Философия науки и техники
2024. Т. 29. № 1. С. 20–33
УДК 165.12

Philosophy of Science and Technology
2024, vol. 29, no. 1, pp. 20–33
DOI: 10.21146/2413-9084-2024-29-1-20-33

Ю.С. Моркина

Сознание как антиномия (антиномичность понятия «сознание» и философия искусственного интеллекта)

Моркина Юлия Сергеевна **–** кандидат философских наук, старший научный сотрудник. Институт философии РАН. Российская Федерация, 109240, г. Москва, ул. Гончарная, д. 12, стр. 1; e-mail: morkina21@mail.ru

В статье показано, что причины противоречий, существующих в философии искусственного интеллекта (ИИ), коренятся в недостаточной отрефлектированности общефилософских оснований. В числе таких оснований – понятия «сознание» и «другое сознание». Проанализированы различные подходы к интерпретации данных понятий, проведена логико-методологическая реконструкция дискуссионного столкновения различных подходов к проблеме ИИ и сознания. Приводится критика мысленных экспериментов, использующихся в философии сознания и ИИ. Продемонстрировано существенное методологическое отличие таких экспериментов от эксперимента Галилея, который является классическим случаем успешного мысленного эксперимента, давшего объективный результат. При концептуализации и соотнесении понятий «сознание», «Я» и «Другой» неизбежны эпистемологические разрывы, обусловленные качественными различиями феноменологических проявлений и способов концептуализации сфер реальности, которые необходимо соотнести между собой. Так, показано, что эпистемологический разрыв неизбежно присутствует между внешним (в том числе вербальным) поведением Другого и его интерпретацией наблюдателем как проявления определенных внутренних содержаний сознания Другого. Также феноменологические проявления для человека его собственного сознания и проявления сознания Другого (в том числе вербальное описание его содержаний сознания) теоретически не соотносимы. Показана антиномичность понятия «сознание» как невозможность заключения о его наличии или отсутствии у системы (Я, Другой, ИИ) при использовании чисто философских методов рассуждения. Понятие сознания, как и представления о сознании Другого, однако, необходимы для практической деятельности и коммуникации, поэтому человек пользуется практическими способами концептуализации сознания, включающими интуитивное восприятие, эмпатию, приписывание и т.д.

© Моркина Ю.С., 2024

Ключевые слова: сознание, искусственный интеллект, философские основания, смыслополагание, аналогизирующая апперцепция, мысленный эксперимент, тест Тьюринга, антиномия

1. Проблема философии ИИ. Тест Тьюринга

Данная работа началась с попытки рефлексии оснований философии искусственного интеллекта (ИИ), но, отправляясь от размышления над возможностями применения теста Тьюринга, она вывела на общефилософские проблемы сознания и «другого сознания».

Сперва я задалась вопросом, какими качествами должна обладать «умная» компьютерная программа, чтобы мы признали за ней наличие интеллекта. В соответствии с определением основоположника философии ИИ А. Тьюринга некую систему можно назвать «умной», если по ответам и реакциям ее невозможно отличить от человека (так называемый тест Тьюринга, впервые предложенный им в работе «Вычислительные машины и разум»[1]) [Тьюринг, 2003].

Сам Тьюринг в своем труде, заложившем фундамент этого «машинного дискурса», подчеркивает, что считает бессмысленным вопрос о том, «может ли машина мыслить», поскольку само понятие мышления также еще нуждается в определении. Он лишь утверждает, что возможно построить машину, которую по ответам и реакциям на задаваемые ей «экзаменатором» вопросы, данный «экзаменатор» не сможет отличить от человека [Там же, с. 47, 50]. Но должно ли устройство действительно обладать интеллектом в каком-нибудь другом смысле для того, чтобы его по ответам и реакциям нельзя было отличить от человека? Например, можно было бы определить интеллект как способность к решению проблем или принятию решений [Майнцер, 2016, с. 469–508].

Думается, чтобы экзаменатор (человек) не мог отличить тестируемое устройство от человека, достаточно, чтобы устройство располагало обширной базой возможных ответов, разделенной на кластеры с привязкой каждого кластера ответов к определенным ключевым словам в вопросах и программой, позволяющей случайным образом выбирать и выдавать любой ответ из каждого кластера. Экзаменатору такие ответы машины будут казаться выдаваемыми к месту, необходимо только, чтобы на протяжении некоторого времени они не слишком часто повторялись. Если обеспечить достаточное разнообразие и относительную связность (за счет привязки к ключевым словам) ответов, экзаменатор, по крайней мере, не сможет провести достаточно четкое для доказательства различие между ответами машины и ответами человека. В чем причина этого? Экзаменатор – человек, т.е. существо, само способное к смыслополаганию. Но такое существо принципиально не может доказать отсутствие такой способности у Другого. Здесь эмпирические и экспериментальные вопросы отодвигаются на второй план, уступая место общефилософским положениям.

