Философия науки и техники
2023. Т. 28. № 2. С. 156–169
УДК 004.8

Philosophy of Science and Technology
2023, vol. 28, no. 2, pp. 156–169
DOI: 10.21146/2413-9084-2023-28-2-156-169

НАУЧНАЯ ЖИЗНЬ

А.Ю. Алексеев, А.А. Ващенко, А.С. Зайкова

Парадоксы и противоречия искусственного интеллекта: 90 лет первой теореме К. Гёделя о неполноте и 60 лет аргументу Дж. Лукаса*

Алексеев Андрей Юрьевич – доктор философских наук, профессор философского факультета. Государственный академический университет гуманитарных наук. Российская Федерация, 119049, г. Москва, Мароновский переулок, д. 26; профессор департамента механики и процессов управления. РУДН; e-mail: aa65@list.ru

Ващенко Александр Александрович – аспирант. Физтех-школа МФТИ. Российская Федерация, 141701, Московская обл., г. Долгопрудный, Институтский пер., д. 9. e-mail: alexanderva@mail.ru

Зайкова Алина Сергеевна – кандидат философских наук, научный сотрудник. ИФПР СО РАН. Российская Федерация, 630090, г. Новосибирск, ул. Николаева, д. 8; e-mail: zaykova.a.s@gmail.com

Сегодня чрезвычайно актуальным оказалось изучение вопроса различения – какая доля работы в произведенном артефакте принадлежит человеку, а какая – компьютеру. В этом направлении важной является разработка компьютерных анализаторов, компьютерных генераторов текстов, картин, музыки, мультипликаций, видео. Фундаментальный теоретико-алгоритмический статус подобным генераторам задает тест Гёделя – Лукаса – Пенроуза. В 1961 г. профессор Оксфорда Дж. Лукас на основании второй теоремы Гёделя сделал вывод о принципиальном превосходстве человеческого разума над всевозможными вычислительными системами. Р. Пенроуз, поддержав такую позицию, заявил, что человеческое сознание не является алгоритмическим и выходит за пределы вычислимости. Этот аргумент вызвал обширную дискуссию, касающуюся его философских оснований, математической обоснованности, физических и нейрофизиологических объяснений, психологической доверительности, техно-

* Статья подготовлена в рамках государственного задания ГАУГН «Цифровизация и формирование современного информационного общества: когнитивные, экономические, политические и правовые аспекты». Регистрационный номер НИОКТР123022000042-0. Код темы FZNF-2023-0004. Регистрационный номер темы 1022040800826-5-5.2.1;6.3.1;5.9.1.

логической реализации. Спор о важности и альтернативных способах интерпретации аргумента не прекращается, составляя одно из базовых положений философии и методологии искусственного интеллекта и когнитивной науки. В октябре 2021 г. НСМИИ РАН провел Всероссийский симпозиум «Искусственный интеллект: парадоксы и противоречия», в ходе которого были изучены логико-философские, формально-вычислительные и культурно-антропологические аспекты аргумента Гёделя-Лукаса, а также показаны возможные пути его трансформации и развития, такие как тест Гёделя – Лукаса – Ватта.

Ключевые слова: философия и методология искусственного интеллекта, теорема Гёделя, аргумент Лукаса, аргумент Гёделя – Лукаса – Пенроуза, комплексный тест Тьюринга, инвертированный тест Тьюринга, тест Ватта, GPT, аргумент Гёделя – Лукаса – Ватта

Введение

Совсем недавно, в 2021 г., исполнилось девяносто лет первой теореме Гёделя о неполноте и шестьдесят лет аргументу Лукаса, основанному на этой теореме. В исследованиях ИИ этот аргумент широко употребляется как *аргумент Гёделя – Лукаса – Пенроуза*. Данный аргумент был всесторонне изучен в рамках всероссийского симпозиума «Искусственный интеллект: парадоксы и противоречия», который состоялся 13 октября 2021 г. Мероприятие было организовано Научным советом при Президиуме Российской академии наук по методологии искусственного интеллекта и когнитивных исследований (НСМИИ РАН) на базе Института философии и права Сибирского отделения РАН в рамках XIX Международной научной конференции молодых ученых «Актуальные проблемы гуманитарных и социальных исследований», г. Новосибирск, Россия. Председатель программного комитета – проф. В.В. Целищев, руководитель организационного комитета – А.Ю. Алексеев, уч. секретарь – А.С. Зайкова.

Обзор дискуссии

В ходе дискуссии были выявлены следующие концептуальные положения.

Лекторский В.А. (д.филос.н., проф., академик РАН, председатель НСМИИ РАН, г. Москва) отметил, что аргументация Лукаса выявляет сходство и различия человеческого и машинного мышления. В свою очередь, это важно не только для развития человеко-машинных систем, но и для прояснения природы мышления самого человека, а это является классической философской проблемой.

Целищев В.В. (д.филос.н., проф., науч. рук. Института философии и права Сибирского отделения РАН, г. Новосибирск) очертил историю споров вокруг теорем Гёделя, которые начались с выходом в 1961 г. статьи профессора Оксфорда Джона Лукаса «Разум, машина, Гёдель». После выхода статьи автор подвергся обвинениям в отсутствии математической логики и многочисленных ошибках. В частности, философ Барух Бенасерраф, заведующий кафедрой в Принстоне, заметил опасность подхода Лукаса относительно устоявшихся

на тот момент механистических идей и опубликовал ряд критических статей. Долгие годы продолжался спор между механицистами и менталистами. Большинство составляли механицисты.

