DOI: 10.30727/0235-1188-2021-64-7-26-45
Оригинальная исследовательская статья
Original research paper

Противоречивость как положительное свойство разума: 90 лет геделевскому аргументу

Д.В. Винник
Финансовый университет при Правительстве Российской Федерации, Москва, Россия

Аннотация

В статье обращается внимание на самобытное и единственное в своем роде систематическое исследование В.В. Целищевым специфической и крайне запутанной проблематики интерпретации результатов геделевских теорем относительно вопроса о природе искусственного интеллекта. Целищев утверждает, что свойство рефлексивности следует рассматривать не только как преимущество человеческого мышления, но и как объективное внутреннее ограничение, проявляющееся при использовании операции добавления геделева предложения к теории для построения новой теории. Анализируется т.н. геделевский аргумент менталистов в пользу принципиального превосходства человеческого интеллекта над машинным и неалгоритмической природы естественного мышления. Утверждается, что полемика относительно геделевского аргумента не является целиком спекулятивной, но содержит новое знание. Примером такого знания являются результаты Р. Смаллиана об уровнях «сознания» вычислительных машин, которые можно интерпретировать в психофизическом смысле. Предлагается понятие «нулевого уровня разумности» для такого рефлексивного свойства, как «осознание самосознания». Рефлексивные ранги ниже осознания самосознания можно считать отрицательными уровнями мышления в том смысле, что, редуцируясь к ним, интеллект существенно теряет свою полноту. Даже самосознание оказывается отрицательным уровнем мышления, т.к. субъект самосознания не ведает о типе мышления, к которому он принадлежит, согласно Смаллиану. Предлагается мысленный эксперимент, позволяющий установить распределение свойств смаллиановской стабильности и нормальности и ответить на вопрос «Влияет ли интуитивная вера в истинность формального доказательства на истинность доказываемого утверждения?». Согласно интуитивизму наиболее неприятным эпистемическим свойством следует считать нестабильность. Убеждения, не основанные на глубоких интуициях, не имеют цены. Согласно конструктивистской

философии математики нестабильность есть менее негативное свойство, чем ненормальность. Тот факт, что высокоранговые убеждения не могут быть погружены до самых оснований, вряд ли имеет большое значение, т.е. нарушение сохранения истинности при понижении ранга рефлексии не является критическим.

Ключевые слова: искусственный интеллект, ментализм, машинный функционализм, алгоритмизация мышления, самосознание, когнитивные функции, нейрофизиология, детекция лжи, полиграф.

Винник Дмитрий Владимирович – доктор философских наук, профессор департамента гуманитарных наук Финансового университета при Правительстве Российской Федерации.

dvinstor@gmail.com

https://orcid.org/0000-0002-3410-8023

Для цитирования: *Винник Д.В.* Противоречивость как положительное свойство разума: 90 лет геделевскому аргументу // Философские науки. 2021. Т. 64. № 7. С. 26–45.
DOI: 10.30727/0235-1188-2021-64-7-26-45

Contradiction as a Positive Property of the Mind: 90 Years of Gödel's Argument

D.V. Vinnik
Financial University under the Government of Russian Federation, Moscow, Russia

Abstract

The article discusses the V.V. Tselishchev's original and unique systematic study of the specific and extremely complicated problems of Gödel results regarding the question of artificial intelligence essence. Tselishchev argues that the reflexive property should be considered not only as an advantage of human reasoning, but also as an objective internal limitation that appears in case of adding Gödel sentence to a theory to build a new theory. The article analyzes so-called mentalistic Gödel's argument for fundamental superiority of human intelligence over machine one and the non-algorithmic nature of natural thinking. The discussion about the Gödel argument is not entirely speculative, but contains new knowledge. An example of such knowledge are the results of R. Smullyan levels of computers "awareness," which are may be interpreted in a psychophysical sense. The concept of "zero level of intelligence" is proposed for such a reflexive property as "awareness of self-consciousness." Reflexive ranks below the awareness of self-consciousness can be considered negative levels of thinking in the sense that the intel-

ligence, being reduced to them, significantly loses its completeness. Even self-consciousness turns out to be a negative level of thinking, since, according to Smullyan, the subject of self-consciousness is unaware of the type of thought to which he belongs. A thought experiment is proposed that allows us to establish the distribution of the properties of Smullyan stability and normality and to answer the question "Does an intuitive belief in the truth of a formal proof affect the truth of a proposition being proved?" According to intuitionism, the most unpleasant epistemic property is instability: beliefs that are not based on deep intuitions have no value. According to the constructivist philosophy of mathematics, instability is a less negative property than abnormality: the fact that high-ranking beliefs cannot be immersed to the very foundations is not significant because violation of truth due to lowering the rank of reflection is not critical.

Keywords: artificial intelligence, Gödel theorem, mentalism, machine functionalism, mind, algoritmisation, self-awareness, cognitive functions, neurophysiology, polygraph, deception detection.

Dmitriy V. Vinnik – D.Sc. in Philosophy, Professor, Financial University under the Government of Russian Federation.
dvinstor@gmail.com
https://orcid.org/0000-0002-3410-8023

For citation: Vinnik D.V. (2021) Contradiction as a Positive Property of the Mind: 90 Years of Gödel's Argument. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 64, no. 7, pp. 26–45.
DOI: 10.30727/0235-1188-2021-64-7-26-45

Введение

В 2021 году исполняется 90 лет с того момента, как Курт Гедель опубликовал свои знаменитые теоремы. Без преувеличения можно утверждать, что немногие научные достижения XX века имеют настолько полемичную и скандальную историю. Эти результаты давно стали частью околонаучной культуры, и ссылки на них наравне с апелляциями к квантовой механике вошли в типичные способы имитации наукообразия и оправдания методологического анархизма [Винник 2016]. В философии математики не утихает полемика как о корректности доказательства теорем, так и о степени спекулятивности различных ее интерпретаций. Однако существуют все основания считать, что геделевские результаты затронули суть проблематики искусственного интеллекта.