---

[1] «Вычислительные машины и разум» (англ. «Computing Machinery and Intelligence») – основополагающая работа в области искусственного интеллекта, опубликованная в 1950 г. в журнале «Mind» и дающая представление о том, что называется тестом Тьюринга.

Именно природа человеческого существа как существа смыслополагающего мешает ему четко отличить человекоподобно реагирующие сущности от Другого, подобного себе. Человек склонен заключать по аналогии (и при этом по аналогии с собой), усматривая в Другом также сознающее, чувствующее, размышляющее существо и делая по внешним человекоподобным реакциям вывод о наличии у Другого сознания и его содержаний. И пусть на гипотезах о содержаниях сознания Другого строится фолк-психология, способность именно так заключать и строить такие гипотезы лежит в основе человеческой социальности и определяет возможность интерсубъективного общения.

Передача знаний о мире (эпистемологический аспект), научный дискурс (социально-эпистемологический аспект), общение (социальный аспект) – все это основано на способности человека к смыслополаганию, а также заключению по аналогии о содержаниях сознания Другого и самом наличии у Другого сознания. Но именно из способности к смыслополаганию вытекает неспособность к усмотрению отсутствия смысла. Сама бессмысленность предстает для смыслополагающего существа как разновидность смысла. (Пример из литературы: абсурд поднимает литературное действие до выражения высшего экзистенциального трагизма). Совсем же бессмысленный ответ машины экзаменатор может истолковать как иронию. Отсутствие смысла оказывается для смыслополагающего существа формой небытия. А небытия нет. Итак, смыслополагающее существо неспособно доказать абсолютную бессмысленность чего бы то ни было.

Рассуждая теоретически, можно прийти к парадоксальному выводу. Философы ИИ спорят о наличии или отсутствии сознания у ИИ-устройств или гипотетической возможности создания искусственного сознания. Но как мы сможем доказать наличие сознания у чего-либо, если мы не можем доказать его отсутствия? То есть гипотеза о наличии у человекоподобно реагирующего устройства сознания, субъективной реальности, способности к смыслополаганию – нефальсифицируемая гипотеза.

Способность к смыслополаганию выявляема только при помощи интроспекции и, по-видимому, нет внешних ответов и реакций, по которым можно было бы безошибочно заключить о наличии или отсутствии такой способности. Иными словами, она может быть сымитирована, и возможность такой имитации заложена в сущности языка и особенностях человеческого восприятия. Смыслополагание (принимаемое нами в качестве человеческой способности, которой не может обладать запрограммированный компьютер) остается за кадром не только поведенческих реакций, но и происходящих (например, в мозге или компьютере) процессов. Таким образом, неспособность ИИ к смыслополаганию, отличающая его от человека как смыслополагающего существа, остается недоказуемым (но и не опровергаемым) положением.

Но если это положение не только недоказуемо, но и не фальсифицируемо, то оно не научно. Тогда таково же и предположение о наличии (или отсутствии) у сущности сознания. Если положение не научно, означает ли это, что оно и философским быть не может?

Проблема состоит еще и в том, что определение интеллекта Тьюринга содержит логический круг. Переформулируя его и не меняя при этом смысла,

можно получить утверждение: *устройство обладает интеллектом, если его по ответам и реакциям нельзя отличить от существа, которому мы априорно приписываем неотъемлемое свойство обладания интеллектом*.

Возможны такие устройства, которые экзаменатор (человек) по ответам и реакциям не сможет отличить от человека (самого себя или подобного себе существа) и которые при этом даже не обладают и подобием интеллекта, если интеллект определить иначе, чем Тьюринг. Эти устройства человек не сможет не признать в чем-то себе подобными, и одновременно в них не будет заложена ни способность решать какие-либо задачи (хотя бы математические), ни способность принимать решения, ни чего-либо иного, что было бы практически полезно человеку. Такие устройства будут преследовать одну цель: демонстрировать человекоподобное поведение.