В полемику в 1989-м г. вступил Роджер Пенроуз с книгой «Новый ум короля» [Пенроуз, 2003], который показал, что способность человека по постижению истины выше, чем у компьютера. Роджер Пенроуз, британский физик и математик, исходил в своих заключениях из идей Тьюринга, из идей алгоритмизации. Суть посыла Лукаса на самом деле была довольно простой: раздельно как бы существует сама идея истинности и отдельно – доказательство истинности. То есть истина и ее доказательность – это разные сущности. Скандал разразился с опубликованием расшифрованных записок самого Курта Гёделя, в которых обнаружилось, что Гёдель и сам придерживался подходов Лукаса и Пенроуза. Соломон Феферман [Feferman, 1995] поставил точку в споре – конечно, у Лукаса есть масса математических ошибок и неточностей, но ни одна из этих ошибок не влияет на суть аргумента. Тогда многие участники споров замолчали. Далее Р. Пенроуз написал вторую книгу – «Тень разума» [Пенроуз, 2005], где выдвигалась более интересная и сложная идея. Ранее все доказательства были основаны на идее непротиворечивости, а теперь Р. Пенроуз предложил рассматривать доказательства, основываясь на обоснованности. Любая доказуемая вещь является истинной. Когда мы сравниваем человека и машину, то мышление человека является обоснованным, а мышление машины – непротиворечивым.

Возникла следующая ситуация. Теорема Гёделя утверждает, что есть истины, которые недоказуемы. Но тогда есть вопрос – а как мы в принципе узнаем, что есть истина, как мы ее определяем? То есть вся гёделевская машинерия сродни некой мистике? Если подводить итог, кто умнее, человек или машина, то ответ – ничья. То есть аргумент Гёделя выступает в форме «дизъюнкции Гёделя»: либо человеческий ум превосходит все возможности машины, либо существуют абсолютно неразрешимые предложения. Нейросети и искусственный интеллект – это совсем не то и не так, как думает живой человек. При сравнивании человека и машины неизбежно возникают вопросы: в какой мере мы вправе использовать математические построения для разрешения этого дискурса? Какой философский смысл у математических теорем? (см. подробнее [Целищев, 2021]).

Ушаков Д.В. (академик РАН, д.псих.н., директор Института психологии РАН, г. Москва) обозначил психологические аспекты аргумента Гёделя – Лукаса – Пенроуза. Психологический подход к обсуждаемой проблеме – это как мы можем описывать естественный интеллект с точки зрения искусственного? Когда мы говорим о моделях когнитивной психологии, то мы говорим о моделях естественного интеллекта, которые работают по принципам вычислительных систем. Этот подход сейчас доминирует. Есть практика психологии, и научно респектабельны представления психики в виде вычислительных моделей. Виталий Валентинович считает, что в споре ничья, так что от итогов этого спора зависит и направление психологии. Второй аспект – анализ личности самого Гёделя, он сам является некой психологической загадкой и психологическим символом. Он, конечно, был удивительным человеком, который умел

схватывать глубинные вещи, которые идут в противофазе. Креативными могут быть только чудаки. Жаль, что мы потеряли и гуманитарного ученого в его лице, ведь он в том числе занимался вопросам религии и мироздания.

Хлебалин А.В. (к.филос.н., зам. директора по научной работе Института философии и права СО РАН, главный редактор журнала «Философия науки», г. Новосибирск) выступил с докладом «Сравнивая человека и машину…». Он отметил, что Арнон Аврон, самый яростный критик подхода Лукаса – Пенроуза, заметил, что «аргументы Лукаса – Пенроуза были опровергнуты достаточно давно, но при этом нельзя не признать, что дискуссия вокруг казалось бы установленных результатов продолжает быть крайне острой». Спор вокруг аргумента Лукаса – Пенроуза напоминает спор вокруг онтологического доказательства Бога. Сравнение человека и машины обострилось с развитием современных технологий. Да, машины играют в шахматы и Го и выигрывают у человека, но существует ли нейтральный (по отношению и к игрокам, и к машине) критерий успеха? Может ли машина опознать истинность? Альберт Виссер отмечал, что аргумент Гёделя – Лукаса – Пенроуза наделяет состязание между человеком и машиной нейтральным критерием успеха при должной степени формулировки. Во всех традиционных соревнованиях человека и машины мы находим очевидную интенсиональную нагруженность каждого компонента в том плане, что состязаются не человек с машиной, а человек с программой, с вычислительными возможностями и с интерпретацией. Пусть есть абстрактный идеализированный человек и есть конкретная машина. Если есть производство истинных предложений у машины, то, рассматривая их, мы неминуемо приписываем интенсиональность, а это приводит к утрате нейтральности. Таким образом, размышления на тему аргумента Гёделя – Лукаса – Пенроуза и сравнение человека и машины помогают нам понять природу самого человека и только на основе этого понять машину. Отвечая на вопрос об алгоритмизации шахматной программы на примере современных систем, которые не знают и не строят алгоритмов ходов, а оценивают варианты вероятности позиции и хода, не понимая, во что именно они играют, А.В. Хлебалин отметил, что создание вычислительных систем проясняет возможности нашего мозга и они, действительно, приближаются к подходам человека. Отвечая на вопрос о творчестве в исполнении машины, докладчик поставил вопросы: в какой степени мы в принципе можем считать творчеством продукт машины? А что такое творчество в человеческом смысле? Дает ли дедукция новое знание? Ответов нет, и споры продолжаются.