Понятие искусственного интеллекта (ИИ или «искин») не является строгим, начиная с его появления в 1956 году на семинаре в Дартмутском университете [Moor 2006]. В предметных областях и даже

научных школах информатики и кибернетики содержание понятия ИИ может значительно различаться. Это справедливо и для философов науки, предпринимающих попытки обобщения знаний, так или иначе представляемых с помощью данного понятия. Отсутствие согласия следует признать закономерным и в известной степени нормальным положением дел. Оно обусловлено тем, что понятие интеллекта само по себе является глубоко полемичным. Его объем и содержание зависит от множества онтологических допущений и концептуальных каркасов, в которые это понятие встроено.

Как известно, философы различают понятия сильного и слабого искусственного интеллект, а также спорят о принципиальной возможности существования сильного ИИ [Wah, Chi 2020]. Принятие понятия сильного иискусственного интеллекта влечет возможность создания искусственной личности, обладающей всеми полноценными атрибутами: самосознанием, творческим мышлением, и, если возвыситься до самых «сокровенных» философских категорий, – разумом. Тех, кто отрицает возможность создания или даже спонтанного зарождения разума в сложных кибернетических системах в результате скачка сложности в процессе технической эволюции последних, обычно относят к сторонникам слабого ИИ. Приверженцы подобной трактовки природы ИИ допускают лишь воспроизводство техническими средствами определенных когнитивных функций без проникновения в сущность человечности и разумности.

Споры о возможности создания полноценной искусственной личности с присущими ей творческими мыслительными способностями ведут с привлечением аргументов из разнообразных дисциплин, от респектабельной эволюционной биологии до вульгарной метафизики. К сожалению, подобные дискуссии редко выходят за пределы спекулятивной аргументации, прямые аналоги которой без труда можно обнаружить в прошлых веках. Действительно, ответ на вопрос «Может ли машина мыслить?» очевиден как для убежденного механициста эпохи Нового времени, так и для любого его противника. Между тем утверждать, что философское осмысление этого фундаментального вопроса не продвинулось вперед в свете знаний современной науки, будет неверным.

Ренессанс перцептроники и ее теоретическая нищета

С одной стороны, в последнее десятилетие мы наблюдаем значительные успехи в моделировании когнитивных функций, относительно которых оптимизм ранних кибернетиков к концу XX века, казалось, был окончательно утерян. В качестве примера приведем такую способность, как перевод с одного естественного

языка на другой. Созданные интенсиональные логики не оправдали возлагаемых на них надежд, т.е. они оказались не лишены фундаментального изъяна, который У.В.О. Куайн именовал онтологической относительностью [Куайн 1996] и ради избавления от которого они изначально задумывались. Более того, за короткий период на существующей вычислительной базе было создано программное обеспечение, позволившее превзойти человека в такой когнитивной функции, которая традиционно считается в психологии одной из самых фундаментальных для любого типа психики, как животной, так и человеческой. Речь идет о распознавании чувственных образов. Сегодня системы полицейского надзора способны успешно распознавать лица, закрытые гигиеническими масками, по верхней части черепа. Очевидным становится, что подавляющее большинство людей решить такую задачу не могут. Сопоставление же человека в толпе с базой данных, содержащей миллионы и даже десятки миллионов изображений, не по силам ни одному из самых тренированных полицейских.

Важно иметь в виду, что эти прорывы в основном обусловлены массовым внедрением машинного обучения на основе симуляций перцептронов, что в некоторой степени умаляет их значение с точки зрения прогресса теоретического знания. Без преувеличения отметим, что этот прогресс почти равен нулю. Идея нейрокомпьютера выдвинута еще в 1943 году [Mcculloch, Pitts 1943]. Сейчас мы наблюдаем ее триумф в прикладной области. Это стало возможным благодаря росту доступных вычислительных мощностей, необходимых для эмуляции нейронных сетей на машинах фон-неймановской архитектуры, доминирующих в техносфере. Удовлетворительная теория нейронных сетей еще не создана, и машинное обучение на их основе, скорее, является своеобразным техническим искусством.

Стоит обратить внимание на то, что машинное обучение на основе нейросетей имитирует живые системы и эволюцию. Замечательно, что человеческий разум оказался способен выполнить нетривиальную задачу, поняв основные принципы работы нейронов и нервной ткани. Однако необходимо помнить о том, что, в отличие от работы классических алгоритмов, результаты работы нейросетевого машинного обучения интеллектуально непрозрачны, т.е. на выходе образуется не более чем некий результат работы и матрица весов логических вентилей. Сама по себе такая матрица не слишком отличается от белого шума и с учетом данной точки зрения является бессмысленной. Это не означает, что не существует способов интерпретации результатов работы

нейрокомпьютеров, исследования в этом направлении ведут давно [Миркес 1998]. Но методы «вербализации» и «прореживания» нейронных сетей для извлечения явного знания носят, скорее, эмпирический характер. В сложившейся ситуации делать весомые онтологические умозаключения о приросте рационального, теоретического знания в области понимания того, что такое искин и каковы пределы его возможностей, не приходится.