Причина этого заложена, как уже было сказано, и в сущности языка. Язык позволяет человеку формулировать свои знания и выражать другие содержания сознания. Если устройство использует язык для имитации реакций человека, то такую имитацию нельзя отличить от подлинной реакции еще и за отсутствием внеязыкового основания для различения. Если машина говорит, что ей больно, и человек говорит, что ему больно, мы можем предполагать, но не рационально доказать, что первая только имитирует реакцию второго. Особенно, если машина заявит нечто вроде: «Мне больно это слышать». Если задаться вопросом, какие поведенческие (в данном случае, относящиеся к вербальному поведению) реакции человека не сможет сымитировать компьютерная программа при заданной программистам цели создания такой имитации, ответ напрашивается пессимистический: по-видимому, никакие. Правда, некоторые философы ИИ сочтут этот ответ оптимистическим…

Итак, почему при столь развитой современной философии ИИ приходится возвращаться к Тьюрингу? Потому что для того, чтобы увидеть корень проблем и причину противостояний разных направлений в обширной литературе по философии ИИ, полезно обратиться к истокам. При этом обнаруживается, сколько проблем, связанных с философией ИИ, уже заложено в самом ее основании.

Но при внимательном анализе оказывается, что основные проблемы, которые пытается решить философия ИИ, коренятся даже не в ее основаниях. Причина сложности проблем, поднимаемых философией ИИ, кроется в «вечности» общефилософских проблем. Это проблемы, связанные с исследованием сознания, полаганием человеческим сознанием Другого как также обладающего сознанием, сомнениями в правомочности применения и вместе с тем неизымаемости интроспекции как метода изучения сознания и его содержаний. Это также тот факт, что некоторые способности человека – к эмпатии, к смыслополаганию, к заключению по аналогии с собой – можно было бы, сменив ракурс рассмотрения, увидеть даже не как склонности, но как неотъемлемые свойства человеческого восприятия.

Попытки трансцендировать сознание как обладающее подобными свойствами восприятия Другого предпринимались в психологии, однако в философии ИИ, скорее, превалируют противоположные тенденции: неявный упор на эти свойства.

2. Критика мысленных экспериментов в философии ИИ

Рассуждая о природе сознания, возможности наличия сознания у ИИ, философы используют такой метод, как мысленный эксперимент. Классический пример успешного мысленного эксперимента, приведшего к бесспорному результату, – мысленный эксперимент Галилея. Этот эксперимент действительно позволил вывести положение, оказавшееся подтвержденным эмпирически и легшее в основу открытого им физического закона.

Философы ИИ тоже проводят мысленные эксперименты. Почему же они не достигают столь четких и бесспорных результатов, настолько же интерсубъективно убедительных, как эксперимент Галилея? Попробуем разобраться, в чем здесь отличие.

Возьмем как пример мысленный эксперимент с заменой живых клеток человеческого мозга на электронные чипы. Философы, ставившие такой эксперимент, не приходят к общему интерсубъективно доказательному результату. Вот, по крайней мере, два одинаково возможных результата:

1) если клетки головного мозга (нейроны) постепенно заменять на электронные чипы так, что в конце концов все они окажутся замененными и мозг человека станет электронным, то сознание полностью исчезнет, а поведение останется прежним, и получится «философский зомби»;

2) при выполнении той же процедуры сознание в ставшем электронным мозге полностью сохранится, поскольку не изменится поведение. И, следовательно, сознание как информационный феномен не зависит от носителя (он может быть как биологическим, так и электронным).

Что здесь не так? Какой результат верен? Очевидно, каждый выберет для себя результат, который ему больше нравится и который согласуется с его общей философской позицией. Но ни один из них не обладает бесспорностью и интерсубъективной убедительностью результата Галилея. Определяющим при выводе и принятии результата такого эксперимента становятся ценностные предпочтения рассуждающего. Но что будет, если поставить такой эксперимент не мысленно, а эмпирически? Пока на современном уровне развития науки это невозможно.

Отличие же и очень серьезное здесь в том, что Галилей в своем мысленном эксперименте получил несомненное логическое противоречие (А и не-А), два взаимоисключающих утверждения: если тяжелые предметы падают быстрее, чем легкие, то два связанных вместе предмета – легкий и тяжелый – падают медленнее одного тяжелого, поскольку легкий предмет замедляет падение тяжелого (А); и система из легкого и тяжелого предметов падает быстрее одного тяжелого, поскольку вес двух предметов в сумме больше, чем одного (не-А).

