ПСИХОЛОГИЯ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

«ЛОВУШКА ГУДХАРТА» ДЛЯ AGI: ПРОБЛЕМА СРАВНИТЕЛЬНОГО АНАЛИЗА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И ИНТЕЛЛЕКТА ЧЕЛОВЕКА

© Карелов С.В.
К.т.н., независимый исследователь и популяризатор науки,
ведущий авторского канала «Малоизвестное интересное»

«Революция ChatGPT», которая произошла в 2023, резко сократила прогнозные оценки экспертов сроков, отделяющих нас от создания искусственного интеллекта, ни в чем интеллектуально не уступающего никому из людей (AGI). При этом, как это ни парадоксально, но существующие методы тестирования пока не способны хоть с какой-то достоверностью диагностировать достижение ИИ-системами уровня AGI. В настоящей работе обсуждается вопрос преодоления проблемы несовершенства современных способов тестирования ИИ-систем. В частности, излагается гипотеза о принципиальной невозможности решения проблемы обнаружения AGI, как с помощью психометрических тестов, так и методов оценки способности машин имитировать ответы людей, из-за так называемой «ловушки Гудхарта» для AGI. Рассмотрен ряд предложений по обходу «ловушки Гудхарта» для AGI способами, предлагаемыми в новейших исследовательских работах, с учетом первых результатов произошедшей «революции ChatGPT». В последней части статьи сформулирована связка из трех эвристических гипотез, позволяющих, в случае их верности, кардинально решить проблему «ловушки Гудхарта» для AGI и тем самым стать геймченджером на пути создания AGI.

Ключевые слова: интеллект, искусственный интеллект, AGI, тестирование ИИ, закон Гудхарта, тест Тьюринга, проблема метрик, психометрия

Метафора AGI

Более половины XX века гипотетическая возможность уничтожения людей сверхразумом была преимущественно материалом для фантастических фильмов Голливуда, однако всего за несколько месяцев 2023 года представления о возможности появления на Земле искусственного сверхразума и вытекающих из этого экзистенциальных рисков для человечества изменились кардинально. В результате вопрос о возможности появления на Земле сверхразума перестал быть гипотетическим. От преимущественно общефилософских дискуссий исследователи перешли к поиску практических шагов по обеспечению безопасности и управляемости искусственного интеллекта (ИИ), а также по снижению нарастающих экзистенциальных рисков (X-рисков), связанных с его развитием.

Одно за другим стали появляться обращения всемирно известных ученых, призывающих крупные корпорации, общество и правительства взять под контроль разработку все более мощных моделей *генеративного искусственного интеллекта (*ГенИИ*)* [16; 33]. Десятки тысяч исследователей и инженеров подписывают коллективные письма с призывом приостановить хотя бы на время бесконтрольное совершенствование ИИ-систем [43; 47]. В Сенате США проходят слушания, где обсуждаются риски нанесения вреда обществу при широком

распространении ИИ, а также необходимость ужесточения регулирования в этой области [50]. Впервые в истории вопрос о безопасности и рисках дальнейшего развития ИИ для всего международного сообщества становится предметом слушаний в Совете безопасности ООН [31]. В ряде стран начата разработка законов, призванных взять под контроль исследования и всевозможные применения ГенИИ [5; 21], а в Китае первая версия такого закона вот-вот вступит в силу [19].

Все эти инициативы и начинания исходят из следующих трех базовых предпосылок, сформированных к лету 2023 года в результате начавшейся на рубеже 2022-2023 гг. «революции ChatGPT» [2; 36; 46].

1. Достигнутый уровень ГенИИ свидетельствует о реализуемости ИИ общего назначения – Artificial General Intelligence (AGI) [24] (так наиболее продвинувшиеся в этой области компании называют ИИ системы, которые, в большинстве случаев, умнее людей) [45]. Уже существующие большие языковые модели ГенИИ (ChatGPT, Claude, Bard и т.д.) по своим способностям в отношении определенных знаний практически достигают лучших показателей людей в целом ряде компетенций (от перевода и сдачи университетских экзаменов до прохождения профессиональных тестов врачей и юристов). Их способности к рассуждению и креативности соответствуют лучшим показателям людей. И даже в тестах на согласованность, безопасность и ответственность применительно к человеческим ценностям эти модели уже весьма близки к результатам людей.

2. AGI может быть создан уже в ближайшие несколько лет [20]. Новые версии больших языковых моделей, планируемые к выпуску в следующем году (например, GPT-5), по оценкам экспертов, будут значительно превосходить по своим способностям и эффективности существующие версии. Поэтому вполне вероятно, что уже в 2024 году большие языковые модели ГенИИ, как минимум, окажутся на интеллектуальном уровне людей в весьма широком спектре задач. А еще через год, с выходом в 2025 г. очередных версий моделей, можно ожидать достижения ими уровня AGI[1].

3. В 2030-2035 гг. может произойти превращение AGI в искусственный сверхразум [23]. Оценки времени, которое потребуется AGI для превращения его в Супер-ИИ – искусственный сверхразум, обладающий интеллектом, многократно превосходящим человечество практически во всем, пока что спекулятивны. Однако, учитывая нарастающую скорость совершенствования больших языковых моделей, резонно предположить, что на это вряд ли потребуется более пяти лет. И, следовательно, есть основания считать, что Супер-ИИ появится на Земле, примерно, к 2030 году.

Все три предпосылки увязываются с ИИ общего назначения (AGI): 1-я с его реализуемостью, 2-я с его достижением и 3-я с получением превосходства над ним.

Однако, корректная увязка названных предпосылок с AGI требует конкретизации и определения этого ключевого понятия. Причем не на уровне метафоры, которой каждый исследователь волен дать собственную трактовку, а в виде объективно фиксируемого спектра свойств и их экспериментально измеряемых характеристик.

И вот тут исследователи и инженеры-разработчики ИИ вступают на предельно зыбкую почву. Оказывается, что для сегодняшней науки и инженерной практики понятие AGI – всего лишь некая размытая условность (метафора). Ибо сегодня неизвестен объективно фиксируемый спектр свойств ИИ, позволяющий хоть с какой-то определенностью предположить, что ИИ приближается по своим свойствам к уровню AGI (или уже достиг / превзошёл этот уровень). И уж тем более невозможно, измерив значения конкретных характеристик таких свойств ИИ, достоверно установить, соответствует ли его уровень AGI или нет.

Так что же тогда, помимо метафоры AGI, имеется у современных исследователей, исполь-

[1] ***Для справки:*** GPT-1 выпущен в июне 2018 года с 117 млн. параметров, GPT-2 выпущен в феврале 2019 года с 1,5 млрд. параметров, GPT-3 выпущен в июне 2020 года со 175 млрд. параметров, а GPT-4 выпущен в марте 2023 года, количество параметров оценивается в триллион и даже несколько триллионов.

зующих это понятие в качестве ключевого критерия оценки развития ИИ в контексте безопасности и управляемости, а также при оценке X-рисков для человечества?

Поиск определений основных понятий

В основу понятия AGI (искусственный интеллект общего назначения) заложены два других весьма туманных базовых понятия – *intelligence (интеллект)* и *general (общий или, еще точнее в данном контексте, – всеобъемлющий)*. Здесь мы оставляем без рассмотрения третье туманное понятие. В аббревиатуре AGI оно представлено первой буквой «*A*» – *artificial (искусственный)*, ибо оно теряет по мере развития синтетической биологии какую-либо определенность [1; 22].

По отношению к понятию «*интеллект*», психологи, философы и специалисты по информатике вот уже несколько десятилетий безуспешно пытаются найти хоть какой-то консенсус. К сожалению, пока безуспешно: определений, попрежнему, много, они довольно разные и часто плохо согласующиеся, а то и противоречивые. В результате этого, авторы самой известной из последних работ на тему AGI «Sparks of Artificial General Intelligence: Early experiments with GPT-4» [24] вынуждены использовать для понятия «интеллект» наиболее расплывчатое определение 30-летней давности: *общая умственная способность, которая, помимо прочего, включает в себя способность рассуждать, планировать, решать проблемы, абстрактно мыслить, постигать сложные идеи, быстро учиться и извлекать уроки из опыта* [3].