Родин А.В. (д.филос.н., доц. СПбГУ, г. Санкт-Петербург) в докладе «О понятии кибернетической машины у Лукаса» посчитал ошибочным аргумент Лукаса, так как Лукас некорректно применяет понятие «механизма» к вычислительным устройствам и их математическим прототипам. Основной тезис Лукаса: «…теорема Гёделя доказывает, что механицизм ложен, то есть разум нельзя объяснить как машину». В противоположность тому, что утверждает Лукас, старая дискуссия о детерминизме и свободе воли не переводится (по крайней мере тем простым способом, который предлагает Лукас) в контекст формальных систем и вычислений. Лукас также утверждает: теорема Гёделя должна применяться к кибернетическим машинам, она по существу

является машиной, т.е. должна быть конкретным воплощением формальной системы. Это сомнительное утверждение. Аналогия формальной системы, к которой применимы теоремы Гёделя о неполноте, – это язык программирования, а не современный (после Konrad Zuse 1941) перепрограммируемый компьютер (hardware), а также не компьютер, который выполняет некоторую фиксированную программу или набор таких программ. Не ясно, проводит ли Лукас различие между последними двумя понятиями. Иногда возникает впечатление, что под кибернетической машиной Лукас понимает устройство, подобное автоматическим станкам первого поколения, которые выполняют фиксированный набор действий по жестко фиксированной программе, которую невозможно изменить. Шахматная аналогия дедуктивной системы: свод правил игры в шахматы (включающий начальную позицию и правила ходов), а не запись отдельной партии и не программа для игры в шахматы. Весьма нерелевантными выглядят аргументы Лукаса о космических лучах и других случайных внешних факторах, которые могут повлиять на работу вычислительного устройства непредсказуемым образом. Основной «непредсказуемый» (для внешнего наблюдателя) фактор, который ключевым образом влияет на работу компьютера – это программист, который создает и запускает на компьютере ту или иную программу, а также вводит те или иные исходные данные, которые обрабатываются программой.

Тем не менее очень важным представляется ответ Лукаса на статью [Turing, 1950], где утверждалось, что математический аргумент в духе Гёделя (гёделевское противоречивое предложение) можно формализовать и реализовать на компьютере. Лукас отвечает, что даже если мы введем в машину «гёделевский оператор», то это не позволит нам обойти вторую теорему Гёделя. Последнее слово всегда остается за человеческим разумом» постольку, поскольку человек контролирует компьютер, в частности, создавая и запуская различные программы. Такой контроль на практике оказывается проблематичным, однако теоремы Гёделя не имеют прямого отношения к этому вопросу. Другая заметная мысль Тьюринга – «при высоком уровне сложности поведение вычислительного устройства становится непредсказуемым». Ответ Лукаса: в этом случае машины станут умными. Но есть обратные примеры, например, программные генераторы паролей, которые невозможно предсказать. Механические интуиции Лукаса про простые и предсказуемые «мертвые механизмы» и сложные и непредсказуемые «живые умы» являются обманчивыми. Так называемые «сложные умы», в свою очередь, часто ведут себя вполне предсказуемо. Таким образом, во-первых, понятие «кибернетической машины» у Лукаса представляет собой неправомочный перенос понятий популярной классической механики в вычислительный контекст. Это плохо построенное понятие оказывается нечувствительным к различию между полнотой и алгоритмической разрешимостью формальной системы. Во-вторых, попытки строить вычислительные модели человеческого мышления (и мышления животных) неправомерно отбрасывать, используя аргумент Лукаса в качестве основания. В-третьих, статья Лукаса – это пример неправомерного использования математического аргумента как основания для философского аргумента. Теоремы Гёделя о неполноте арифметики не имеют

тех далеко идущих философских следствий, которые им приписывает Лукас и другие философы.