С другой стороны, существует область фундаментального знания, в которой вопрос о соотношении понятий естественного интеллекта и искусственного интеллекта поставлен на принципиально новом теоретическом уровне, немыслимом для прошлых эпох. В данном случае речь идет о математической логике и теории вычислений. Скептики могут указать на то, что постановка этого вопроса восходит к парадоксу лжеца и различным версиям теоретико-множественных парадоксов, проистекающих из использования самореферентных высказываний. В известной мере это так, но между пониманием сути парадокса лжеца и сути теорем Геделя о неполноте – огромная дистанция. К сожалению, знаменитые математические результаты Геделя стали предметом множества спекуляций, выходящих далеко за пределы математической логики и математики в целом. Как справедливо указывает Р. Докинз, некорректные ссылки на геделевские результаты послужили одним из четырех самых популярных способов создания наукообразия: она вошла в джентельменский набор постмодернистов наравне со ссылками на квантовую механику и теорию хаоса, известную на постсоветском пространстве как синергетика [Докинз 2013].

Метафизика от математики на страже церебральной разумности

Указанное выше явление негативно сказалось на восприятии тех работ, которые действительно демонстрировали нетривиальные философские следствия знаменитых геделевских результатов. Речь идет об изданной в 1961 году статье «Умы, машины и Гедель» оксфордского профессора Дж. Лукаса [Lukas 1961]. Согласно утверждению Лукаса, приведенному в этой работе, из второй теоремы Геделя следует, что человеческий разум принципиально превосходит возможности разума машинного. Длительное время профессор Лукас подвергался мощнейшей критике с разных сторон, фактически оставаясь в одиночестве. Впоследствии его поддержал лауреат Нобелевской премии Роджер Пенроуз [Пенроуз и др. 2004]. Однако это не охладило пыл многочислен-

ных критиков, и даже авторитетного Пенроуза упрекали в том, что он вторгся в несвойственную для себя предметную область и что его выводы носят поспешный характер. У этой драматической полемики был очень неожиданный результат.

Развитие полемики, изощренные аргументы, которые пущены в ход, включая, например, теорему Гудстейна, использующую для доказательства тривиально выразимых арифметичеких истин «нечеловеческую математику», – всё это изложено в книге профессора В.В. Целищева из Новосибирского научного центра «Алгоритмизация мышления. Геделевский аргумент» [Целищев 2021, 260]. Без сомнения, это – единственная работа, которая во всей своей полноте проливает свет на вопрос о соотношении понятий человеческого интеллекта и машинного интеллекта в контексте проблем, проистекающих из результатов осмысления наследия Геделя, а также всех теоретических проблем, следующих из использования самореферентных высказываний и различных форм того, что философами обычно именуется рефлексивностью. Впервые книга издана в 2005 году. Однако она не была замечена должным образом широкой философской общественностью, что во многом обусловлено как сложностью проблематики, так и общей тенденцией к спаду интереса в отношении точных наук в России. С тех пор ситуация изменилась. В частности, в 2021 году исправленная, дополненная работа пережила второе издание и уже привлекла внимание специалистов в сфере ИИ.

В предисловии к своей книге Целищев обращает внимание на то, что массовая, порой уничижительная критика в отношении Лукаса и Пенроуза внезапно прекратилась, как только из расшифрованных записных книжек К. Геделя стало очевидным, что он осознавал философские следствия своей теоремы и их понимание, что находится, видимо, на их стороне. Сама по себе эта ситуация может рассматриваться как факт психологического или социологического порядка, говорящий скорее о нравах научного сообщества и человеческой природе, чем о содержании геделевских результатов самих по себе. Тот факт, что Гедель трактовал свои результаты подобно его некоторым последователям, причем, в отличие от них, делал это скупо и обобщенно, не является решающим аргументом в пользу истинности этой интерпретации. Тем не менее авторитета отца проблематики полноты арифметики и «пророка» существования недоказуемых истин оказалось достаточно, чтобы большинство критиков отступили.

Для создания более ясного представления о том, какая проблема находится на острие вопроса о соотношении естественного и

искусственного в области природы разума, Целищев прибегает к понятию «алгоритмизация мышления»: «Во-первых, алгоритмизация мышления часто увязывается с очень широким вопросом, который романтически формулируется как вопрос о том, “может ли машина мыслить”. В такой формулировке проблема приобретает столь большую неопределенность и метафоричность, что практически уходит из сферы рационального. Правда, известный тест Тьюринга ставит проблему в определенные рамки, но и в этом случае она приобретает в значительной степени скорее психологический характер» [Целищев 2021, 8]. Автор обращает внимание на то, что в т.н. когнитивных исследованиях, посвященных созданию компьютерных моделей мозга, обычно прибегают к использованию теорий и концепций из математики, физиологии, теории сложности, компьютерных исследований и др. Все эти и смежные дисциплины еще недавно пытались объединить под названием «кибернетика», которое следует понимать как «практически необъятное поле действия, в значительной мере связанное с программами построения систем искусственного интеллекта» [Целищев 2021, 8]. Намеренно отстраняясь от подобного крайне синкретичного и эмпирически контаминированного дискурса, Целищев поднимает вопрос о необходимости исследования принципиальной возможности алгоритмизации мышления, уходя от изучения практических попыток такой алгоритмизации. Действительно, фундаментальный вопрос о возможности создания сильного ИИ в наиболее ясной и острой форме ставят в философии логики, и он имеет непосредственное отношение к основаниям математики.

Противоположных ответов на этот вопрос придерживаются сторонники механицизма и ментализма, к которым и следует отнести Лукаса, Пенроуза и Геделя, инспирировавшего проблематику в целом. Механицисты следуют редукционистской точке зрения о том, что природный ум эквивалентен вычислительной машине, что он может быть формализован. Из этого следует, что разнообразные усилия по моделированию ума осмысленны и создание искусственного разума теоретически возможно. И среди тех, кто так или иначе вовлечен в указанную проблематику, механицистов – большинство.