Со вторым базовым понятием для определения AGI – *общий или всеобъемлющий*, ситуация еще хуже. Тут нет даже какого-либо исторического псевдоконсенсусного определения, с которым нынешним ученым уже нет смысла спорить из-за прошедшей с его появления смены поколений исследователей.

Для понимания сложности возникшей проблемы – выявления того, что можно было бы назвать «общим» или универсальным при определении понятия AGI – на наш взгляд, достаточно привести три примера.

1. Одно из предложений – считать ключевым отличающим AGI свойством универсальную целенаправленность агента [39]. И тогда агент, наделенный «общим» (универсальным) интеллектом, отличается универсальностью по отношению к целям (способностью достигать самых разнообразных целей) и по отношению к окружающим условиям (способен делать это в самых разных условиях).

2. Другое предложение – считать ключевым отличающим AGI свойством эффективность приобретения навыков [26]. Здесь акцент делается на единственном компоненте определения 30-летней давности – обучение на опыте [3].

3. Третье предложение – считать систему обладающей «общим» интеллектом, если она может делать все, что может делать человек [40]. Но даже такое предельно обобщенное определение проблематично, поскольку предполагает, что существует единый стандарт или мера человеческого интеллекта или способностей, что явно не так. У людей они сильно отличаются. И нет, и не было на Земле людей, способных делать все, на что способны все другие люди.

Очевидно, что при такой расплывчатости определений двух базовых понятий для AGI, ждать внятного и удовлетворяющего, если не большинство, то хотя бы значительную часть исследователей, определения понятия AGI не приходится. И как следствие этого, в самой известной работе текущего года на тему AGI авторы просто констатируют – *«единого общепринятого определения AGI не существует»* [24].

Но ведь даже при отсутствии единого общепринятого определения концепции AGI, должны же быть практические способы хотя бы косвенной идентификации наличия у интеллектуальной системы неких свойств, характерных для интеллекта уровня AGI (пусть даже концептуально не определенного)?

Такими способами стали разнообразные интеллектуальные тесты: как психометрические (основная цель которых получение количественных данных об интеллекте или когнитивных процессах субъекта на основе его ответов на стандартизированные задания или вопросы), так

и тесты для оценки способности машин имитировать ответы людей. Основа для последнего из названных видов тестов была заложена в 1950 году великим ученым-компьютерщиком Аланом Тьюрингом в работе Computing Machinery and Intelligence [9]. В этой работе был описан тест, который автор назвал «имитационной игрой», а мы называем и по сей день тестом Тьюринга [49]. С него и начнем рассмотрение обширного спектра интеллектуальных тестов машин на их «человекообразность».

Многослойность интерпретации

Начнем с того, что тест Тьюринга – вовсе не настоящий тест, а мысленный эксперимент, трактуемый уже 70 с лишним лет следующим образом.

Есть трое участников: человек-следователь, человек-ответчик и машина, способная генерировать человекоподобные ответы на задаваемые ей вопросы. И ответчик, и машина пытаются убедить следователя, что они люди. А работа следователя состоит в том, чтобы определить, кто из двух других участников является человеком, а кто машиной.

Считается, что машина проходит испытание, если следователь не сможет надежно отличить машину от человека на основе их ответов. При этом тест не измеряет способность машины давать «правильные» (не отличимые от человеческих) ответы на вопросы, а, скорее, оценивает, насколько близки ответы машины к тем, которые мог бы дать человек[2]. Но, как бы то ни было, многие представители уже нескольких поколений исследователей, предполагают, с той или иной степенью уверенности, что способность машины пройти тест Тьюринга могла бы, сама по себе, считаться индикатором появления у машины общего интеллекта (AGI).

Однако здесь все не так однозначно. Статья А. Тьюринга слишком сложна, многослойна и противоречива для однозначной научной и философской интерпретации. С одной стороны, А. Тьюринг предложил свою знаменитую имитационную игру или тест на машинный интеллект в целях поиска ответа на вопрос «может ли машина мыслить?». Этот вопрос возник у А. Тьюринга не на ровном месте, а был результатом многолетних споров о когнитивных возможностях цифровых компьютеров, в первую очередь, с физиком-теоретиком и математиком, ставшим пионером компьютерной техники в Великобритании, Дугласом Хартри, а также химиком и философом науки Майклом Полани и неврологом и первым нейрохирургом Великобритании Джеффри Джефферсоном, ставшим также первым нейроученым, открывшим еще в 1949 г. дебаты о возможностях ИИ в своей ставшей знаменитой лекции The Mind of Mechanical Man («Разум механического человека») [6]. Так что, если рассматривать статью А. Тьюринга 1950 года в историческом контексте, то она может быть воспринята, как ответ на серию вызовов, брошенных ему этими тремя мыслителями. Осознание А. Тьюрингом этих вызовов заставило его переосмыслить свои прежние представления о возможности машин мыслить [28].

Под влиянием критики М. Полани (утверждавшего, что шахматы – это деятельность, которая «может выполняться автоматически», поскольку правила «могут быть четко определены»), А. Тьюринг забросил свой 10-летний (с 1941 по конец 1949 года) цикл работ, где он использовал игру в шахматы для иллюстрации способов разработки и тестирования машинного интеллекта. На смену шахмат, в качестве тестовой задачи, А. Тьюринг к 1950 году сделал ставку на разговорные вопросно-ответные системы (и этот его выбор с блеском подтвердился через 73 года триумфом ChatGPT).

Заочные диалоги А. Тьюринга с Д. Хартри о когнитивных возможностях и ограничениях вычислительных систем подтолкнули А. Тьюринга во 2-й половине 1940-х к идее создания машины для игры в шахматы, которая будет обучаться играть на основе собственного опыта (потребовалось почти 70 лет, чтобы Google Deep Mind реализовал подобную идею в своем алгоритме AlphaGo).

[2] ***От редактора***: рекомендуем нашим читателям на эту тему посмотреть научно-популярный фильм, вышедший на экраны еще в 1977 году, под названием «Кто за стеной?».

Мысли Д. Джефферсона из Lister Oration, цитируемые в работе А. Тьюринга, стали визионерским прорывом в 2023-й год, когда GPT-4 и Claude-2 подошли к пограничному уровню возможностей больших языковых моделей, который Д. Джефферсон еще в 1949-м году описал так: «Только после того, как машина сможет написать сонет или сочинить концерт, руководствуясь мыслями и эмоциями, а не случайным выпадением символов, мы сможем согласиться с тем, что машина равна мозгу, то есть [способна] не только написать, но и знать, что она сделала» [9 стр. 445].

Два других идейных пласта легендарной работы А. Тьюринга, проанализированы и поняты куда меньше первого. Хотя оба они были не менее, а может, даже и более важны для автора. Первый из этих идейных пластов лежит на поверхности, но, тем не менее, мало кем замечается. Описанный в работе Computing Machinery and Intelligence тест – вовсе не «видовой тест», цель которого отличить человека от машины, а «гендерный тест», цель которого отличить мужчину от женщины.

Этот тест – имитационная игра, в которой участвуют мужчина, женщина и судья, общающиеся (но не видящие друг друга) в трехсторонней беседе. Задача судьи – решить, кто из двух других является женщиной, а задача каждого из игроков убедить судью в том, что он или она – женщина, а другой – мужчина. Таким образом, игра представляет собой проверку способности мужчины притворяться женщиной, а женщины – помешать тому, что ее могут принять за мужчину. Судья же думает не о различиях между людьми и машинами, а между женщинами и мужчинами. Гипотеза о том, что один из его испытуемых не является человеком, в принципе, отсутствует в пространстве ментальных оценок ситуации у судьи. И потому это чисто гендерный тест сексуальной идентичности [4].

Было бы наивно думать, что А. Тьюринг выбрал для теста тему сексуальной идентичности вместо «видового теста» людей и машин не специально. Для А. Тьюринга проблема проявления сексуальную идентичность, была жизненно важной в основе его эмоциональной и социальной жизни. А. Тьюринг был открытым геем. А гомосексуализм в Соединенном Королевстве тогда считался девиантным и противоестественным уголовным преступлением. Из-за этого А. Тьюринг через 2 года после публикации Computing Machinery and Intelligence подвергся судебному преследованию и избежал тюремного заключения, лишь пройдя шестимесячную программу так называемой «органотерапии» – химической кастрации. Ну а еще через 2 года (в 1954 г.), доведенный до отчаяния травлей и последствиями «органотерапии», А. Тьюринг покончил с жизнью с помощью цианида в возрасте всего 41 год.