Но результаты большинства мысленных экспериментов, связанных с проблемой сознания и ИИ, не содержат таких противоречий. Они зависят только от способности или неспособности экспериментатора вообразить какую-либо ситуацию. Так, кто-то из философов как эмпирический субъект, личность с определенным складом ума, стилем мышления и убеждениями (в том числе, философскими) может вообразить себе такую сущность, как «философский зомби», а кто-то не может (проблема «мыслимости» «философских зомби»).

Итак, мысленные эксперименты подобного рода не выводят на логические противоречия, вообразимость же ситуации, описываемой в них, зависит от личностных качеств философа.

Разберем понятие «философский зомби». Под «философским зомби» понимается мыслительный конструкт (в связи с ним при этом возникает проблема его «мыслимости»), который используется философами сознания и философами ИИ в мысленных экспериментах. «Философский зомби» определяется как сущность (система), функционально неотличимая от человека, но при этом не обладающая всего одним из человеческих свойств – сознанием.

По отношению к этому понятию А.Ю. Алексеев подразделил всех философов, которые рассуждают о «философских зомби», на три лагеря. «Зомбисты» утверждают самое меньшее – мыслимость «философских зомби», самое большее – возможность реального существования систем, представляющих собой таких «зомби». Алексеев приводит некоторые фундаментальные выводы из подобной позиции: если возможны поведенческие зомби, то ложен бихевиоризм; если возможны физические зомби, ложен физикализм; проблема Другого неразрешима, т.к. мы оцениваем не субъективную реальность Другого, но лишь его двойника-зомби (внешнее поведение); сознание несущественно для эволюции природы; изучение сознания для прогресса когнитивно-компьютерных технологий неинтересно. «Антизомбисты» стоят на позиции принципиальной немыслимости «философских зомби». «Нейтральные зомбисты» требуют внести ясность в проблематику сознания и терминологию, прежде чем заключать о возможности или невозможности «зомби» [Алексеев, 2013, с. 126–133].

Если (предположим) и в самом деле возможно существование «философских зомби», т.е. таких сущностей, которые мы по поведению неспособны отличить от нас как от сознательных существ (а некоторые философы даже рассуждают в том ключе, что мы и сами такие «философские зомби»), возникает экзистенциальный вопрос: если все в нас могло бы отлично функционировать без нас, все наше поведение могло бы быть точно таким же, если бы нас в нас не было (включая творческие проявления и созидание культуры), тогда *что мы в нас делаем*? При таком предположении сознание остается за скобками. Человек оказывается неспособным отличить себя ни от явлений природы, ни от компьютерных программ с человекоподобным поведением, ни от «философских зомби», которых преподносит воображение философов. Даже если на самом деле лишь он сам наделяет смыслом явления природы и поведение машин, то он не может это теоретически доказать.

Но тогда в чем же отличие (именно теоретическое) наделения смыслом поведения компьютерных программ от приписывания сознания другим людям? Даже тест Ватта (инвертированный тест Тьюринга) на самом деле не работает. С. Ватт утверждает, что в отличие от человека машина неспособна приписывать сознание другой сущности, такой способностью обладает только человек и по этому принципу мы можем отличить человека от машины [Там же, с. 81]. Интуитивно с Ваттом можно согласиться. Действительно, только человек способен к смыслополаганию и приписыванию сознания отличным от себя или сходным с собой сущностям.

Проблема возникает из-за того, что и такое приписывание (как и сам феномен сознания или смыслополагания) неуловимо непосредственным образом и выражается в определенных формах поведения, в том числе вербального. А это значит, и феномен приписывания сознания можно симулировать, если машина будет генерировать выражения вроде «я понимаю, что наши сознания непохожи и мое сознание отлично от человеческого». И в самом деле, современные боты могут генерировать в том числе и такие рассуждения, опираясь всего лишь на анализ человеческого вербального поведения и черпая материал из интернета.

А что же тогда мы? Это «мерцание внутри», которое мы в себе чувствуем и которое так склонны видеть во многих сущностях, даже и не обязательно похожих на нас? Если на нем *сосредоточиваться долго, еще и обладая современными философскими знаниями, то мы и* сами для себя распадаемся на некий набор параметров и функций. Иначе философы всерьез бы не выдвигали такие концепции сознания, как, например, лингвистическая. Нам приходится признать, что они в интроспекции нашли возможным проблематизировать наличие сознания и Я у самих себя.