Одна из наиболее распространенных философских форм механицизма – функционализм, к сторонникам которого причисляют П. Черчленд, Дж. Фодора, Н. Блока. К самым известным механицистам, очевидно, следует отнести раннего Х. Патнэма, выдвинувшего редукционистскую концепцию машинного функциона-

лизма, согласно которой ментальные состояния тождественны состояниям машины Тьюринга [Putnam 1960].

Функционализм имеет любопытные следствия. Если принять эту концепцию, то необходимо признать, что вычислительные устройства могут быть как искусственными, так и естественными по своему происхождению. Для универсального алгоритма это не имеет значения: он способен успешно функционировать на субстратах разного типа. В пределе в роли вычислительного устройства может работать любой природный объект, обладающий достаточным количеством логических элементов и связей между ними для формирования полноценной элементной базы. Влиятельность функционализма объясняется именно тем, что он предстал в виде онтологии принципиально нового типа, а именно – онтологией отношений в противовес субстанциональной онтологии вещей или феноменалистской онтологии свойств [Уёмов 1963].

В 1961 году в упомянутой выше статье Лукас атаковал машинный редукционизм Патнэма. Интересным представляется то, что Лукас трактовал машинный функционализм как разновидность материализма, к которому он испытывал недоверие. Действительно, в современной философии сознания функционализм считается одной из самых разработанных версий материалистического учения о сознании. Итак, немногочисленные менталисты отвергли тезис об эквивалентности человеческого мышления и машинного, настаивая на том, что человеческое мышление не может быть представлено алгоритмом. Более того, ими утверждается принципиальное превосходство человеческого интеллекта над машинным, которое следует понимать и как невозможность автоматизации интуиции математика.

Гедель в своей лекции в честь Дж.У. Гиббса пришел к следующим менталистским выводам: «1) человеческий ум не способен к формулировке (или механизации) всех математических интуиций; если они сформулированы, появляются новые, например, о непротиворечивости; 2) либо человеческий ум превосходит все машины (может решить больше теоретико-числовых проблем), либо существуют такие теоретико-числовые проблемы, которые неразрешимы для человеческого ума» [Целищев 2021, 25]. Гедель отвергал второй член дизъюнкции о неразрешимых теоретико-числовых проблемах, содержащейся во втором тезисе. Характерным является, что Гедель трактовал математическую интуицию в классическом метафизическом стиле, как прямой доступ к платонистскому универсуму математических объектов [Martin 1993].

Как отмечалось в начале статьи, теорема о неполноте вызвала огромное количество спекуляций, как онтологического, так и откровенно лингвистического характера. Между тем при корректном понимании она ставит фундаментальный вопрос, который Целищев формулирует следующим образом: «Главным исследуемым вопросом является возможность алгоритмизации мышления, если мы исходим из превосходства человека, и алгоритмической природы человеческого мышления, если мы исходим их эквивалентности машины человеку или даже превосходства машины над человеком» [Целищев 2021, 18]. Мышление, как он пишет, в данном случае понимается именно как «математическое мышление», как способность рассуждать с математической определенностью; а теоремы Геделя о неполноте «говорят о фундаментальных ограничениях при попытке полностью формализовать интуитивное знание» [Целищев 2021, 18].

Аргументация механицистов и менталистов вращается вокруг сравнения возможностей конструирования т.н. «геделева предложения» (неразрешимого предложения в богатой формальной системе) для машины и для человека. Лукас полагал, что компьютер не способен построить такое предложение, в отличие от человека. Уверенность Лукаса в том, что именно геделево предложение является ахиллесовой пятой машины, выступает причиной распространённого убеждения, в соответствии с которым теорема о неполноте выходит на первый план при обсуждении проблемы «машина *versus* ум» [Целищев 2021, 44].

Согласно Лукасу, человек способен сконструировать истинное предложение, которое никогда не сможет напечатать компьютер. Иными словами, геделево предложение представляет собой проблему для компьютера. Лукас также считал, что только человеческий разум способен порождать ординальные числа, что способность распознавать ординалы превосходит способность любого формального алгоритма выполнять эту задачу [Lukas 1961, 113].

Изощренная аргументация Лукаса не осталась без внимания Х. Патнэма и П. Бенацераффа, которые обратили внимание на ряд неясностей и нарочитые диалектические рассуждения первого. Одна из претензий состояла в том, что, по утверждению Лукаса, человеческий разум «видит» или «знает» истинность геделева предложения. Подобные эпистемологические модальности способны сбить с толку математика, особенно если они подкрепляются мистическими намеками на трансцендентальную природу восприятия математических объектов и истин. Р. Пенроуз в этом контексте говорит о феномене математического понимания, ко-

торое не может быть полностью сведено к вычислительным методам и к некоему набору правил. Более того, согласно Пенроузу, понимание является функцией нашего сознания. Это дает нам веские основания считать, что сознательное восприятие – процесс «невычислимый» [Penrose 2016, 14]. О разных типах математического мышления – аналитическом и геометрическом – писал еще А. Пуанкаре [Пуанкаре 1990, 399]. Именно геометрическому типу более свойственно находить решения исходя из собственного понимания. Крайним проявлением подобного типа математического мышления, очевидно, были убеждения самобытного индийского математика С.С. Рамануджана, который к моменту его прибытия в Великобританию не понимал, что такое доказательство и зачем оно нужно [Борвейн 1988].