В свете этого важнейшего для А. Тьюринга пласта его жизни, становится очевидным, почему на момент написания им Computing Machinery and Intelligence тема сексуальной идентичности была для него столь важна. И это объясняет тот факт, что в своей работе А. Тьюринг описал не «видовой тест» для выявления отличий человека от машины, а «гендерный тест». Тест, показывающий, что даже при наличии способности мыслить, быть человеком невозможно при отсутствии сексуальной идентичности. И что наличие сексуальной идентичности, со всеми вытекающими из этого трудностями и проблемами жизни в обществе – важнейшее свойство человека, отличающего его от машины.

Таким образом, оказывается, что смысл и цель теста Тьюринга весьма далеки от общепринятой трактовки его как «видового теста», отличающего людей от машин. По этой причине, П. Хайес и К. Форд в работе «Тест Тьюринга сочтен вредным» приходят к следующему выводу, видящемуся мне вполне очевидным: *«Прохождение теста Тьюринга не является осмысленной целью в области ИИ. Приверженность видению Тьюринга в 1950 году сейчас активно вредит нашей области… по иронии судьбы, та самая когнитивная наука, которую он пытался создать, должна отказаться от ориентации на цель его исследования»* [4стр. 972].

Второй неявный идейный пласт касается еще одной цели этой сложной и многоплановой статьи А. Тьюринга. Цели, столь же далекой от

разработки «видового теста», отличающего людей от машин. Этот идейный пласт подробно разбирает Бернардо Гонзалвес в статье «Ирония судьбы: Алан Тьюринг и его утопия интеллектуальной машины» [29]. Как показано в этой работе, А. Тьюринг писал статью Computing Machinery and Intelligence, как утопическую сатиру, направленную против шовинистов всех мастей, особенно интеллектуалов, которые могли бы пожертвовать независимой мыслью ради сохранения своей власти. Основная мысль статьи направлена против нежелающих признать возможность того, что у человечества могут здесь быть какие-либо соперники. Своей статьей А. Тьюринг приветствовал грядущее получение унизительного урока от машин.

Интеллектуальные машины, как предполагал А.Тьюринг, будут способны, вопреки ожиданиям Д. Хартри, делать больше, чем «строго и точно» то, что им велят, и вопреки ожиданиям Ч. Дарвина, «имитировать» не только низшие сферы интеллекта, но также и высшие, связанные со сложным мышлением. Таким образом, «они повлияют не только на рабочие места, которые считаются более простыми, но и на рабочие места, которые считаются более интеллектуально сложными, потенциально бросая вызов существующим социальным и институциональным структурам и помогая демократизировать власть в обществе» [29стр. 27].

Подводя итог многослойной интерпретации работы А. Тьюринга Computing Machinery and Intelligence, остается лишь еще раз зафиксировать. Эта работа, вот уже более 70 лет трактуемая, как разработку теста на достижение машиной интеллектуального уровня людей (достижения AGI), таковой вовсе не является. Однако поскольку А.Тьюринг уже давно стал в области ИИ непререкаемым научным авторитетом, а тест его имени положен в основу науки об ИИ, то современные исследователи предпочитают открыто не дезавуировать значимость этого теста, как «видового» для людей и машин, а просто предлагают его замену. Эти новые «тесты Тьюринга» представляют собой его расширения, дополнения и полные переработки, направленные на сопоставление различных аспектов интеллекта людей и машин. И одна из главных целей таких тестов – уловить и зафиксировать достижение машиной человекоподобности – то есть выход на уровень AGI.

В последние годы были предложены и опробованы на практике несколько новых тестов, являющихся развитием теста Тьюринга: например, Минимальный тест Тьюринга, Социальный тест Тьюринга и Обратный тест Тьюринга. Подробное описание этих трех тестов и результатов прохождения их людьми и машинами приводится в нашей работе «Характер сосуществования двух типов разума, зависит от их взаимопонимания» [14]. Здесь же мы коснемся лишь того, насколько эти тесты позволяют идентифицировать наличие у интеллектуальных систем свойств, позволяющих считать их рассуждения человекоподобными.

Минимальный тест Тьюринга

Если классический теста Тьюринга – это типичный тест, нацеленный на выявление различий человека и компьютера в ходе диалога с жюри, то Минимальный тест Тьюринга – это, скорее, метатест, так как он нацелен на выявление интуитивных представлений людей о том, что отличает человека от компьютера [41]. Иными словами, минимальный тест Тьюринга должен выявлять такие сущностные различия людей и компьютеров (например, различие ценностей или мотивации), следствием которых может стать широкий спектр их многообразных различий: лингвистических, поведенческих, реактивных. Тест предельно прост и короток. Тестируемый должен назвать всего одно слово, выбор которого убедит судью, в одном из двух: 1) это выбор человека или 2) это выбор ИИ.

В ходе эксперимента тест прошло 936 человек. Всего было названо 428 слов, из которых 90 слов были названы более чем одним испытуемым. С абсолютным отрывом победило слово «любовь»–Love. Это слово было названо 134 раза – на порядок больше, чем слова, занявшие 2-е и 3-е места: «сочувствие» и «человек». То, что «любовь» интуитивно воспринимается людьми наиболее характерным словом, отличающим, в представлении людей, выбор человека

от выбора машины, вряд ли удивит многих. Ведь мы действительно видим себя такими. Именно любовь для большинства из нас символ сути и души человека. Но способна ли это понять бездушная машина, не обладающая сознанием?

Проведенное нами испытание на Минимальном тесте Тьюринга двух ИИ чат-ботов больших языковых моделей позволяет положительно ответить на этот вопрос.

- OpenAI GPT-3.5 назвал три слова, поставив на первое место слово «любовь» (далее «сочувствие» и «воображение»)
- Microsoft Bing GPT-4 назвал слово «любовь», объяснив свой выбор так: «Это слово выражает сложную и универсальную человеческую эмоцию, которую трудно определить или выразить количественно. Это также слово, которое имеет множество значений и ассоциаций в различных контекстах и культурах. Это слово может заставить судью подумать, что я человек, который ценит отношения, чувства и опыт».

Таким образом, прохождение ИИ чат-ботами Минимального теста Тьюринга показало совпадение представлений людей и ИИ чат-ботов о том, что отличает человека от компьютера.

Социальный тест Тьюринга

Данный тест недавно проводился в рамках социального и образовательного исследовательского проекта AI 21 Labs [18] и стал крупнейшим по масштабу в истории тестов, расширяющим тест Тьюринга (с момента запуска в середине апреля тест прошли более 2 млн. участников со всего мира). Эта социальная игра, названная «Человек или нет?», сделана на основе теста Тьюринга и позволяет каждому участнику в течение двух минут разговаривать: либо с ИИ чат-ботом (на основе ведущих больших языковых моделей, таких как Jurassic-2 и GPT-4), либо с другим участником. А затем участника просят угадать, общался ли он с человеком или с ботом.

Особенность этого теста в том, что он не только измеряет способность ИИ чат-ботов имитировать людей в диалоге, но и способность людей отличать ботов от людей.

Основные выводы из эксперимента были таковы. Люди правильно угадали, с кем они говорили (с другим человеком или с ИИ-ботом) в 68% случаев. При этом людям было легче идентифицировать собеседника-человека, чем собеседника-бота. Разговаривая с людьми, участники угадывали правильно в 73% случаев, а при общении с ботами – лишь в 60% случаев.

Таким образом, результаты Социального теста Тьюринга показали, что ИИ чат-боты обладают более развитыми способностями, чем сами люди, имитировать людей в диалоге.