Можно удивляться философам, в интроспекции не находящим у себя самих сознания или, по крайней мере, такого сознания, в наличии которого нельзя теоретически усомниться. Можно считать, что это уж слишком. Но присмотритесь к собственному сознанию и спросите себя: как вы самому себе докажете, что оно у вас есть? И что при этом оно едино и непрерывно (даже автору этой статьи в интроспекции видно, что это может оказаться не так). К тому же все зависит от изначальных мировоззренческих установок проводящего такую интроспекцию человека. Если я считаю себя «завихрением атомов» или набором простых восприятий, с которыми оперирует нечто вроде компьютера, выдавая мои собственные состояния как реакции, то тогда что же представляет собой мое сознание? И обязательно ли оно для того, чтобы я вел себя так, как себя веду? Если я только игрок в языковых играх – зачем мне оно? И все указывает на то, что его у меня нет. «Честно говоря, мне даже неясно, на что похоже быть мною в данный момент» [Хофштадтер, Деннет, 2003, с. 370]. И доказать, что меня во мне нет, что сознания не существует или же оно несущественно, выстраивая рациональную систему рассуждений, возможно, даже намного легче, чем доказать обратное. Теоретически парировать такие аргументы затруднительно. Особенно, если при этом находиться на сходных мировоззренческих позициях.

3. Аналогизирующая апперцепция

Остановимся на феноменологическом понятии аналогизирующей апперцепции. Это понятие, предвосхищенное Э. Махом [Max, 2021, с. 64], подробно разработано Э. Гуссерлем [Гуссерль, 2000, с. 433–515]. Его суть – в заключении по аналогии, производимом трансцендентальным сознанием, обладающим, однако, телесностью, когда оно наблюдает телесные проявления Других, сходные со своими телесными проявлениями. Аналогия приводит такое сознание к предположению о наличии за этими сходными телесными проявлениями сходной сознательности:

> В трансцендентальной теории опыта Другого смысловой образец как матрица смыслов всех возможных Других представляет собою не просто человеческое тело, но нераздельное психофизическое единство, уникальное взаимоотношение сознания с собственной телесностью. Поэтому аналогизирующая апперцепция… уподобляет друг другу не тела, а процессы непрерывно возобновляющейся данности тела самому себе [Смирнова, 2014, с. 191].

При этом Гуссерль подчеркивает спонтанность аналогизирующей апперцепции как проявления одного из свойств человеческого мышления. Для него это ни в коем случае не индуктивный вывод и не рассуждение по аналогии – это вообще не вывод и не рассуждение. Оно сродни ассоциативному мышлению. Но при этом Гуссерль придает большое значение телесному и поведенческому сходству (поведение играет роль верифицирующей презентации): «Узнаваемые телесные движения (рук, ног, глаз) постоянно подтверждают стиль собственного чувственного процесса. Впоследствии над этим надстраиваются образования высшей психической сферы, проявляющиеся в телесных реакциях веселья, гнева или печали» [Там же, с. 192].

Итак, на философской сцене – Марк-зверь Третий из очень интересного эксперимента, который ставят авторы книги «Глаз разума» Д. Деннет и Д. Хофштадтер. Сцена, приводимая Деннетом и Хофштадтером, принадлежит перу писателя-фантаста Т. Миданера.

Марк-зверь Третий, первое знакомство: «Хант… извлек… нечто, напоминающее крупного алюминиевого жука с маленькими цветными лампочками индикаторов и несколькими выступами на гладкой металлической поверхности» [Хофштадтер, Деннет, 2003, с. 99]. Когда он подключается к электрической розетке, он мигает зелеными лампочками и издает мурлыкающий звук. «Убейте его», – говорит герой-ученый героине. Уже на этом уровне восприятия языкового высказывания возникает реакция: убить можно только что-то живое, испытывающее боль и боящееся смерти. «Убить» означает причинить этому живому боль и смерть. «Почему я должна его убивать… ломать… эту машину?» – сопротивляется героиня некорректности высказывания ученого. Но ученый заставляет героиню «убить» Марка-зверя. При этом он употребляет по отношению к механическому предмету слова, которые мы употребляем только по отношению к живым существам: «это животное беззащитно и не может вам повредить», «положите его на спинку, так оно будет совершенно беспомощно» и т.д. Сам автор текста также добавляет антропоморфных выражений: когда удар молотка сломал колесико, «машина остановилась, страдальчески мигая огоньками» [Там же, с. 100]. Под конец машина издает плач младенца. Течет смазочная жидкость, красная, как кровь. Героиня отказывается добивать машину, и это делает ученый, не обращая внимания на ее мольбы починить Марка-зверя.