Однако, по словам критиков, главный порок аргументации Лукаса состоит в принятии убеждения собственной непротиворечивости, т.е. непротиворечивости человека, который «видит» истинность геделева предложения, а также обладает способностью различать истину и ложь. Критики сомневаются в ценности таких посылок, не подкрепленных доказательствами непротиворечивости. Согласно Бенацераффу человек может быть просто машиной Тьюринга, но машиной противоречивой [Benaceraff 1967].

С точки зрения Д. Маккаллока, неспособность самоопределиться относительно собственной обоснованности не является принципиальным недостатком разума, независимо от его природы. Для использования теоремы Геделя в получении более общих теорий человеческое математическое мышление нуждается во всё большей собственной формализации, а затем совершает некий прыжок к заключению о том, что такая формализация непротиворечива. Но, если математик формализует слишком много, включая указанные прыжки, то результирующая теория будет способна формализовать себя. В итоге математик неизбежно сделает прыжок к заключению о том, что собственное геделево утверждение истинно. Как мы знаем, это умозаключение немедленно приведет к противоречию.

Таким образом, Маккаллок утверждает, что всякое математическое мышление стоит перед фундаментальным выбором: «Итак, либо (1) математик в какой-то момент перестает формализовать все свои рассуждения (в этом случае совокупность всех фактов, которые он может доказать, будет аксиоматизируемой теорией), либо (2) он формализует все свои рассуждения, и полученная теория будет противоречивой (она сможет доказать свою непротиворечивость)» [McCullough 1995, 64]. Он приходит к неожи-

данному выводу. В соответствии с ним аргументы Р. Пенроуза о том, что наше мышление может быть формализовано, в некотором смысле правильны. Однако приведенные выше ограничения свидетельствуют не об «ущербности» машин и интеллектуальном «превосходстве» человека, а о внутренних ограничениях нашей способности рассуждать о собственном процессе рассуждений. Важно иметь в виду, что «это ограничение не связано с недостатком интеллекта с нашей стороны, а является неотъемлемой частью в любой системе рассуждений, способной рассуждать о себе» [McCullough 1995, 64]. Согласно Целищеву, эти аргументы демонстрируют, что при анализе проблемы непротиворечивости в математическом мышлении различие между человеком и машиной несущественно [Целищев 2021, 187].

Речь идет о таком внутренне присущем человеческому мышлению свойстве, как рефлексивность. Менталисты используют это свойство как свидетельство превосходства естественного интеллекта над искусственным, благодаря которому можно в своих рассуждениях выходить за рамки формальных систем, абстрагироваться от объектного языка на уровень метаязыка, формулировать в метаязыке предложения, недоказуемые в объектном и т.п. Однако Целищев разделяет точку зрения о том, что свойство рефлексивности работает и как внутреннее ограничение, присущее человеческому мышлению как таковому. Существование этого ограничения можно обнаружить при построении трансфинитных объектов. Рефлексивность используется как принцип для построения трансфинитных последовательностей, для рекурсивных ординальных чисел. Суть данного принципа заключается в добавлении геделева предложения для получения непротиворечивой или обоснованной системы. Он позволяет присоединять к теории истинное, но недоказуемое в ней предложение, что ставит вопрос о формальной возможности постоянной итерации принципа рефлексии через рекурсивные ординалы. Все это означает в конечном счете, что «попытка получить гарантии непротиворечивости и обоснованности путем формализации мышления обречена на неудачу» [Целищев 2021, 18].

Лукас парировал аргументы о возможности противоречивой природы человеческого интеллекта, трактуя ее не как фундаментальный онтологический порок, а как простой и досадный сбой в его работе, банальную ошибку: «Когда непротиворечивость человека ставится под сомнение, наивная реакция заключается в том, чтобы рьяно на ней настаивать: но это, в свете второй теоремы Геделя, принимается некоторыми философами как

свидетельство его действительной непротиворечивости. Профессор Патнэм предположил, что люди – это машины, но машины противоречивые… И мы не можем упрекнуть его в непоследовательности – разве люди не противоречивы? Конечно, женщины и политики; и даже мужчины не политики иногда противоречат самим себе и, единственного противоречия достаточно, чтобы сделать противоречивой всю систему. Нельзя отрицать тот факт, что все мы иногда противоречивы, но из этого не следует, что мы равнозначны противоречивым системам. Наши противоречия – это скорее ошибки, чем изначальные установки (*set policies*). Они соответствуют периодическим сбоям в работе машины, а не ее нормальной схеме работы. Свидетельством тому является то, что мы избегаем противоречий, когда распознаем их в качестве таковых. Если бы мы действительно были бы противоречивыми машинами, мы бы не корректировали содержание из-за наших противоречий и с радостью утверждали бы оба противоречивых суждения. Более того, можно было бы утверждать все, что угодно, но мы этого не делаем. Было легко продемонстрировано, что в противоречивой формальной системе доказуемо все, и требование непротиворечивости оборачивается просто тем, что не все в ней можно доказать – это явно не тот случай, когда "все сгодится". Это, безусловно, характерная черта умственных операций людей: они являются избирательными. Они действительно различают между предпочтительным, истинным, неблагоприятным и ложным высказываниями: когда человек готов утверждать все, что угодно, и готов противоречить самому себе без всяких зазрений совести или отвращения, он считается "сошедшим с ума". Люди, хотя и не совершенно непротиворечивы, они и не настолько противоречивы, сколько подвержены ошибкам» [Lukas 1961, 120–121].