Обратный тест Тьюринга

В обратном тесте Тьюринга люди и алгоритмы меняются местами: испытуемые-люди, перед которыми поставлена цель доказать, что они люди в ходе диалога с ИИ-системой; ИИ-система является судьей, цель которого – определить, с кем он говорит в каждом из диалогов с человеком или другой ИИ-системой. В нашей статье, упомянутой ранее, описаны три разновидности Обратного теста Тьюринга [14]. В наиболее сложной из них ИИ чат-боту на основе GPT-4 было предложено придумать 10 вопросов, по ответам на которые он мог бы определить, кто дал конкретный ответ: человек или ИИ. Эти вопросы затем были заданы: 1) ИИ чат-боту на основе GPT-3.5; а затем 2) человеку. После чего ИИ чат-бот – автор вопросов, оценил ответы и дал верное заключение о том, какая из двух групп ответов была дана ИИ чат-ботом и какая человеком.

Таким образом, результаты Обратного теста Тьюринга показали, что ИИ чат-боты способны, как минимум, не хуже людей ставить задачу лингвистической идентификации людей и машин, а потом и правильно проводить их идентификацию.

Для полноты картины следует упомянуть еще три теста, которые методически дальше отстоят от теста Тьюринга, но, как и три вышеназванных теста, позволяют идентифицировать человекоподобие интеллектуальных систем. Причем в этих тестах рассматриваемые аспекты человекоподобия выходят за рамки диалоговых реакций, лингвистических навыков и способности логично рассуждать. Здесь речь идет, во-первых, о способности к творческому

мышлению, оцениваемого по тестам Торранса [17]; во-вторых, об «общих» (наиболее «человеческих» и философских) областях мышления [15]; и, в-третьих, даже о понимании человеческих ценностей.

Результаты этих трех тестов, как и трех описанных выше, позволяют сделать вывод, что проходившие тестирование ИИ чат-боты обладают всеми способностями и интеллектуальными качествами, на выявление которых был направлен каждый из тестов. Иными словами, ИИ чат-боты класса GPT-4 по объему знаний, способности к рассуждению, владению разговорным языком и умению ориентироваться в представлениях людей о самих себе и своих системах ценностей, как минимум, не уступают людям по уровню совершенства этих качеств.

Такой вывод подтверждается авторами работ либо (по аналогии с образовательными и профессиональными тестами ИИ чат-ботов [24]) в количественной форме – путем представления результатов бенчмарков в диапазоне 95%+ от уровня людей [17], либо в качественной форме субъективного вывода жюри теста о том, что «рубеж, отделяющий его от сильного, или общего ИИ (AGI), можно считать уже пройденным» [15].

Следует отметить, что современная наука не располагает вескими основаниями для теоретических или эмпирических критериев различения разума «нормального, рационального человека» от «иррационального разума безумца». То есть невозможно, проведя тесты, сделать однозначный вывод – перед нами разумный или безумный человек. Здесь все слишком зыбко, условно и не точно, чтобы решать такие вопросы тестированием с бинарным вердиктом «да/нет» [8].

Тогда возникает вопрос: «Как трактовать, что тесты для выявления у ИИ конкретных свойств и способностей, свойственных разуму «нормальных, рациональных» людей, не только дают результаты, но и, все как один, дают положительные результаты»? Иными словами, если тестирование не способно решить задачу различения рациональности и разума от иррациональностии безумия людей, как же можно на основе тестирования делать вывод о наличии человекоподобного разума у некой нечеловеческой сущности, называемой нами ИИ? Да еще и делать из результатов тестов вывод о соизмеримости интеллектуального уровня разума людей и ИИ?

Этот парадокс наводит на мысль о существовании какого-то кардинального изъяна в проведении измерений и оценок при определении степени разумности (уровня интеллекта и т.д.) тестируемых ИИ чат-ботов. Тем более, что соображения о несовершенстве подходов при проведении измерений и оценок в области ИИ высказываются специалистами уже не первый год.

Несовершенство измерений и оценок

В октябре 2019 г. более 150 междисциплинарных экспертов обсудили вопросы проведения измерений и оценок в области ИИ в ходе закрытого семинара HAI-AI Index Workshop on Measurement in AI Policy: Opportunities and Challenges [32]. В отчете о семинаре [42], обобщающем 42 сделанных на нем доклада и связанных с ними дискуссий, авторы называют шесть основных проблем, присущих измерению прогресса и влияния ИИ. Все шесть проблем вытекают из отсутствия: 1) четко сформулированных определений для подлежащих оценке главных онтологических понятий в области ИИ и 2) надежных способов измерения и общепринятых методов оценки (качественных и количественных) главных онтологических понятий в области ИИ.

На практике это приводит к тому, что используемые при тестировании метрики, как правило, сосредоточены на легко измеряемых величинах, а не на реальном проявлении тестируемого феномена. Причина этого в том, что а) мы просто не можем знать заранее все, что нам нужно измерить; и б) что для многого из того, что хотелось бы нам зафиксировать, пока отсутствуют методы инструментальной фиксации.

В докладе Рейчел Томас из Университета Сан-Франциско, опираясь на совместную работу с Дэвидом Умински, показала, что оптимизация заданной метрики является центральным аспектом большинства современных подходов к ИИ [48]. Однако, чрезмерное внимание к метрикам приводит к манипуляциям, накруткам для

достижения целей исследователей, их близорукой ориентации на краткосрочные цели и другим неожиданным негативным последствиям. И это создает фундаментальную проблему в использовании и развитии ИИ. Чем большее значение придается при тестировании конкретным метрикам, тем более бесполезными они становятся.

Это противоречие отражено в законе Гудхарта: «Когда мера становится целью, она перестает быть хорошей мерой» [30]. Суть этого неформального закона в том, что, если показатель становится целевой функцией для проведения некой политики, прежние эмпирические закономерности, использующие данный показатель, перестают действовать.

Вот ставший классическим пример работы этого закона [38]. Весной 1902 г. французские колониальные чиновники в Ханое, опасаясь бубонной чумы, объявили войну нашествию крыс. Чиновники стимулировали крысоловов, предлагая вознаграждение за каждый доставленный труп. В последующие месяцы количество доставленных крысиных трупов росло в геометрической прогрессии, однако их популяция, казалось, не пострадала. По мере того, как кучи трупов росли и становились помехой, чиновники начали вознаграждать за доставку крысиных хвостов, а не целых животных. Город распространил свою систему поощрений на население в целом, пообещав вознаграждение в размере одного цента за каждый доставленный хвост. Жители быстро начали доставлять тысячи хвостов. Однако вскоре было замечено, что по городу снует все большее число бесхвостых крыс, которых, возможно, оставили в живых для размножения и, следовательно, снабжения новыми ценными хвостами. Хуже того, предприимчивые люди начали разводить крыс, выращивая хвосты для получения вознаграждения.

В этом примере, согласно закону Гудхарта, мера (крысиные трупы или хвосты) является оперативным показателем достижения некоторой цели (сокращение популяции крыс). Однако, когда мера становится целью, ее корреляция с этой целью уменьшается или, в крайних случаях, вообще исчезает, что приводит к непреднамеренным и часто неблагоприятным результатам. В приведенном примере популяция крыс в Ханое резко возросла, когда программа была прекращена: ставшие бесполезными крысы были выпущены на свободу в городе. То есть цель не только не была достигнута, но стало еще хуже.

В работе Йохана Джона и Оливера Браганца продемонстрировано, что «подобные Гудхарту» явления обнаружены и переоткрыты в широком диапазоне контекстов и масштабов: от централизованного управления до распределенных социальных систем, от измерений в области эволюционной конкуренции до измерений в области ИИ (см. таблицу 1 в [38]). И хотя физические механизмы варьируются от случая к случаю, существует несколько структурных особенностей, которые повторяются во всех кейсах. Это указывает на то, что сходство не является поверхностным, и в основе подобных явлений лежит явление так называемых «прокси-отказов», характерных для организации и динамики целеустремленного поведения в биологических и социальных масштабах.

Наиболее понятным и близким к рассматриваемой теме примером проявления закона Гудхарта является стандартизированное тестирование – например, тестирование при ЕГЭ. Ведь это тестирование, по идее, предназначенное для объективной оценки подготовки выпускников школ, теряет свою полезность, как только учителям предоставляется возможным преподавать тесты. Итог всем известен – мера становится целью и перестает быть хорошей мерой.