Можно подумать, что перед нами эксперимент, подобный мысленному, – литературный, который ставится писателем-фантастом Миданером над героиней. На самом деле это эмпирический эксперимент, который ставят Деннет и Хофштадтер над читателем, позволяя ему познакомиться с литературным отрывком – сценой «убийства» Марка-зверя Третьего – и испытать искреннее сопереживание железной игрушке и героине, которую вынуждают ее «убивать».

А затем авторы книги ловят читателя на этом сопереживании: получается, чтобы вызвать искреннее соболезнование к предмету, достаточно описать этот предмет, как «мурлыкающий», «издающий звук, похожий на крик страха», могущий «отпрянуть», издающий детский плач. Последний штрих – вытекающая «кровь». Хотя на самом деле это смазочная жидкость, читателю становится всерьез не по себе. «Не всегда легко понять, кто или что имеет чувства», – говорит герой Миданера. Он утверждает, что когда в первый раз «убил» такую машину, то «узнал кое-что о значении жизни и смерти» [Хофштадтер, Деннет, 2003, с. 99].

Что же узнает из эксперимента о себе читатель? Хотя по выражению Деннета и Хофштадтера «все это шито белыми нитками», «мы чувствуем, что нами манипулируют, и все же, несмотря на раздражение, не можем превозмочь инстинктивного чувства жалости» [Там же, с. 101]. Отметим, читатель узнает в этом эксперименте только свою реакцию на представленный текст, автор которого при описании Марка-зверя намеренно для создания художественного эффекта пользуется антропоморфными выражениями. Деннет и Хофштадтер замечают, что Марк-зверь, только и умеющий, что мурлыкать, плакать и избегать молотка, имитируя реакцию страха, «прошел тест Тьюринга». Но прошел он его в художественном произведении.

Что же было бы, столкнись мы с устройством, похожим на Марка-зверя, в ситуации реального эксперимента? Тронуло бы нас мурлыканье? Казалось бы нам, что устройство своими лампочками на нас смотрит? Вполне возможно, что казалось бы. Но это только предположение. Хотя звук детского плача вызывает у человека реакцию на бессознательном уровне. Да и от вида красной текущей жидкости некоторых замутило бы. Так что совсем не исключено, что прошел бы он тест Тьюринга и в реальной ситуации.

Обратимся к архаическому сознанию и древнему восприятию мира человеком. Из источников известно, что архаическое сознание не отделяет себя от окружающей природы. Природные явления кажутся архаическому человеку такими же, как он сам, т.е. имеющими волю, эмоции, чувства, намерения. Природные силы сердятся или умиляются, карают или милуют. Для такого человека тест Тьюринга прошел бы весь мир – весь мир как целое и каждая его часть (каждая гора, каждый камень, каждая река) в отдельности. Ибо целое подобно части, а часть – целому. Мир – это просто огромный человек, а человек – мир, только локализованный в конкретном теле.

Подумаем, что бы человек с архаическим сознанием мог сказать о нашей проблеме другого сознания и о тесте Тьюринга? Ведь он не просто знает, что все вокруг него сознательно, он чувствует сознания гор и рек так же, как свое собственное, поскольку находится в состоянии слитности со всем. И зачем ему доказывать, что у молнии есть намерение покарать? Он это просто знает.

Возможно, понятие аналогизирующей апперцепции как понятие феноменологии не распространимо на определенную часть жизненного мира. Оно объясняет конституирование Других сознаний трансцендентальным сознанием, но не сознаниями эмпирических субъектов.

В эмпирическом мире сознание может приписываться человеком предметам, совершенно на него не похожим. И бывает труднее удержаться от акта

такого приписывания, чем решиться на него. Смысл и глубинные механизмы этого явления – тема отдельной работы. Но приписывание окружающему воли и сознательности, наделение всего смыслами, помогает человеку не только определенным образом концептуализировать мир, но и находить способы действовать в нем. А значит, с этой точки зрения, представляет собой определенную изначальную практическую установку.