Несмотря на то, что механицисты отступили от уничижительных атак против менталистов после привлечения авторитета Курта Геделя, Целищев оценивает спор о превосходстве человеческого интеллекта над машинным незавершенным. Поскольку этот спор касается одного из самых фундаментальных вопросов, лежащих в основании математики и логического мышления вообще; причем, – к сожалению многих он может быть признан диалектическим, надежд на то, что в нем может быть поставлена точка, – немного. Р. Смаллиан посвятил геделевской проблематике книгу с говорящим названием «Вовеки неразрешимое», в которой сконструировал новые, – ранее «немыслимые» свойства мышления и даже предложил иные формы

разумности с крайне странными эпистемическими свойствами [Smullyan 1987].

Смаллиановская стабильность и «нулевой» уровень разума

Исследуя вопрос о рефлексивных свойствах машин, Смаллиан приходит к неожиданному выводу о том, что подлинная разумность начинается на более высоких рефлексивных уровнях, чем те, которым в человеческом ментальности соответствуют сознание и даже самосознание. Следует отдавать отчет в том, что, размышляя над геделевской проблематикой с помощью средств эпистемической логики, Смаллиан использует специфически понимаемую модальность «верит» и стандартный инструментарий пропозициональной логики. Он применяет эпистемический оператор «B», так что под выражением «Bp» следует понимать предложение, в которое «верит» некий Мыслитель. Одновременно «Bp» является предложением, которое доказуемо в системе. Обратим внимание на то, что отождествление веры и доказуемости выступает самым слабым звеном рассуждений с точки зрения интерпретации последующих результатов. Однако этот вопрос требует отдельного тщательного исследования [Smullyan 1987, 166–167].

Иными словами, именно такое рафинированное свойство разума, как осознание самосознания, является наиболее фундаментальной формой мышления, своеобразным «нулевым» уровнем разумности, по отношению к которому даже самосознание оказывается отрицательным уровнем мышления. Причина этого состоит в том, что, начиная с четвертого рефлексивного ранга, мыслители (которых можно понимать и как машины или формальные дедуктивные системы) могут средствами логики высказываний доказать себе, что они принадлежат к четвертому типу. Мыслители более низких рангов не осведомлены о рефлексивном ранге, к которому они принадлежат. Рефлексивные ранги ниже осознания самосознания можно считать отрицательными уровнями мышления в том смысле, что, редуцируясь к ним, интеллект существенно теряет свою полноту.

Мы привыкли рассматривать истину как понятие эпистемологическое. Но существует разновидность истины, которую можно трактовать онтологически. Как ни странно, это – истина эпистемическая, т.е. истинность утверждения формы «*x* верит что *p*». Истинность конкретного убеждения или знания можно понимать как состояние конкретной личности или машины. Теоретически это состояние можно детектировать, что практически

делают с неплохой вероятностью в отношении людей средствами полиграфа.

Однако в настоящее время детекция лжи используется исключительно для выяснения объективной истины – первопорядковой. Но что будет, если, имея в распоряжении идеальный полиграф, испытуемым задавать самореферентные вопросы, ведущие к противоречиям и парадоксам? Очевидно, мы получим некое распределение ответов. Будет ли это знание иметь ценность? Если будет, то какое?

Снова уместно обратиться к модели Смаллиана. Он ввел в оборот два странных эпистемических свойства – стабильность и нормальность: «Мы называем Мыслителя стабильным, если для каждого предложения *p*, если он верит в *p*, тогда он на самом деле верит в *p*. Мы называем мыслителя нестабильным, если он не является стабильным, т.е., если имеется, по крайней мере, одно предложение *p* такое, что Мыслитель верит, что он верит в *p*, но на самом деле он не верит в *p*… нестабильность представляется столь же странной психологической характеристикой, как и странность. <…> Мы отметим, что стабильность обратна по отношению к нормальности. Если нормальный Мыслитель верит в *p*, тогда он верит в *Bp*, в то время как если стабильный Мыслитель верит в $\underline{Bp}$, тогда он верит в *p*» [Smullyan 1987, 166–167].

Если Мыслитель *нормальный*:
(1) B*p*→BB*p*
Если Мыслитель *стабильный*:
(2) BB*p*→B*p*.

Действительно, нестабильность является странной психологической характеристикой, но настолько ли невозможной? Убеждения могут возникать в метаязыке и на высоких рефлексивных рангах мышления. Означает ли это, что они «выживут» при снижении ранга рефлексии и упрощении языка до объектного уровня? Подобного рода драматический процесс отказа от первоначальных убеждений можно нередко наблюдать при попытках натурализации утверждений высокого уровня абстракции. Например, если некто утверждает, что он верит в троичность бога, обычно это означает, что он считает себя верующим и фактически верит в то, что он верит в Троицу. При внимательном рассмотрении часто оказывается, что эта вера не имеет достаточных оснований, – ее носитель вообще не понимает, что такое догмат о Троице и, следовательно, его веру следует считать ложной. Возможна и иная

ситуация: «Очевидно, что существуют математики, которые верят, что они верят, что первая теорема Геделя истинна. Может оказаться, что проверка таких математиков на детекторе лжи выявит, что некоторые из них на самом деле не верят в истинность этой теоремы. Это вполне представимо, т.к. известно, что человек может иметь убеждение на сознательном уровне, подкрепляемое множеством дополнительных сознательных допущений, но при этом не иметь глубокой убежденности, испытывать серьезные подсознательные сомнения в истинности своего убеждения» [Винник 2015, 82].

Характеристика стабильности может быть применена и к техническим экспертным системам комбинированного типа, состоящих из двух контуров: 1) интеллектуально непрозрачного нейросетевого перцептрона, распознающего и классифицирующего данные; 2) классического алгоритмического, осуществляющего логические выводы на основании данных первого уровня и иных вводных. Несложно представить себе ситуацию, если алгоритмические выводы вступят в логический конфликт с информацией на выходе нейронной сети, утверждающей нечто без прозрачных рациональных оснований. Такое состояние системы можно понимать как нестабильность. Если использовать это в качестве аналогии, то человеческую психику следует отнести к комбинированным системам подобного рода, на что обратил внимание еще Аристотель в учении о трех видах души.