Бессонов А.В. (д.ф.н., проф., НГУ, г. Новосибирск) выступил с докладом «Об обоснованности аргумента Гёделя – Лукаса». Аргумент Гёделя – Лукаса основывается на доказанной Гёделем первой теореме о неполноте. В этой теореме утверждается неразрешимость гёделевой формулы G в формальной арифметике (*PA*), при этом считается, что на неформальном уровне мы можем установить истинность G. А.В. Бессонов попытался опровергнуть последнее утверждение. При истолковании теорем Гёделя происходит подмена понятий: смешиваются понятия истинности и неразрешимости относительно конкретной фиксированной нумерации синтаксиса *PA* и истинности и неразрешимости в арифметике как таковой. В действительности Гёделем доказаны лишь неразрешимость-в-нумерации и истинность-в-нумерации формулы G. Однако определения истинности (и неразрешимости) в арифметике как таковой нумерационно независимы: в этих определениях вообще нет упоминания какой-либо нумерации. То есть при обычной интерпретации теорем Гёделя истинность-в-нумерации смешивается с истинностью-в-арифметике. Но данные понятия не совпадают. Предложенная самим Гёделем нумерация языка *PA* отнюдь не является единственно возможной: выражения *PA* не несут «на своих лбах» свои Гёделевы номера. Истинность-в-нумерации для некоторых формул вообще не определена. При этом они могут быть вполне себе истинными-в-арифметике. Нетрудно привести пример формулы, построенной с использованием предиката «быть Гёделевым номером какого-то выражения арифметики», которая истинна и доказуема в одной нумерации, но ложна и недоказуема в другой. Формула, построенная с использованием предиката «быть Гёделевым номером какого-то выражения арифметики» (а Гёделева формула именно такова), может быть истинной-в-одной-нумерации и ложной-в-другой. Это полностью разрушает фундамент фатального теоретико-познавательного вывода о том, что в «в любой достаточно богатой теории есть истинные, но недоказуемые суждения». Таким образом, основанный на теоремах Гёделя о неполноте аргумент Гёделя – Лукаса некорректен.

А.В. Родин не согласился с докладчиком, поскольку, по его мнению, вопрос гёделевской нумерации не является основой последующих теорий и, следовательно, ему не стоит уделять такое большое внимание.

Винник Д.В. (д.ф.н., проф. Финансового университета при Правительстве РФ, г. Москва) рассмотрел т.н. «эффект последовательности» как аргумент против кибернетической природы ума. Этот термин известен психологам и социологам. Давно замечено, что если перепутать вопросы в анкетах местами, то ответы будут меняться, даже у одних и тех же людей. На это есть многочисленные объяснения. Десять лет назад к вопросу стали подходить математически, как к некоторой части теории рационального выбора.

В 2001 г. была опубликована работа группы американских авторов под руководством Джерома Буземайера «Контекстуальные эффекты, продуцированные последовательностями вопросов, вскрывающие квантовую природу человеческих суждений» [Busemeyer, 2014]. Была сформулирована идея QQQ (Quantum-Question-Quality). Авторы смогли предсказать результаты опросов.

Далее авторы развили свое исследование. Провели анкетирование представителей 70 стран и задали простые вопросы, касающиеся политики. Была обнаружена изменяемость ответов в зависимости от порядка вопросов. То есть природа человеческих суждений является квантовой, не рациональной. Это был удар по сторонникам теории рационального выбора. Люди, например, часто не имеют в принципе никакого мнения по вопросу, пока он не задан, и формируют свое мнение в результате ответа на вопрос.

А.В. Родин обратил внимание на то, что описываемый феномен в самом деле можно назвать «квантовым», ведь слово «квантовый» применимо не только к физике и механике микромира, но и к физике и механике макромира. На это ранее обращал внимание и проф. С.Ф. Сергеев (г. Санкт-Петербург) при анализе социальных массовых явлений, которые организованы по принципу «социального лазера» (термин А.Ю. Хренникова, Швеция).

Зайкова А.С. (к.филос.н, н.с. Института философии и права Сибирского отделения РАН, г. Новосибирск) посвятила доклад книге Стюарта Рассела «Совместимость с человеком» [Russel, 2019] и начала рассуждения с рассмотрения стандартной модели ИИ: 1) Задача ИИ (интеллектуальной системы) – наилучшая реализация цели (целей), поставленной разработчиком; 2) Цели жестко определены; 3) Цели статичны; 4) Цели полностью определяются разработчиками. Однако Стюарт Рассел с этим не согласен. Он считает, что именно в этом главное отличие ИИ от человека: «Когда Вы просите кого-нибудь принести кофе, то это не значит, что вы хотите кофе любой ценой». Проблемы, обнаруженные Стюартом Расселом: 1) Нет возможности скорректировать цели; 2) При задании целей нет возможности учесть всевозможные человеческие ценности; 3) Жесткая зависимость целей от принципов и ценностей разработчика; 4) Нет возможностей предсказать появление новых целей. С. Рассел предлагает ввести ряд принципов, которые нужно ввести в ИИ: 1) Главной целью машины должна быть максимизация реализации человеческих предпочтений; 2) Машина изначально не обладает информацией о человеческих предпочтениях; 3) Единственный источник информации о человеческих предпочтениях – это анализ поведения людей. А.С. Зайкова предложила применить принципы Стюарта Рассела в разрезе аргумента Лукаса: 1) Для каждой машины существует истина, которая не может быть произведена как истина, но которую может распознать как таковую разум; 2) Это касается и целей; 3) Машина не стремится произвести истину, но пытается к ней приблизиться.

Камиль Салямов (бакалавр философского факультета МГУ, г. Москва), заявил тему «Клюджевое программирование и аргумент Лукаса». Дж. Лукас, – считает докладчик, – понимает случайность как контингентность, как аномалию, а аномалии в машине невозможны. Камиль не согласен: 1) Существует такой феномен, как «клюдж». Это ситуация, когда программа выполняет больше, чем от нее ожидается; 2) Если «клюдж» имеется в программе, то программа может быть полной и непротиворечивой в контингентном смысле, т.е. пока наблюдатель не может определить, какой член дизъюнкции ложен; 3) Следовательно, возможна такая программа, которая будет полна и непротиворечива в контингентном смысле. Возможно ли клюджевое

программирование? 1) Программа не является чисто формальной системой и может быть представлена графом, в который эта система имплементирована; 2) Граф имеет морфизмы в топологическом смысле; 3) Мы можем представить графы как миры, а достижимости между ними – как морфизмы; 4) Противоречие в одном мире не гарантирует противоречие в другом мире, достижимом из данного; 5) Следовательно, программа не выйдет из строя до тех пор, пока противоречие не станет необходимым. Проблема тезиса Лукаса о том, что существует определенное количество различных видов операций, которые машина может выполнять, но которые непредставимы в теоретико-алгоритмическом формате.