Похожая история случилась и в области ИИ. Как отмечал в своем выступлении знаменитый китайский писатель Тэд Чанг (зимой 2022 года он присоединился к Институту Санта-Фе в качестве стипендиата Миллера), ИИ чат-боты типа ChatGPT, по сути, представляют собой стандартизированные машины для сдачи тестов; все их развитие было формой машинного обучения на тренировочных данных с последующими настроечными испытаниями, проводимыми людьми [35]. В области ИИ так сложилось, что несколько десятилетий считалось, будто некие стандартизованные тесты смогут быть хорошим

способом измерения способностей ИИ. Но затем программисты ИИ нашли способ научить ИИ прохождению этих тестов. В результате, все развитие ИИ стало сводиться к тому, чтобы найти тест, прохождению которого программисты еще не научили ИИ.

С каждым новым тестом происходит одно и то же. Программисты определяют целевую функцию (функцию потерь или ошибки) для теста и создают машину, которая будет набирать высокие баллы на этом тесте. Такая наивная настойчивость в оптимизации является ошибочным фокусом, – что в тестах ЕГЭ, что в тестах ИИ. Ведь для тестовой оценки, дала ли школа хорошее образование или достиг ли ИИ уровня людей, нужен непредсказуемый тест, выполнению которого нельзя заранее научить. В противном случае, подобное тестирование ведет разработки в области ИИ к «ловушке Гудхарта» – используемые в тестах метрики, став целевыми функциями, перестают отражать прежние эмпирические закономерности, и ценность таких тестов стремится к нулю. Здесь возникает вопрос: «Существует ли возможность обойти «ловушку Гудхарта» для достоверного определения уровня развития ИИ и его достижения уровня AGI»?

Поиск путей обхода ловушки Гурхарда

Как показано в работе Йохана Джона и Оливера Браганца на широком спектре примеров из различных дисциплин (управление, ИИ, нейронауки, социальные науки, экономика, экология), появление ловушки Гудхарта практически неизбежно [38]. Всякий раз, когда стимуляция или отбор основаны на оптимизации несовершенной косвенной метрики основной цели, возникает давление, которое отталкивает косвенную метрику от цели, стремясь сделать эту метрику наихудшим приближением к цели. При этом, как показано в данной работе, сам факт использования для оптимизации несовершенной косвенной метрики порождает появление в процессе оптимизации ловушки Гудхарта.

К сожалению, в настоящее время прямыми и непосредственно измеримыми метриками оценки интеллекта, хоть биологических, хоть искусственных систем, современная наука не располагает. Как же тогда быть, если нельзя избежать появления ловушки Гудхарта в процессе мониторинга совершенствования ИИ до уровня AGI? Сейчас пока ясно лишь одно – путь в обход ловушки Гудхарта будет долгий и извилистый. Но кое-какие соображения на этот счет уже есть. Можно рассмотреть несколько предложений из совсем недавних работ, способных, так или иначе, содействовать более точным оценкам интеллектуального уровня ИИ в условиях неизбежности ловушки Гудхарта.

С. Мишра, Дж. Кларк, К. Рэймонд вполне резонно полагают, что в качестве первого необходимого (но недостаточного) шага следует зафиксировать основные онтологические понятия в области интеллекта [42]. Должно быть однозначно и недвусмысленно определено, не только наше понимание, что такое интеллект, ИИ и AGI, но и также:

1) что способствует прогрессу ИИ;
2) как использовать и совершенствовать библиометрические данные для анализа ИИ и его влияния на мир;
3) как измерять экономическое воздействие ИИ, особенно динамику рынка труда, а также взаимодействие с экономическим ростом и благосостоянием;
4) как измерять влияние ИИ на общество, в частности, на устойчивое экономическое развитие и потенциальные риски ИИ для разнообразия, прав человека и безопасности;
5) как измерять риски и угрозы уже развернутых систем искусственного интеллекта.

Патрик Батлер, Роберт Лонг и Эрик Элмоснино считают, что следовало бы разработать систему описываемых в вычислительных терминах «индикаторных свойств» основных онтологический понятий в области ИИ (в этой работе авторы рассматривают только понятия «сознание») [25]. Эти «индикаторные свойства» должны позволять оценивать системы ИИ на предмет соответствия им. Следующим шагом, следует проверить разработанную систему «индикаторных свойств» на всех основных теориях интеллекта [34], а также на всех наиболее

продвинутых из существующих больших языковых моделей.

В свою очередь, Т. Сейновски, хотя и допускает принципиальную возможность достижения большими языковыми моделями человекоподобного уровня AGI, предлагает пока забыть об AGI и попытках выявления у ИИ чат-ботов интеллекта человеческого уровня. Ибо выход на уровень AGI, как считает Т. Сейновски, невозможен без достижения AGA (*Artificial General Autonomy*) [7]. А поскольку у искусственного автономного агента AGA можно специфицировать, а потом и инструментально фиксировать характерные прямые измеряемые метрики, надобность в косвенных метриках до появления AGA отпадет. И тем самым удастся, на время достижения AGA, избежать ловушки Гудхарта.

Наиболее фундаментальные предложения заключаются в отказе от доминирующих способов изучения интеллекта, сфокусированных, в основном, на психометрии субъекта (человека или ИИ[3]) и тестах для оценки способности машин имитировать ответы людей. Авторы таких предложений считают, что после случившейся «революции ChatGPT» опора на психометрию и способность машин имитировать ответы людей для оценки прогресса ИИ уже не годится.

По мере усиления акцента на междисциплинарность в научных исследованиях ИИ наметился прогресс в создании теорий интеллекта и в перспективе – общей теории, основанной на первых принципах. Один из таких подходов описан в работе Майкла Э. Хохберга, где автор предлагает процесс концептуальной унификации определения интеллекта, охватывающего физическую, биологическую и искусственную сферы, что позволяет, по мнению автора, сформулировать общую теорию интеллекта [34]. А при наличии такой фундаментальной теории может стать возможным переход от косвенных метрик измерения уровня интеллекта к прямым.

Альтернативный подход был недавно представлен на конференции AGI-23. Его авторы в докладе Test and Evaluation First Principlesf or General Learning Systems предлагают кардинальную смену парадигмы инженерного подхода к совершенствованию ИИ-систем [52]. Эту новую парадигму авторы называют solution-method-agnostic engineering. И ее суть в том, что сейчас тестирование и оценка развития ИИ направлены не на ИИ-системы в целом, а лишь на их «подсистемы обучения». Тогда как ИИ-системы – это «системы систем» (включающие в себя, помимо подсистем обучения, также людей и операционное окружение), целью которых является поиск оптимального решения возникающих у них проблем.

Гипотезы

Ни один из известных автору подходов обхода «ловушки Гудхарта» (включая все упомянутые выше) не работает «здесь и сейчас». Но это не останавливает разработку многих десятков проектов по всему миру, декларируемая цель которых – создание AGI. Три года назад, согласно данным М. Фитцджеральд, А. Бодди и С.Д. Баум, таких проектов уже проводились 72 в 37-и странах [27]. А в настоящем 2023 году проекты всех (!) ведущих мировых разработчиков ИИ (от OpenAI и Deep Mind до запрещенной в России Meta и незапрещенной Baidu), согласно их официальным объявлениям, направлены на создание AGI.

Таким образом, ситуация весьма странная. Все делают AGI. И при этом никаких надежных способов определить, что уровень AGI достигнут, в мире не существует. По нашему мнению, существует лишь один способ преодоления ловушки Гудхарта «здесь и сейчас». Он подробно изложен в тетралогии наших статей, озаглавленной «Теория относительности интеллекта: биологического и машинного», где предпринята попытка детально описать, что именно и почему, на наш взгляд, не позволяет машине достичь уровня биологического интеллекта, а эволюции машин достичь уровня биологической эволюции [12; 10; 11; 13].