Представляется возможным задаться вопросом: как работает аналогизирующая апперцепция, если сами себе мы даны хотя и как психофизическое единство («первое творение»), но при этом превалирует взгляд изнутри. Мы не можем наблюдать со стороны многих своих телесных проявлений, например, выражения лица, движения глаз. Мы хотя и представляем себе во внутреннем восприятии и видим движения своего тела, но все равно не можем знать, как выглядят наши движения в восприятии Другого, в восприятии «оттуда». Одновременно Другого мы наблюдаем именно во внешних проявлениях. Таким образом, это нельзя записать как уравнение такого характера:

$$X^{I} \text{ (внутреннее)} \leftrightarrow Y^{I} \text{ (внешнее) Я}$$
$$? \text{ (внутреннее)} \leftrightarrow Y^{II} \text{ (внешнее) ДРУГОЙ}$$

Скорее, это выглядит так:

$$X^{I} \text{ (внутреннее)} \leftrightarrow ? \text{ (внешнее) Я}$$
$$? \text{ (внутреннее)} \leftrightarrow Y^{II} \text{ (внешнее) ДРУГОЙ}$$

Но если мы, вспомнив о существовании изначальной неразделенности себя и Другого в архаическом сознании, предположим некую априорность полагания сознания у внешних по отношению к нам сущностей, тогда эта формула может выглядеть и так:

$$X^{I} \text{ (внутреннее)} \leftrightarrow ? \text{ (внешнее) Я}$$
$$X^{II} \text{ (априорное полагание сознания, эмпатия)} \leftrightarrow Y^{II} \text{ (внешнее) ДРУГОЙ}$$

Из последней записи следует нечто зеркально симметричное аналогизирующей апперцепции: из факта изначального признания нами сознательности действий и поведения Другого, наблюдая за его внешним поведением, мы отчасти можем судить, как выглядят для Другого наше собственное поведение и телесные проявления.

Итак, наличие сознания у Другого мы одновременно полагаем априорно и воспринимаем в аналогизирующей апперцепции (вспомним, что это вид ассоциативного мышления, проявляющегося спонтанно). Но о характере собственных телесных проявлений мы во многом узнаем в силу такой же ассоциации себя с Другими, полагая их сознательными и наблюдая за внешними проявлениями их взаимоотношения с собственной телесностью. Представляется продуктивным видеть такие взаимосвязи не как логическую петлю в рассуждениях, но как описание определенного вида обратной связи, действующей при восприятии нами себя и Других в качестве видов психофизического единства.

4. Эпистемологические разрывы

До сих пор мы описывали восприятие сознания Другого, подразумевающее некоторую возможность заключения от внешнего к внутреннему и от внутреннего к внешнему. Но такие заключения кажутся нам тем бесспорнее, чем обычнее и обыденнее ситуация, и в конце концов оказываются практическим элементом. Необходимо отметить на уровне теории наличие эпистемологических разрывов. Эпистемологические разрывы обусловливаются качественными различиями феноменологических проявлений и концептуализации сфер реальности, которые нужно соотнести. Различия концептуализации означают: то, что мы пытаемся соотнести, теоретически описывается в несоотносимых терминах и системах знаний. Итак, эпистемологический разрыв неизбежно присутствует между внешним (в том числе вербальным) поведением Другого и его интерпретацией моим сознанием.

Другой эпистемологический разрыв тектонически проходит между сознанием (моим или Другого) и его внешними проявлениями (в том числе вербальными). Так, следует отметить, что в европейской традиции целостные состояния сознания обычно не носят названий. Названия имеют эмоции, чувства (и то под вопросом остается, все ли они и каждый ли их оттенок имеют название). В результате сознающему часто кажется: то в себе, что он может определить словом, и является единственным содержанием его сознания («мое состояние сознания в данный момент – печаль»). Если же человек чувствует, что в таком определении что-то не так или эмоция не называется однозначно, ему приходится переходить на язык метафор, мало понятных слушающему (приватный язык).

На самом деле в субъективном опыте человека, особенно если он часто рефлектирует над собственным субъективным опытом, действительно много невыразимого вербально вследствие недостаточности языка и относительности любой метафоры. Вследствие этого феноменологические проявления для человека его собственного сознания и таковые других людей действительно могут быть концептуализированы только в несоотносимых терминах. Во внутренней сфере сознания многое остается неконцептуализированным, неназванным, невыразимым, а потому может приниматься внешним наблюдателем за бессознательное, будучи ясно осознаваемым для самого человека. Концептуализация проявленности Других более полна (в смысле меньшего количества неназванного в ней), будучи при этом и более бедной.

5. Итоги

Итак, мы рассмотрели проблему сознания (в том числе в философии ИИ) с разных сторон. Исследовали возможные подходы к ней. Но вывод, который придется сделать, возможно, не порадует тех, кто считает, что сознание и квалиа поддаются изучению, что вообще существование сознания у меня или Другого можно доказать теоретическими или практическими методами.