А.В. Родин и В.В. Целищев раскритиковали термин «клюдж»: в изучении конкретных способов практической реализации логического вывода всегда требуются оговорки и интерпретации, зачем возводить некорректные и незапланированные способы функционирования в самостоятельный предмет исследования. А.Ю. Алексеев, наоборот, отметил валидность понятия «клюдж»: в практике программирования часто встречаются ситуации, когда программа работает, хотя программист не может понять, почему она работает. «Клюджем» и принято называть программу, которая бессистемно создается вне скоординированной дисциплины программирования, в противоречиях с другими разработчиками, а чаще всего – одиночкой-программистом. И программа, как ни странно, работает. Хотя, конечно, чаще бывает наоборот: непонятно то, почему программа не работает.

Антипова А.В. (магистрантка философского факультета МГУ, г. Москва) обозначила тему «Лукас и суперкомпьютер». Тезис Гёделя – Лукаса – Пенроуза говорит о ложности механистического подхода в проблематике ИИ, позволяет сделать заключение о том, что машина никогда не сможет быть моделью человеческого разума, и вывод о слабости ИИ по сравнению с ЕИ. Дж. Лукас считает, что для того, чтобы мы могли сравнить способности человека и компьютера, компьютер должен суметь выполнить любое задание, на которое способен человек. Если мы найдем такую операцию, которая на логическом уровне будет невозможна для не-человека, то тезис Лукаса – Пенроуза будет доказан. Нашему сознательному уму присуще то, что он может размышлять над собой и над своими действиями и для этого не требуется ничего сверх рефлексии подобного рода. Аргумент Лукаса – Пенроуза постоянно критиковался. Например, Дугласом Хофштадтером, который показал, что с увеличением сложности систем наша человеческая способность к «гёделизации» начинает ослабевать. Человеческий мозг, столкнувшись со слишком сложной системой, также не может сообразить, что делать дальше. В таком случае формальные системы, пусть даже и не полные, могут быть приравнены к человеческому разуму. Другой критический контраргумент против Лукаса: согласно Лукасу, гёделевский аргумент должен применяться к кибернетическим системам, поскольку сущностью машин является быть конкретным воплощением формальной системы. То есть если какой-то системе удается обойти проблему Гёделя, то такая система более не является кибернетической машиной. Таким образом мы попадаем в логический круг определений «машинности». Посылки, которые лежат в основе аргумента Лукаса: 1) Не существует способа достичь

человеческого уровня интеллекта; 2) ЕИ метафизически лучше ИИ. Однако первый аргумент не валиден: существуют способы достичь человеческого уровня интеллекта, и Чат GPT это воочию демонстрирует. Второй аргумент снимается понятием «суперкомпьютер», которое как большое число параллельно работающих компьютеров. Они решают сложные задачи, хотя для этого не привлекаются ни сознание, ни эмоции. Необходима такая теория интеллекта, которая бы и учитывала градацию интеллектуальных способностей, и характеризовала специализацию этих градаций.

Толоконников Г.К. (к.ф.-м.н., с.н.с. ВИМ, г. Москва), по сути, уточнил возможность представления предложенных А. Антиповой «спецификаций градаций». В этой связи показательной ему показалась идея представимости оснований арифметики, изложенная в статье Лори Кирби и Джефа Париса «Доступные независимые результаты арифметики Пеано» [Kirby, Paris, 1982]. В ней рассматривались числовые последовательности и формула Гудстейна о прекращении этих последовательностей. Это утверждение и есть еще одна интерпретация аргумента Гёделя.

Алексеев А.Ю. не согласился с выводом докладчика. Этот математический парадокс для интерпретации теоремы Гёделя предлагает Р. Пенроуз в начале книги [Пенроуз, 2003]. Исследование парадокса важно, так как проясняет спор между механицистами и менталистами (В.В. Целищев). Парадокс математически воспроизводит миф о победе Геракла над гидрой. Надо запустить такую числовую последовательность, которая бы запрещала эту последовательность. То есть надо сформулировать такую стратегию отрубания голов гидры (удаления чисел), чтобы обрубить бессмертную голову: вычислить и удалить некоторое число как корень постоянно растущего графа числовой последовательности. Однако теорема Гёделя утверждает более фундаментальное противоречие. Геракл обладает априорными знаниями о стратегии порождения последовательности, запрещающей порождение последовательности. Для гёделевского противоречия таких априорных знаний нет. Эти знания надо создать изнутри этих самых знаний. Последовательность Гёделя имеет семантический характер: лжец, лгущий о своей лжи, существенно отличается от брадобрея, не бреющего свою бороду. Парадокс лжеца имеет семантические и, следовательно, теоретические основания, а парадокс брадобрея (и Геракла) таких оснований не имеет, так как лежит в плоскости синтаксических комбинаций. Парадокс Кирби – Париса (парадокс Геракла, парадокс гидры) предназначен для алгоритмической обработки. Возможно, потребуется суперкомпьютерная обработка, как в современной криптографии, но рано или поздно этот парадокс разрешится. Парадокс же Гёделя неразрешим на синтаксическом уровне. Он предназначен для квазиалгоритмической обработки, т.е. для систем ИИ. Формализовать неформализуемую систему – а именно таковой является система ИИ – в этом состоит величие проблемы Гёделя, подмеченное Д. Лукасом.