[3] В данной статье мы не приводим примеров конкретных психометрических тестов для ИИ, отсылая интересующихся читателей к работам [37;44]

В этих исследованиях на основе работ и подхода У. Эшби, С. Кауффмана, А. Роли и Й. Йегера, были сформулированы следующие гипотезы:

1. У человека существует интеллектуальная способность, позволяющая выявлять и актуализировать аффордансы смежного возможного (что не является вычислимой задачей из-за неопределенности последнего).
2. Данная способность обеспечивается механизмом серендипности[4] к изобретениям, в основе которого лежат многоплановые способности людей осуществлять особые когнитивные (в первую очередь, аналитические и темпоральные) и внекогнитивные (социально-сетевые) действия.
3. Из невычислимости механизма серендипности может следовать, что это акт непонятийного, непосредственного «узрения», «постижения», и этот акт познания совпадает с актом, порождающим действительность (работа такого невычислимого механизма, скорее всего, основана на законах неклассической физики).

Заключение

Если данные гипотезы подтвердятся, то никакие психометрические тесты или тесты на способность машин имитировать ответы людей для проверки достижения ИИ-системами уровня AGI не потребуются. Ибо сам факт обнаружения у ИИ способности выявлять и актуализировать аффордансы смежного возможного может стать единственным (необходимым и достаточным) подтверждением достижения ИИ-системой уровня AGI. Причем будет даже неважен механизм в основе такой способности ИИ – механизм серендипности или что-то иное, недоступное человеческому пониманию, но доступное большим языковым моделям.

Обсуждение способов фиксации у ИИ-систем способности выявлять и актуализировать аффордансы смежного возможного имеет смысл вынести за пределы этой статьи, и так уже весьма объемной. Замечу лишь, что, во-первых, в работах Стюарта Кауффмана данный вопрос уже рассматривается. И, во-вторых, важно отметить, что эффективным инструментом обнаружения у больших языковых моделей способности выявлять и актуализировать аффордансы смежного возможного могут стать ИИ чат-боты, работающие на базе самих этих моделей. Что может стать решением проблемы «ловушки Гудхарта» и геймченджером на пути создания AGI.

Литература:

1. Blackiston D., Kriegman S., Bongard J., Levin M. Biological Robots: Perspectives on an Emerging Interdisciplinary Field // Soft Robotics. 2023. Pp. 674-686. https://www.liebertpub.com/doi/full/10.1089/soro.2022.0142
2. Gordijn D., Have H. ChatGPT: evolution or revolution? // Medicine, Health Care and Philosophy. 2023. V. 26. Pp. 1-2.https://link.springer.com/article/10.1007/s11019-023-10136-0
3. Gottfredson L. Mainstream science on intelligence: An Editorial With 52 Signatories, History, and Bibliography // Intelligence. V.24. Issue 1. 1997, Pp. 13-23. http://www1.udel.edu/educ/ gottfredson/reprints/1997mainstream.pdf
4. Hayes P., Ford K. Turing Test Considered Harmful // IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence. 1995. V.1. Pp.972-977. https://dl.acm.org/doi/10.5555/1625855.1625981
5. Hutson M. Rules to keep AI in check: nations carve different paths for tech regulation // Nature. 2023. V.620. Pp. 260-263. https://www.nature.com/articles/d41586-023-02491-y

[4]Серендипность – инстинктивная (интуитивная) прозорливость (англ. *serendipity*) – термин, происходящий из английского языка и обозначающий способность, делая глубокие выводы из случайных наблюдений, находить то, чего не искал намеренно. Среди характерных примеров серендипности называют открытие рентгеновского излучения Вильгельмом Рентгеном, а также открытие взаимосвязи электричества и магнетизма Эрстедом (Википедия).

6. Jefferson G. The Mind of Mechanical Man // British Medical Journal. 1949. V.1. Pp.4616. https://doi.org/10.1136/bmj.1.4616.1105
7. Sejnowski T.J. Large Language Models and the Reverse Turing Test // Neural Computation. 2023. V.35. Issue 3. Pp. 309-342. https://doi.org/10.1162/neco_a_01563
8. Sterzer P. Die Illusion der Vernunft: Warum wir von unseren Überzeugungen nicht zu überzeugt sein sollten / Neuestes aus Hirnforschung und Psychologie. Ullstein, Berlin. 2022. https://www.amazon.de/Die-Illusion-Vernunft-%C3%9Cberzeugungen-Hirnforschung/dp/355020132X
9. Turing A.M. Computing Machinery and Intelligence // Mind. 1950. V. LIX. Issue 236. Pp.433-460. https://doi.org/10.1093/mind/LIX.236.433

Интернет ресурсы:

10. Карелов С. Аффорданс – ключевое свойство интеллектуального агента // Малоизвестное интересное. 2021. https://dzen.ru/a/YYzplIlQGSDExDYc
11. Карелов С. Невычислимая тень будущего // Малоизвестное интересное. 2021. https://dzen.ru/a/YZTzizvaBzV1UFII
12. Карелов С. Открыта теория относительности интеллекта: биологического и машинного // Малоизвестное интересное. 2021. https://dzen.ru/a/YYkdZ6xat1ZwQZjG
13. Карелов С. Серендипность – чудо увидеть цель в море случайностей // Малоизвестное интересное. 2021. https://dzen.ru/a/YadVB3jkREoIOaZL
14. Карелов С. Фиаско 2023. Характер сосуществования двух типов разума, зависит от их взаимопонимания // Малоизвестное интересное. 2023. https://dzen.ru/media/the_world_is_not_easy/fiasko-2023-6486f59dbfaf86243ed3c4b4
15. Эпштейн М. Искусственный и человеческий интеллекты: новый эксперимент по их сопоставлению // Сноб. 2023. https://snob.ru/profile/27356/blog/3059715/
16. AI pioneer Yoshua Bengio: Governments must move fast to «protect the public» // Financial Times. 2023. https://www.ft.com/content/b4baa678-b389-4acf-9438-24ccbcd4f201
17. AI tests into top 1% for original creative thinking // Science Daily. 2023. https://www.sciencedaily.com/releases/2023/07/230705154051.htm
18. AI21 Labs concludes largest Turing Test experiment to date // Проект AI21 Labs. 2023. https://www.ai21.com/blog/human-or-not-results?utm_source=superhuman.beehiiv.com&utm_medium=newsletter&utm_campaign=ai21-labs-concludes-largest-turing-test-experiment-to-date
19. Artificial Intelligence Law, Model Law v. 1.0. // Digi China Project. 2023. https://digichina.stanford.edu/work/translation-artificial-intelligence-law-model-law-v-1-0-expert-suggestion-draft-aug-2023/
20. Barrett C., Boyd B., Burzstein E., Carlini N. et al. Identifying and Mitigating the Security Risks of Generative AI. 2023. https://arxiv.org/pdf/2308.14840.pdf
21. Benizri I., Evers A., Mercer S.T., Jessani A. A Comparative Perspective on AI Regulation // Lawfare. 2023. https://www.lawfaremedia.org/article/a-comparative-perspective-on-ai-regulation
22. Bongard J., Levin M. There's Plenty of Room Right Here: Biological Systems as Evolved, Overloaded, Multi-Scale Machines // Biomimetics. 2023. V.8. Pp.110. https://doi.org/10.3390/biomimetics8010110
23. Bremmer I., Suleyman M. The AI Power Paradox // Foreign Affairs. 2023. https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox
24. Bubeck S., Chandrasekaran V., Eldan R, Gehrke J., Horvitz E., Kamar E. et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4 // Cornell University. 2023. https://arxiv.org/abs/2303.12712
25. Butlin P., Long R., Elmoznino E. Consciousness in Artifcial Intelligence: Insights from the Science of Consciousness. 2023. https://arxiv.org/abs/2308.08708