Средства детекции лжи обычно применяют для нужд следствия и кадровых целей. Представляется возможным использование этих средств для оценки искренности экспертов в экспертных опросах, например, с целью повышения надежности дельфийского метода. Если допустить, что истина объектных суждений зависит от веры экспертов, что они верят в истинность этих суждений, то полиграф теоретически можно использовать для калибровки экспертных вопросов.

Высказывания ВВ p могут трактоваться как публичные утверждения экспертов, что p. Если имеет место ВВp, то результаты экспертного мнения принимаются как «истинно, что p». В основе этого лежит допущение о том, что эксперты высказывают только формально доказуемые утверждения. Утверждения типа Вp можно понимать как внутренние убеждения. Теоретически сверхточный полиграф способен установить Вp или ¬Вp, т.е. *принимает* он это убеждение как истинное или нет.

Мы не владеем знаниями о распределении свойств *смаллиановских нормальности и стабильности* в реальной мыслительной

деятельности. Детекция ложного, но правильно обоснованного убеждения будет означать, что эксперт нестабилен. Рассмотрим случай, когда эксперт верит в истинность собственного заключения, являющегося формально ложным, т.е. неправильно обоснованным. В этом случае эксперта следует признать ненормальным в смаллиановско м смысле.Не вполне понятен вопрос о том, какое свойство хуже – ненормальность или нестабильность, исходя из требований доказуемости. Ответ зависит от того, какую философию математики следует признать истинной – интуитивистскую или конструктивистскую.

Согласно интуитивизму наиболее неприятным эпистемическим свойством следует считать нестабильность. Убеждения, не основанные на глубоких интуициях, не имеют цены. Ненормальность с точки зрения этой философии есть менее скверное свойство, за исключением одной «мелочи»: «Иное дело, когда такое имеет место при повышении ранга рефлексии, т.е. когда субъект не способен транслировать свои убеждения на более высокий рефлексивный уровень…» [Винник 2015, 94].

Если придерживаться конструктивистской философии математики, очевидным становится, что нестабильность следует признать менее негативным свойством, чем ненормальность. Для конструктивиста не является важным, по каким причинам эксперт не верит в свои убеждения, даже если доказательство корректно. Он может не принимать истинность аксиом: «Гораздо важнее, что истинность суждений подкреплена на более высоком рефлексивном ранге, в которые “вмонтирована” мощная формальная система обоснования предложений. То, что эти убеждения не могут быть погружены до самых оснований, вряд ли имеет большое значение, т.е. нарушение сохранения истинности при понижении ранга рефлексии не является критическим феноменом» [Винник 2015, 94].

Заключение

В.В. Целищев видит в сопоставлении человеческого мышления и «мышления» компьютера крайность с точки зрения здравого смысла и даже определенную парадоксальность. Сначала было принято допущение относительно того, что мозг подобен компьютеру. Однако в дальнейшем стало очевидным, что наши знания о вычислительных машинах гораздо более точны и полны, чем наши знания о мозге. По этой причине сравнения малопродуктивны: «Но существует и более радикальная позиция, согласно которой сопоставление вычислительной машины и человека

совершенно неправомерно. Т.к. человек не обладает ни полнотой, ни непротиворечивостью, утверждение, что этими качествами не обладает и компьютер, ничего не дает при попытке их сопоставления» [Целищев 2021, 16]. Действительно, наличие общих логически негативных свойств у двух объектов не может быть достаточным основанием для вывода об их общей природе. Можно сделать вывод, что т.н. геделевский аргумент взят на вооружение критиками вычислительного редукционизма, но оказался обоюдоострым оружием. Его легко применили против них самих, причем в специфической манере, предусматривающей вменение необходимости доказательства собственной непротиворечивости. К тому же со стороны механицистов это вменение носит поистине иезуитский характер, поскольку исходит не из попыток переложить бремя доказывания, а из расчета на неизбежные парадоксальные следствия.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Борвейн, Борвейн 1988 – *Борвейн Дж., Борвейн П.* Рамануджан и число π // В мире науки. 1988. № 4. С. 58–66.

Винник 2015 – *Винник Д.В.* Эпистемическая ложь как онтологическое понятие // Философия науки. 2015. № 2. С. 70–88.

Винник 2016 – *Винник Д.В.* «Эпистемический искупитель»: Свод приемов легитимации бессмыслицы // Философия науки. 2016. № 4. С. 76–95.

Докинз 2013 – *Докинз Р.* Разоблачение постмодернизма // *Докинз Р.* Капеллан дьявола. Размышления о надежде, лжи, науке и любви. – М.: Аст, 2013. С. 78–89.

Куайн 1996 – *Куайн У.В.О.* Онтологическая относительность // Современная философия науки. – М.: Логос, 1996. С. 40–61.

Миркес 1998 – *Миркес Е. М.* Логически прозрачные нейронные сети и производство явных знаний из данных // Нейроинформатика. – Новосибирск: Наука. Сибирское предприятие РАН, 1998.

Пенроуз и др. 2004 – *Пенроуз Р., Шимони А., Картрайт Н., Хокинг С.* Большое, малое и человеческий разум. – М.: Мир, 2004.

Пуанкаре 1990 – *Пуанкаре А.* Наука и метод // О науке. – М.: Наука, 1990. С. 368–522.

Уёмов 1963 – *Уёмов А.И.* Вещи, свойства и отношения. – М.: Издательство Академии наук СССР, 1963.