Сергеев С.Ф. (д.псих.н., проф. СПбГУ, г. Санкт-Петербург) исследовал мировоззренческие основания проблемы Гёделя – Лукаса. Что такое исследования проблем ИИ? Это попытка человека дойти до истины, т.е. попытка системы создать такую модель, которая превзойдет свой оригинал, т.е. саму эту систему.

Ващенко А.А. (аспирант Физтех-школы МФТИ, г. Москва) обратил внимание на искусственность строгой дизъюнктивной оппозиции естественного и искусственного интеллектов. С усложнением алгоритмов и языков программирования и увеличением вычислительной мощности машин споры вокруг аргумента Гёделя – Лукаса – Пенроуза не окончились, а только вышли на другой уровень, определив, по сути, новое качество дискуссий. Современные компьютерные системы больше не являются прямыми родственниками «кибернетической машины» в понимании Лукаса. Программирование все чаще и чаще называется искусством. Начиная с первых разработок в области объектно-ориентированного программирования прошлого века до современных алгоритмов проектирования нейросетей человек получает от машины все менее и менее предсказуемый результат. Программа может пойти по пути, который человек неспособен предусмотреть. И машина в споре с человеком выступает не в одиночку. За ней стоят коллективы программистов. Даже простые задачи «обучения с подкреплением» алгоритмов машинного обучения знаменуют новый этап в разработке программ: это *интеракционное кообучение*, когда машина работает в связке с людьми и каждый член оппозиции непредставим вне другого члена. Без синтеза ЕИ и ИИ невозможен чат-GPT.

Пожарев Тодор (электронный художник, аспирант философского факультета МГУ, г. Нови Сад, Сербия) отметил, что в современных исследованиях проблемы творчества в ИИ, точнее, в последних исследованиях аргумента (теста) Лавлейс [Алексеев, Пожарев, 2020] актуальной является проблема идентификации «смысла» в артефактах. Эта задача явно прописана в тесте Лавлейс 3.0 с различными вариациями. Тест Ватта позволяет вместе с инвертированием теста Тьюринга обратить и тематику теста Лавлейс, т.е. создать новое семейство тестов Лавлейс. Мы «поворачиваем» тьюринговое тестирование с анализа «смысла» в артефактах к процессу приписывания артефакту «смысла». То есть важной герменевтической задачей ИИ является не только интерпретация смысла, но и его презентация.

Алексеев А.Ю. (д.филос.н., в.н.с. философского факультета МГУ, г. Москва) в докладе «Противоречия киборга: аргумент Гёделя – Лукаса – Ватта» отметил то, что в современных философских и научных дискуссиях упорно просматривается стратегия, обозначаемая как «Гёдель – Лукас – Пенроуз – Хамерофф». Это метафизическая физикалистская концепция сознания, конечным продолжателем которой является анестезиолог Стюарт Хамерофф. 10 октября 2016 г. он выступал с докладом на семинаре «Нейрофилософия» НСМИИ РАН в МГУ (рук. А.Ю. Алексеев, модератор проф. В.Г. Кузнецов, оппонент – Д.И. Дубровский) [Хамерофф, 2016]. Выступление подробно и разгромно критически проанализировано Д.И. Дубровским на страницах настоящего журнала [Дубровский, 2017]. Алексеев А.Ю. обратил внимание на принципиальную невозможность формулирования теории сознания, когерентно существующей в космических просторах Вселенной. Любая теория сознания не объективируема, она зависит от того, как *Я* понимаю сознание и как приписываю сознание и самосознание другим, в том числе компьютерным системам. Формальное выражение дистинкции *сознание/самосознание* четко представлено в аргументе Гёделя. Однако зачем в философии

и методологии ИИ воспроизводить паранаучные идеи о сознании Р. Пенроуза? Для систем ИИ интересной представляется ветвь «Гёдель – Лукас – Ватт», которая использует идеи Стюарта Ватта о тьюринговом приписывании сознания и самосознания объектам природы и артефактам техники [Watt, 1996]. Решение гёделевских проблем приписывания машиной машине сознания и интеллекта требует специальных исследований и будет исследовано далее в совместной работе А.Ю. Алексеева и Т. Пожарева.