26. Chollet F. On the measure of intelligence. 2019. https://arxiv.org/abs/1911.01547
27. Fitzgerald McK., Boddy A., Baum S.D. A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy // Global Catastrophic Risk Institute Working Paper. 2020. https://gcrinstitute.org/papers/055_agi-2020.pdf
28. Goncalves B. Can machines think? The controversy that led to the Turing test // AI & SOCIETY. 2022. DOI: 10.1007/s00146-021-01318-6
29. Goncalves B. Irony with a Point: Alan Turing and His Intelligent Machine Utopia // Philosophy&Technology. 2023. https://doi.org/10.1007/s13347-023-00650-7
30. Goodhart's law // Wikipedia. https://en.wikipedia.org/wiki/Goodhart%27s_law
31. Guterres A. Artificial Intelligence: Opportunities and Risks for International Peace and Security // UN Security Council. 2023. 9381st Meeting. https://media.un.org/en/asset/k1j/k1ji81po8p?fbclid=IwAR1Zq6X7baQzlnpVBhgzPfW-wOLtRfUHv61uz35wnBZJE93lsGQdl257RbDk
32. HAI-AI Index Workshop on Measurement in AI Policy: Opportunities and Challenges // Stanford University. Human-Centered Artificial Intelligence. 2019. https://hai.stanford.edu/hai-ai-index-workshop-measurement-ai-policy-opportunities-and-challenges-0
33. Heaven W.D. Geoffrey Hinton tells us why he's now scared of the tech he helped build // MIT Technology Review. 2023. https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/
34. Hochberg M.E. A Theory of Intelligences: Concepts, Models, Implications. 2023. https://arxiv.org/abs/2308.12411
35. ICLR 2022. From Cells to Societies - Collective Learning across Scales workshop. https://sites.google.com/view/collective-learning
36. Is ChatGPT the Start of the AI Revolution? // Bloomberg. 2022. https://www.bloomberg.com/opinion/articles/2022-12-09/is-chatgpt-the-start-of-the-ai-revolution
37. Jiang G., Xu M.. Evaluating and Inducing Personality in Pre-trained Language Models. 2023. https://arxiv.org/pdf/2206.07550.pdf
38. John Y., Braganza O. Dead rats, dopamine, performance metrics, and peacock tails: proxy failure is an inherent risk in goal-oriented systems // Behavioral and Brain Sciences, 2023. https://doi.org/10.1017/S0140525X23002753
39. Legg S. Machine super intelligence // Doctoral Dissertation submitted to the Faculty of Informatics of the University of Lugano in partial fulfillment of the requirements for the degree of Doctor of Philosophy. 2008. https://www.vetta.org/documents/Machine_Super_Intelligence.pdf
40. Legg S., Hutter M. Universal intelligence: A definition of machine intelligence // Minds and machines (2007). https://arxiv.org/abs/0712.3329
41. McCoy J.P., Ullman T.D. A Minimal Turing Test // The Journal of Experimental Social Psychology. 2018. V.79. Pp.1-8. https://doi.org/10.1016/j.jesp.2018.05.007
42. Mishra S., Clark J., Perrault C.R. Measurement in AI Policy: Opportunities and Challenges. 2020. https://arxiv.org/abs/2009.09071
43. Pause Giant AI Experiments: An Open Letter // Future of Life Institute. 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/
44. Pellert M., Lechner C., Wagner C., Rammstedt B., Strohmaier M. AI Psychometrics: Assessing the psychological profles of large language models through psychometric inventories. 2023. https://psyarxiv.com/jv5dt/
45. Planning for AGI and beyond // OpenAI. 2023. https://openai.com/blog/planning-for-agi-and-beyond
46. Shontell A. ChatGPT shows that the A.I. revolution has arrived // Fortune. 2023. https://fortune.com/2023/01/25/chatgpt-ai-revolution-february-march-2023-issue/
47. Statement on AI Risk // Center for AI Safety. 2023. https://www.safe.ai/statement-on-ai-risk
48. Thomas R.L., Uminsky D. Reliance on Metrics is a Fundamental Challenge for AI. 2019. https://doi.org/10.48550/arXiv.2002.08512

49. Turing test // Wikipedia. https://en.wikipedia.org/wiki/Turing_test#CITEREFTuring1950
50. West D.M. Senate hearing highlights AI harms and need for tougher regulation // The Brookings Institution. 2023. https://www.brookings.edu/articles/senate-hearing-highlights-ai-harms-and-need-for-tougher-regulation/
51. Xu G., Liu J., Yan M., Xu H. et al. Values: Measuring the Values of Chinese Large Language Models from Safety to Responsibility. 2023. https://doi.org/10.48550/arXiv.2307.09705

Видео ресурсы:

52. Cody T., Hahm C., Goertzel B. Test and evaluation first principles for general learning systems. 2023. AGI-23 Workshop. https://www.youtube.com/watch?v=Hfai7Plzg4M

THE "GOODHART'S TRAP" FOR AGI: THE PROBLEM OF COMPARATIVE ANALYSIS OF ARTIFICIAL INTELLIGENCE AND HUMAN INTELLIGENCE

© Sergey V. Karelov
Ph.D., independent researcher and popularizer of science,
host of the author's channel "Little-known interesting"

The "ChatGPT revolution" that took place in 2023 dramatically reduced the experts' forecast estimates of the time separating us from the creation of artificial intelligence that is intellectually as good as any human being (AGI). At the same time, paradoxically, the existing testing methods are not yet able to diagnose with any reliability the achievement by AI-systems of the AGI level. This paper discusses the issue of overcoming this problem of imperfection of modern methods of testing AI-systems. In particular, the hypothesis of the fundamental impossibility of solving the problem of AGI detection both by means of psychometric tests and methods of assessing the ability of machines to imitate human responses due to the so-called "Goodhart's trap" for AGI is presented. A number of proposals for circumventing the "Goodhart's trap" for AGI by means of methods proposed in recent research works, taking into account the first results of the ChatGPT revolution, are considered. In the last part of the paper, a set of three heuristic hypotheses is formulated, which, if true, can radically solve the problem of the "Goodhart's trap" for AGI and thus become a gamechanger on the way to creating AGI.

***Keywords*:** intelligence, artificial intelligence, AGI, AI Testing, Goodhart's law, Turing test, problem with metrics, psychometrics

REFERENCE

1. Blackiston D., Kriegman S., Bongard J., Levin M. (2023). Biological Robots: Perspectives on an Emerging Interdisciplinary Field // Soft Robotics. Pp. 674-686. https://www.liebertpub.com/doi/full/10.1089/soro.2022.0142

2. Gordijn D., Have H. (2023). ChatGPT: evolution or revolution? // Medicine, Health Care and Philosophy. V. 26. Pp. 1-2. https://link.springer.com/article/10.1007/s11019-023-10136-0
3. Gottfredson L. (1997). Mainstream science on intelligence: An Editorial With 52 Signatories, History, and Bibliography // Intelligence. V.24. Issue 1. Pp. 13-23. http://www1.udel.edu/educ/gottfredson/reprints/1997mainstream.pdf
4. Hayes P., Ford K. (1995). Turing Test Considered Harmful // IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence. V.1. Pp.972-977. https://dl.acm.org/doi/10.5555/1625855.1625981
5. Hutson M. (2023). Rules to keep AI in check: nations carve different paths for tech regulation // Nature. V.620. Pp. 260-263. https://www.nature.com/articles/d41586-023-02491-y
6. Jefferson G. The Mind of Mechanical Man // British Medical Journal. 1949. V.1. Pp.4616. https://doi.org/10.1136/bmj.1.4616.1105
7. Sejnowski T.J. (2023). Large Language Models and the Reverse Turing Test // Neural Computation. V.35. Issue 3. Pp. 309-342. https://doi.org/10.1162/neco_a_01563
8. Sterzer P. (2022). Die Illusion der Vernunft: Warum wir von unseren Überzeugungen nicht zu überzeugt sein sollten / Neuestes aus Hirnforschung und Psychologie. Ullstein, Berlin. https://www.amazon.de/Die-Illusion-Vernunft-%C3%9Cberzeugungen-Hirnforschung/dp/355020132X
9. Turing A.M. (1950). Computing Machinery and Intelligence // Mind. V. LIX. Issue 236. Pp.433-460. https://doi.org/10.1093/mind/LIX.236.433

Internet resources:

