французский исследователь искусственного интеллекта Жан-Габриэль Ганасси отмечает, что самолет не летает, махая крыльями, как птица, но это не делает его менее способным к полету. Аналогично, машинный интеллект может отличаться от человеческого, но при этом быть подлинным и эффективным.

Тест Тьюринга фокусируется исключительно на лингвистических способностях и не оценивает другие аспекты интеллекта, такие как восприятие, моторные навыки, эмоциональный интеллект или способность к физическому взаимодействию с миром. Это ограничивает его применимость для оценки полноценного искусственного интеллекта.

Кроме того, тест не проверяет способность к обучению, адаптации и творчеству — качества, которые многие исследователи считают важными компонентами интеллекта.

Рассел и Норвиг предложили концепцию «полного теста Тьюринга» (Total Turing Test), который включает видеосигнал, позволяющий судье оценивать способности системы к восприятию объектов, а также возможность передавать физические объекты через «люк». Для прохождения такого теста компьютер должен обладать:

- Компьютерным зрением для восприятия объектов
- Робототехникой для манипуляции объектами и перемещения

Этот расширенный тест охватывает более широкий спектр когнитивных и физических способностей, приближаясь к комплексной оценке искусственного интеллекта.

В 2011 году Гектор Левеск предложил альтернативу тесту Тьюринга, известную как Winograd Schema Challenge [4]. Этот тест состоит из вопросов, требующих понимания контекста для разрешения неоднозначности местоимений в предложениях.

Пример схемы Винограда: «Городской совет отказал демонстрантам в разрешении, потому что они опасались насилия. Кто опасался насилия?»

Ответ зависит от понимания причинно-следственных связей и знания о том, как обычно функционируют социальные институты. Изменение одного слова в предложении может изменить правильный ответ, что делает невозможным использование статистических методов без подлинного понимания.

Преимущество схем Винограда заключается в объективности оценки (ответ либо правильный, либо нет) и в том, что они труднее поддаются обману через простое сопоставление шаблонов.

Марк Ридл и его коллеги предложили тест, названный в честь Ады Лавлейс, который оценивает творческие способности искусственного интеллекта. Согласно этому тесту, система должна создать оригинальное произведение (рассказ, стихотворение, изображение), причем сама система должна быть способна объяснить процесс создания.

Ключевое требование заключается в том, что создатели системы не должны быть способны объяснить, как именно система создала конкретное произведение, что свидетельствует о подлинной креативности, а не о воспроизведении заложенных алгоритмов.

Гэри Маркус, критикуя поверхностность теста Тьюринга, предложил оценивать глубину понимания через способность системы читать и понимать сложные тексты, отвечать на вопросы о причинах и следствиях, а также демонстрировать гибкость мышления.

Маркус подчеркивает, что современные системы часто полагаются на статистические корреляции в данных, не обладая подлинным пониманием концепций и причинно-следственных связей, что ограничивает их способность к обобщению и адаптации к новым ситуациям.

Тест Тьюринга оказал глубокое влияние на развитие искусственного интеллекта как научной дисциплины. Он предоставил конкретный, измеримый критерий для оценки прогресса и стимулировал исследования в области обработки естественного языка, представления знаний и машинного обучения.

Многие современные технологии, включая виртуальных ассистентов, чатботов и системы автоматического перевода, являются прямыми потомками исследований, вдохновленных тестом Тьюринга. Даже если эти системы не стремятся пройти тест в его классической форме, они используют принципы и методы, разработанные в процессе работы над этой задачей.

По мере того, как системы искусственного интеллекта становятся все более способными к имитации человеческого поведения, возникают важные этические вопросы. Если машина может убедительно имитировать человека, как это влияет на наше понимание идентичности, ответственности и прав?

Существует опасность, что люди могут быть обмануты системами ИИ, что поднимает вопросы о необходимости прозрачности и раскрытия информации о природе собеседника. Некоторые юрисдикции уже вводят требования о том, чтобы боты идентифицировали себя как таковые при взаимодействии с людьми.

Современные исследователи все чаще признают, что единого универсального теста на интеллект не существует. Вместо этого предлагается использовать набор специализированных тестов, оценивающих различные аспекты когнитивных способностей: рассуждение, обучение, восприятие, творчество, социальное взаимодействие и т.д.

Франсуа Шолле, исследователь из Google, предложил концепцию оценки интеллекта через способность к обобщению и эффективность обучения, а не через выполнение конкретных задач [5]. Этот подход фокусируется на измерении того, насколько быстро и с каким количеством примеров система может освоить новые навыки.

Тест Тьюринга, предложенный более семидесяти лет назад, остается одной из самых влиятельных идей в истории искусственного интеллекта. Несмотря на многочисленную критику и появление альтернативных подходов, он продолжает служить отправной точкой для дискуссий о природе интеллекта и критериях его оценки.

Анализ ключевых исследований — от ELIZA и PARRY до современных языковых моделей — демонстрирует значительный прогресс в создании систем, способных к убедительной имитации человеческого диалога. Однако философские вопросы, поднятые Сёрлом и другими, остаются открытыми, напоминая, что технический прогресс в имитации диалога не тождествен решению проблемы подлинного понимания и сознания.

Список использованных источников:

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

2. Searle, J. R. Minds, brains, and programs / J. R. Searle // Behavioral and Brain Sciences. – 1980. – Vol. 3, № 3. – P. 417-424.

3. Bohannon, Eugene Goostman chatbot passes Turing test / J. Bohannon // Science. – 2014. – Vol. 344, № 6189. – P. 1230-1231.

4. Levesque, H. J. The Winograd Schema Challenge / H. J. Levesque, E. Davis, L. Morgenstern // Principles of Knowledge Representation and Reasoning. – 2012. – P. 552-561.

5. Chollet, F. On the Measure of Intelligence / F. Chollet // arXiv preprint arXiv:1911.01547. – 2019.

UDC 004.8

TURING TEST: CONCEPT, HISTORY OF CREATION, ANALOGS AND KEY RESEARCH IN THE FIELD OF ARTIFICIAL INTELLIGENCE

Yarovskaya I.A., student gr.578103, Lukashevich H.E., student gr.376741

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Petrovich Y.Y. – senior lecturer, master's degree

Annotation. The article is devoted to the analysis of the Turing test as a fundamental criterion for evaluating artificial intelligence. It examines the history of the "imitation game" concept and its evolution from the first programs, such as ELIZA and PARRY, to modern language models like GPT-4 and Gemini. Particular attention is paid to the critique of this approach, including John Searle's "Chinese Room" thought experiment, and the analysis of modern alternatives: the Winograd Schema and the Ada Lovelace Test. The article explores the ethical aspects of simulating human behavior and the prospects for transitioning to comprehensive methods for assessing the cognitive abilities of machines.

Keywords. Artificial intelligence, Turing test, imitation game, machine thinking, natural language processing, large language models, cognitive abilities, human imitation, deep learning, intelligence criteria.