Заключение

С каждым годом увеличивается степень участия интеллектуальных систем в человеческой деятельности. Ставшие популярными не так давно, не более года назад, генеративные модели способны обеспечить осмысленный диалог, составление рефератов, дипломных работ и диссертаций; сдачу экзаменов, сочинение музыки и театральных сценариев с визуальным и звуковым сопровождением; имитацию художественного полотна; и делать многие другие вещи, ранее выполняемые творческими людьми. Однако можно ли такие программы отнести к системам искусственного интеллекта? Эти интеллектуальные системы используют лишь то, что сформулировал и представил человек. Ничего нового от себя они не продуцируют. Никакой новой теоремы. Никакой новой классификации. Никакого нового понятия, тем более категории. Почему люди подвержены обману интеллектуальности таких неинтеллектуальных систем, каковым являются чат-боты GPT? Это культуропсихосоциополитологический парадокс. Доклады, прозвучавшие на симпозиуме, изучали основания подобных парадоксов и противоречий ИИ.

Несмотря на то, что аргумент Лукаса – Гёделя – Пенроуза существует давно и не раз обсуждался, в ходе симпозиума были отмечены и проанализированы новые модификации, новые аргументы, проработаны логико-философские, формально-вычислительные и культурно-антропологические аспекты. Было показано, что из-за бурного развития систем искусственного интеллекта используемые понятия и концепции претерпевают значительные трансформации, однако тест Гёделя – Лукаса и другие подобные тесты теоретически важны для осмысления тех систем имитации человеческой деятельности, которые уже сегодня функционируют и соперничают с людьми во всех сферах общественной жизни.

Список литературы

Алексеев, 2013 – *Алексеев А.Ю.* Комплексный тест Тьюринга: философско-методологические и социокультурные аспекты. М.: ИИнтеЛЛ, 2013. 304 с.

Алексеев, Пожарев, 2015 – *Алексеев А.Ю., Пожарев Т.* Креативные мультимедиа: физикалистский и менталистский подходы // Философия творчества: материалы Всероссийской научной конференции, 8–9 апреля 2015 г., Институт философии РАН, г. Москва. М.: ИИнтелл, 2015. С. 418–424.

Дубровский, 2017 – *Дубровский Д.И.* Критический анализ теории сознания Пенроуза – Хамероффа. Часть 1 // Философия науки и техники. 2017. Т. 22. № 1. С. 125–136.

Пенроуз, 2003 – *Пенроуз Р.* Новый ум короля. О компьютерах, мышлении и законах физики. М.: УРСС, 2003. 416 с.

Пенроуз, 2005 – *Пенроуз Р.* Тени разума. В поисках науки о сознании. М.; Ижевск: Институт компьютерных исследований, 2005. 688 с.

Хамерофф, 2016 – *Хамерофф С.* Оркестрируемая объектно-редуцируемая (ORCH OR) теория сознания как квантового вычисления в микротрубочках мозга: 20 лет спустя. Тезисы доклада на семинаре НСМИИ РАН «Нейрофилософия», 10 октября 2016 г., МГУ; URL: https://scmai.ru/2016/10/10/

Целищев, 2021 – *Целищев В.В.* Алгоритмизация мышления: Гёделевский аргумент. 2-е изд., испр. М.: ЛЕНАНД, 2021. 304 с.

Busemeyer, 2014 – *Busemeyer J., Busemeyer J., Solloway T., Shiffrin R., Wang Z.* Context effects produced by question orders reveal quantum nature of human judgments // PNAS. 2014. Vol. 111. No. 26. P. 9431–9436.

Chalmers, 1995 – *Chalmers D.J.* Mind, Machines, and Mathematics // PSYCHE. 1995. Vol. 2. No. 9. P. 11–20.

Feferman, 1995 – *Feferman S.* Penrose's Godelian Argument // PSYCHE. 1995. Vol. 2. No. 7. P. 21–32.

Goodfellow, 2020 – *Goodfellow I. et al.* Generative adversarial networks // Communications of the ACM. 2020. Vol. 63. No. 11. P. 139–144.

Kirby, Paris, 1982 – *Kirby L., Paris L.* Accessible independence results for Peano arithmetic // Bulletin of the London Mathematical Society. 1982. Vol. 14. Iss. 4. P. 285–293.

Russel, 2019 – *Russell S.* Human Compatible: AI and the Problem of Control. London: Penguin, 2019. 349 p.

Turing, 1950 – *Turing A.M.* Computing machinery and intelligence // Mind. 1950. No. 59. P. 433–460.

Watt, 1996 – *Watt S.* Naive Psychology and the Inverted Turing Test // PSYCOLOQUY. 1996. Vol. 7. No. 14.

Paradoxes and contradictions of artificial intelligence: 90 years of K. Gödel's first incompleteness theorem and 60 years of J. Lucas's argument

Andrey Yu. Alekseev

State Academic University for the Humanities. Russian Federation, 119049, Moscow, 26 Maronovsky lane; RUDN; e-mail: aa65@list.ru

Alexander A. Vashchenko

MIPT. Russian Federation, 141701, Moscow region, Dolgoprudny, 9 Institutsky lane; e-mail: alexanderva@mail.ru

Alina S. Zaykova

IPL SB RAS. Russian Federation, 630090, Novosibirsk, 8 Nikolaeva str.; e-mail: zaykova.a.s@gmail.com

Today, the study of the issue of distinction has turned out to be extremely relevant – what proportion of the work in the produced artifact belongs to a person, and what proportion belongs to a computer. In this direction, it is important to develop computer analyzers of computer generators of texts, pictures, music, animations, and videos. The fundamental theoretical-algorithmic status of such generators is given by the Gödel–Lucas–Penrose test. In 1961, Oxford professor J. Lucas, on the basis of Gödel's second theorem, concluded that the human mind is fundamentally superior to all kinds of computing systems.