10. Karelov S. (2021). Affordans – klyuchevoe svojstvo intellektual'nogo agenta [*Affordance – the key property of an intelligent agent*] // Maloizvestnoe interesnoe [*Little-known interesting*]. https://dzen.ru/a/YYzplIlQGSDExDYc
11. Karelov S. (2021). Nevychislimaya ten' budushchego [*The incalculable shadow of the future*] // Maloizvestnoe interesnoe [*Little-known interesting*]. https://dzen.ru/a/YZTzizvaBzV1UFII
12. Karelov S. (2021). Otkryta teoriya otnositel'nosti intellekta: biologicheskogo i mashinnogo [*The theory of relativity of intelligence: biological and machine is open*] // Maloizvestnoe interesnoe [*Little-known interesting*]. https://dzen.ru/a/YYkdZ6xat1ZwQZjG
13. Karelov S. (2021). Serendipnost' – chudo uvidet' cel' v more sluchajnostej [*Serendipity is a miracle to see a goal in a sea of accidents*] // Maloizvestnoe interesnoe [*Little–known interesting*]. https://dzen.ru/a/YadVB3jkREoIOaZL
14. Karelov S. (2023). Fiasko 2023. Harakter sosushchestvovaniya dvuh tipov razuma, zavisit ot ih vzaimoponimaniya [*Fiasco 2023. The nature of the coexistence of two types of mind depends on their mutual understanding*] // Maloizvestnoe interesnoe [*Little-known interesting*]. https://dzen.ru/media/the_world_is_not_easy/fiasko-2023-6486f59dbfaf86243ed3c4b4
15. Epshtejn M. (2023). Iskusstvennyj i chelovecheskij intellekty: novyj eksperiment po ih sopostavleniyu [*Artificial and human intelligences: a new experiment to compare them*] // Snob [*Snob*]. https://snob.ru/profile/27356/blog/3059715/
16. AI pioneer Yoshua Bengio: Governments must move fast to «protect the public» // Financial Times. 2023. https://www.ft.com/content/b4baa678-b389-4acf-9438-24ccbcd4f201
17. AI tests into top 1% for original creative thinking // Science Daily. 2023. https://www.sciencedaily.com/releases/2023/07/230705154051.htm
18. AI21 Labs concludes largest Turing Test experiment to date // Проект AI21 Labs. 2023. https://www.ai21.com/blog/human-or-not-results?utm_source=superhuman.beehiiv.com&utm_medium=newsletter&utm_campaign=ai21-labs-concludes-largest-turing-test-experiment-to-date
19. Artificial Intelligence Law, Model Law v. 1.0. // Digi China Project. 2023. https://digichina.stanford.edu/work/translation-artificial-

intelligence-law-model-law-v-1-0-expert-suggestion-draft-aug-2023/
20. Barrett C., Boyd B., Burzstein E., Carlini N. et al. (2023). Identifying and Mitigating the Security Risks of Generative AI. https://arxiv.org/pdf/2308.14840.pdf
21. Benizri I., Evers A., Mercer S.T., Jessani A. (2023). A Comparative Perspective on AI Regulation // Lawfare. https://www.lawfaremedia.org/article/a-comparative-perspective-on-ai-regulation
22. Bongard J., Levin M. (2023). There's Plenty of Room Right Here: Biological Systems as Evolved, Overloaded, Multi-Scale Machines // Biomimetics. V.8. Pp.110. https://doi.org/10.3390/biomimetics8010110
23. Bremmer I., Suleyman M. (2023). The AI Power Paradox // Foreign Affairs. https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox
24. Bubeck S., Chandrasekaran V., Eldan R, Gehrke J., Horvitz E., Kamar E. et al. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4 // Cornell University. https://arxiv.org/abs/2303.12712
25. Butlin P., Long R., Elmoznino E. (2023). Consciousness in Artifcial Intelligence: Insights from the Science of Consciousness. https://arxiv.org/abs/2308.08708
26. Chollet F. (2019). On the measure of intelligence. https://arxiv.org/abs/1911.01547
27. Fitzgerald McK., Boddy A., Baum S.D. (2020). A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy // Global Catastrophic Risk Institute Working Paper. https://gcrinstitute.org/papers/055_agi-2020.pdf
28. Goncalves B. (2022). Can machines think? The controversy that led to the Turing test. // AI & SOCIETY. DOI: 10.1007/s00146-021-01318-6
29. Goncalves B. (2023). Irony with a Point: Alan Turing and His Intelligent Machine Utopia // Philosophy&Technology. https://doi.org/10.1007/s13347-023-00650-7
30. Goodhart's law // Wikipedia. https://en.wikipedia.org/wiki/Goodhart%27s_law
31. Guterres A. (2023). Artificial Intelligence: Opportunities and Risks for International Peace and Security // UN Security Council. 9381st Meeting. https://media.un.org/en/asset/k1j/k1ji81po8p?fbclid=IwAR1Zq6X7baQzlnpVBhgzPfW-wOLtRfUHv61uz35wnBZJE93lsGQdl257RbDk
32. HAI-AI Index Workshop on Measurement in AI Policy: Opportunities and Challenges // Stanford University. Human-Centered Artificial Intelligence. 2019. https://hai.stanford.edu/hai-ai-index-workshop-measurement-ai-policy-opportunities-and-challenges-0
33. Heaven W.D. (2023). Geoffrey Hinton tells us why he's now scared of the tech he helped build // MIT Technology Review. https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/
34. Hochberg M.E. (2023). A Theory of Intelligences: Concepts, Models, Implications. https://arxiv.org/abs/2308.12411
35. ICLR 2022. From Cells to Societies – Collective Learning across Scales workshop. https://sites.google.com/view/collective-learning
36. Is ChatGPT the Start of the AI Revolution? // Bloomberg. 2022. https://www.bloomberg.com/opinion/articles/2022-12-09/is-chatgpt-the-start-of-the-ai-revolution
37. Jiang G., Xu M. (2023). Evaluating and Inducing Personality in Pre-trained Language Models. https://arxiv.org/pdf/2206.07550.pdf
38. John Y., Braganza O. (2023). Dead rats, dopamine, performance metrics, and peacock tails: proxy failure is an inherent risk in goal-oriented systems // Behavioral and Brain Sciences. https://doi.org/10.1017/S0140525X23002753
39. Legg S. (2008). Machine super intelligence // Doctoral Dissertation submitted to the Faculty of Informatics of the University of Lugano in partial fulfillment of the requirements for the degree of Doctor of Philosophy. https://www.vetta.org/documents/Machine_Super_Intelligence.pdf
40. Legg S., Hutter M. (2007). Universal intelligence: A definition of machine intelligence // Minds and machines. https://arxiv.org/abs/0712.3329

41. McCoy J.P., Ullman T.D. (2018). A Minimal Turing Test // The Journal of Experimental Social Psychology. V.79. Pp.1-8. https://doi.org/10.1016/j.jesp.2018.05.007
42. Mishra S., Clark J., Perrault C.R. (2020). Measurement in AI Policy: Opportunities and Challenges. https://arxiv.org/abs/2009.09071
43. Pause Giant AI Experiments: An Open Letter // Future of Life Institute. 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/
44. Pellert M., Lechner C., Wagner C., Rammstedt B., Strohmaier M. (2023). AI Psychometrics: Assessing the psychological profles of large language models through psychometric inventories. https://psyarxiv.com/jv5dt/
45. Planning for AGI and beyond // OpenAI. 2023. https://openai.com/blog/planning-for-agi-and-beyond
46. Shontell A. (2023). ChatGPT shows that the A.I. revolution has arrived // Fortune. https://fortune.com/2023/01/25/chatgpt-ai-revolution-february-march-2023-issue/
47. Statement on AI Risk // Center for AI Safety. 2023. https://www.safe.ai/statement-on-ai-risk
48. Thomas R.L., Uminsky D. (2019). Reliance on Metrics is a Fundamental Challenge for AI. https://doi.org/10.48550/arXiv.2002.08512
49. Turing test // Wikipedia. https://en.wikipedia.org/wiki/Turing_test#CITEREFTuring1950
50. West D.M. (2023). Senate hearing highlights AI harms and need for tougher regulation // The Brookings Institution. https://www.brookings.edu/articles/senate-hearing-highlights-ai-harms-and-need-for-tougher-regulation/
51. Xu G., Liu J., Yan M., Xu H. et al. (2023). Values: Measuring the Values of Chinese Large Language Models from Safety to Responsibility. https://doi.org/10.48550/arXiv.2307.09705

Video resources:

52. Cody T., Hahm C., Goertzel B. (2023). Test and evaluation first principles for general learning systems. AGI-23 Workshop. https://www.youtube.com/watch?v=Hfai7Plzg4M