Целищев 2021 – *Целищев В.В.* Алгоритмизация мышления: Геделевский аргумент / изд. 2-е, испр. – М.: ЛЕНАНД.

Benacerraf 1967 – *Benacerraf P.* God, Devil, and Gödel // Monist. 1967. Vol. 51. No. 1. P. 9–32.

Lucas 1961 – *Lucas J.* Minds, machines, and Gödel // Philosophy. 1961. Vol. 36. No. 137. P. 112–127.

Martin 1993 – *Martin D.* How Subtle is Gödel's Theorem? // Behavioral and Brain Sciences. 1993. Vol. 16. No. 3. P. 611–612.

McCullough 1995 – *McCullough D.* Can Humans Escape Gödel? A Review of *Shadows of the Mind* by Roger Penrose // Psyche. 1995. Vol. 2. No. 4. P. 57–65.

McCulloch, Pitts 1943 – *McCulloch W.S., Pitts W.* A Logical Calculus of the Ideas Immanent in Nervous Activity // The Bulletin of Mathematical Biophysics. 1943. Vol. 5. No. 4. P. 115–133.

Moor 2006 – *Moor J.* The Dartmouth College Artificial Intelligence Conference: The Next Fifty years // AI Magazine. 2006. Vol. 27. No. 4. P. 87–91.

Ng, Leung 2020 – *Ng W.C., Leung W.C.* Strong Artificial Intelligence and Consciousness // Journal of Artificial Intelligence and Consciousness. 2020. Vol. 7. No. 1. P. 63–72.

Penrose 2016 – *Penrose R.* The Emperor's New Mind Concerning Computers, Minds, and the Laws of Physics. – Oxford: Oxford University Press, 2016.

Putnam 1960 – *Putnam H.* Minds and Machines // Dimensions of Minds / ed. by S. Hook. – New York: New York University Press, 1960. P. 138–164.

Smullyan 1987 – *Smullyan R.* Forever Undecided: A Puzzle Guide to Godel. – Oxford: Oxford University Press, 1987.

REFERENCES

Benacerraf P. (1967) God, Devil, and Gödel. *Monist.* Vol. 51, no. 1, pp. 9–32.

Borwein J.M. & Borwein P.B. (1988) Ramanujan and Pi. *Scientific American.* Vol. 258, no. 2, pp. 112–117 (Russian translation: S*cientific American.* 1988. No. 4, pp. 58–66).

Dawkins R. (2013) Postmodernism Disrobed. In: Dawkins R. *A Devil's Chaplain: Reflections on Hope, Lies, Science and Love* (pp. 78–89). Moscow: Ast (Russian translation).

Lucas J. (1961) Minds, machines, and Gödel. *Philosophy.* Vol. 36, no. 137, pp. 112–127.

Martin D. (1993) How Subtle is Gödel's Theorem? *Behavioral and Brain Sciences.* Vol. 16, no. 3, pp. 611–612.

McCullough D. (1995) Can Humans Escape Gödel? A Review of *Shadows of the Mind* by Roger Penrose. *Psyche.* Vol. 2, no. 4, pp. 57–65.

McCulloch W.S. & Pitts W. (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics.* Vol. 5, no. 4, pp. 115–133.

Mirkes E.M. (1998) Logically Transparent Neural Networks and The Production of Explicit Knowledge from Data. In: Gorban A.N., Dunin-Barkovsky V.L., & Kirdin A.N. (Eds.) *Neuroinformatics* (pp. 283–292). Novosibirsk: Nauka (in Russian).

Moor J. (2006) The Dartmouth College Artificial Intelligence Conference: The Next Fifty years. *AI Magazine.* Vol. 27, no. 4, pp. 87–91.

Ng G.W. & Leung W.C. (2020) Strong Artificial Intelligence and Consciousness. *Journal of Artificial Intelligence and Consciousness.* Vol. 7, no. 1, pp. 63–72.

Penrose R. (2016) *The Emperor's New Mind Concerning Computers, Minds, and the Laws of Physics.* Oxford: Oxford University Press.

Penrose R., Shimony A., Cartwright N., & Hawking S. (2000). *The Large, the Small and the Human Mind.* Cambridge, UK: Cambridge University Press (Russian translation: Moscow: Mir, 2004).

Poincaré J.H. (1990) Science and Method. In: Pontryagin L.S. (Ed.) *On Science* (pp. 368−522). Moscow: Nauka (Russian translation).

Putnam H. (1960) Minds and Machines. In: Sidney Hook (Ed.) *Dimensions of Minds* (pp. 138−164). New York: New York University Press.

Quine W.V.O. (1996) Ontological Relativity. In: Pechyonkin A.A. (Ed.) *Modern Philosophy of Science: Knowledge, Rationality, Values in the Works of Western Thinkers: An Educational Textbook* (pp. 40–61). Moscow: Logos (Russian translation).

Smullyan R. (1987) *Forever Undecided: A Puzzle Guide to Godel*. Oxford: Oxford University Press.

Tselishchev V.V. (2021) *Algorithmization of Thinking: The Gedelian argument.* Moscow: LENAND (in Russian).

Uemov A.I. (1963) *Things, Properties and Relations.* Moscow: USSR Academy of Sciences Press (in Russian).

Vinnik D.V. (2015) Epistemic Lie as an Ontological Conception. *Filosofiya nauki*. No. 2, pp. 70–88 (in Russian).

Vinnik D.V. (2016) "Epistemic Redeemer": The Code of Tricks Used to Legitimate Nonsense. *Filosofiya nauki.* No. 4, pp. 76−95 (in Russian).