R. Penrose, supporting this position, stated that human consciousness is not algorithmic and goes beyond computability. This argument has generated extensive discussion regarding its philosophical underpinnings, mathematical validity, physical and neurophysiological explanations, psychological credibility, and technological implementation. The dispute about the importance and alternative ways of interpreting the argument does not stop, constituting one of the basic provisions of the philosophy and methodology of artificial intelligence and cognitive science. In October 2021, the SCMAI RAS held the All-Russian Symposium "Artificial Intelligence: Paradoxes and Contradictions", during which the logical-philosophical, formal-computational and cultural-anthropological aspects of the Gödel-Lucas argument were studied, and possible ways of its transformation and development were shown, such as the Gödel-Lucas-Watt test.

Keywords: philosophy and methodology of artificial intelligence, Gödel's theorem, Lucas argument, Gödel-Lucas-Penrose argument, comprehensive Turing test, inverted Turing test, Watt test, GPT, Gödel-Lucas-Watt argument

Acknowledgments: The article was prepared within the framework of the state task of the GAUGN "Digitalization and the formation of a modern information society: cognitive, economic, political and legal aspects". Registration number NIOKTR 123022000042-0. Subject code FZNF-2023-0004. Subject registration number 1022040800826-5-5.2.1;6.3.1;5.9.1.

References

Alekseev, A.Y. *Kompleksnyi test T'yuringa: filosofsko-metodologicheskie i sotsiokul'turnye aspekty* [Complex Turing Test: philosophical, methodogical and sociocultural aspects]. Moscow: IInteLL Publ., 2013, 304 pp. (In Russian)

Alekseev, A.Yu., Pozharev, T. 'Kreativnye mul'timedia: fizikalistskii i mentalistskii podkhody' [Creative multimedia: physicalist and mentalist approaches], in: *Filosofiya tvorchestva: materialy Vserossiiskoi nauchnoi konferentsii, 8–9 aprelya 2015 g., Institut filosofii RAN, g. Moskva* [Philosophy of creativity: Papers of the All-Russian scientific conference, April 8-9 2015, RAS Institute of Philosophy, Moscow]. Moscow: IInteLL Publ., 2015, pp. 418–424. (In Russian)

Busemeyer, J., Busemeyer, J., Solloway, T., Shiffrin, R., Wang, Z. 'Context effects produced by question orders reveal quantum nature of human judgments', *PNAS*, 2014, vol. 111, no. 26, pp. 9431–9436.

Chalmers, D.J. 'Mind, Machines, and Mathematics', *PSYCHE*, 1995, vol. 2, no. 9, pp. 11–20.

Dubrovskii, D.I. 'Kriticheskii analiz teorii soznaniya Penrouza – Khameroffa' [The critical analysis the Penrose–Hameroff theory of consciousness. Part 1], *Filosofiya Nauki i Tekhniki/Philosophy of Science and Technology*, 2017, vol. 22, no. 1, pp. 125–136. (In Russian)

Feferman, S. 'Penrose's Godelian Argument', *PSYCHE*, 1995, vol. 2, no. 7, pp. 21–32.

Goodfellow, I. et al. 'Generative adversarial networks', *Communications of the ACM*, 2020, vol. 63, no. 11, pp. 139–144.

Hameroff, S. Orkestriruemaya ob"ektno-redutsiruemaya (ORCH OR) teoriya soznaniya kak kvantovogo vychisleniya v mikrotrubochkakh mozga: 20 let spustya. Tezisy doklada na seminare NSMII RAN "Neirofilosofiya", 10 oktyabrya 2016 g., MGU [A Review of the 'Orch OR' Theory of Consciousness. Paper abstract at the NSMII RAS seminar "Neurophilosophy", 10 October 2016, MSU]; URL: https://scmai.ru/2016/10/10 (In Russian)

Kirby, L., Paris, L. 'Accessible independence results for Peano arithmetic', *Bulletin of the London Mathematical Society*, 1982, vol. 14, iss. 4, pp. 285–293.

Penrose, R. Novyi um korolya. O komp'yuterakh, myshlenii i zakonakh fiziki [The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics], Moscow: URSS Publ., 2003, 416 pp. (In Russian)

Penrose, R. Teni razuma. V poiskakh nauki o soznanii [Shadows of the Mind: A Search for the Missing Science of Consciousness], Moscow-Izhevsk: Institut komp'yuternykh issledovanii Publ., 2005, 688 pp. (In Russian)

Russell, S. Human Compatible: AI and the Problem of Control. London: Penguin, 2019. 349 p.

Tselishchev, V.V. *Algoritmizatsiya myshleniya: Gedelevskii argument* [Algorythmization of thought: Godelian argument]. 2nd ed. Moscow: LENAND Publ., 2021, 304 pp. (In Russian)

Turing, A.M. 'Computing machinery and intelligence', *Mind*, 1950, no. 59, pp. 433–460.

Watt, S. 'Naive Psychology and the Inverted Turing Test', *PSYCOLOQUY*, 1996, vol. 7, no. 14.