Проблема другого сознания, как и проблема сознания вообще, представляет собой кантианскую антиномию. Поскольку она не сводится фактически

ни к одной из кантовских антиномий, приходится говорить, что это не кантовская, а кантианская антиномия. В ее основе – специфика данности сознания в опыте, порождающая проблемы с его теоретической концептуализацией: сознание дано нам в опыте так и только так.

Антиномичность понятия сознания приводит к тому, что проблема сознания не имеет научного решения (и столько философских решений, сколько философствующих о ней субъектов). В то же время она действительно не решаема на собственно философском (окончательного и единственного философского решения не имеет) и на методологическом уровне, пока мы рассматриваем ее как проблему для применения чистого разума. Но мы именно так ее и рассматриваем, когда надеемся дать определенные и безусловно верные ответы на вопросы, как можно доказать существование сознания – у себя и у других, в каких терминах его надо описывать и какими методами возможно изучать.

Единственная возможная причина того, что человеческая мысль столько лет движется по кругу, но не находит решения этой проблемы, – кантианская антиномичность последней. Ни сознание Другого, ни единство моего собственного Я не даны нам в опыте [Кант, 2003, с. 249–253]. Нам кажется, что это не так: мое сознание мне дано, и поэтому метод интроспекции должен иметь неоспоримую легитимность среди методов философского и психологического исследования. Сознание Другого дано нам изначально в первичной неразделимости мира и схватывается при помощи эмпатии. Тем не менее, рассуждая подобным образом, мы на самом деле покидаем сферу чистого разума и переходим в сферу приложения разума практического.

Понять и обосновать, что проблема не имеет решения на определенном уровне и с помощью определенных методов, – это тоже решение данной проблемы. Проблема сознания и Другого сознания не решаема на уровне чистого разума при помощи философских методов. Никакое из предлагаемых ее решений не только нельзя доказать рационально как теорему или при помощи мысленных экспериментов (мы уже показали, что здесь у каждого будет получаться свой результат), но нельзя указать и на условия опыта, в которых что-либо, связанное с проблемой сознания или других сознаний будет подтверждено или опровергнуто. Эмпирические эксперименты бесполезны по той же причине: если что-то не дано в опыте, опытным путем мы это не исследуем. Окажется, что результаты любого возможного эксперимента можно будет истолковать в свете противоположных теорий или гипотез, которые будут «недоопределены» фактами. Итак, приходится сделать вывод: сознание нельзя не только изучить, но даже изучать (в нашем понимании научного и философского изучения с позиций нахождения твердого основания).

Сказанное, конечно, не означает, что на практике с сознанием нельзя иметь дела. Всем нам приходится действовать практически, исходя из предположения о существовании нашего сознания и существовании у Других сознаний, аналогичных нашему. Многие науки не могли бы существовать без базовых предположений об этом. Просто, исходя из этих положений практически, нужно понимать, что теоретически они недоказуемы.

Список литературы

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

Гуссерль, 2003 – *Гуссерль Э.* Логические исследования. Картезианские размышления. Кризис европейских наук и трансцендентальная феноменология. Кризис европейского человечества и философии. Философия как строгая наука. Минск: Харвест; М.: АСТ, 2000. 752 с.

Кант, 2003 – *Кант И.* Критика чистого разума. Симферополь: Реноме, 2003. 464 с.

Майнцер, 2016 – *Майнцер К.* Исследуя сложность: от искусственной жизни и искусственного интеллекта к киберфизическим системам // Инновационная сложность / Отв. ред. Е.Н. Князева. СПб.: Алетейя, 2016. С. 469–508.

Мах, 2021 – *Мах Э.* Анализ ощущений и отношение физического к психическому / Пер. с нем., вступ. ст. А.А. Богданова; предисл. А.Ф. Зотова. М.: ЛЕНАНД, 2021. 288 с.

Смирнова, 2014 – *Смирнова Н.М.* Трансцендентальная интерсубъективность и проблема «чужих сознаний» // Интерсубъективность в науке и философии / Под ред. Н.М. Смирновой. М.: Канон+, РООИ «Реабилитация», 2014. С. 183–272.

Тьюринг, 2003 – *Тьюринг А.М.* Вычислительные машины и разум // *Хофштадтер Д., Деннет Д.* Глаз разума. Самара: Бахрах-М, 2003. С. 47–59.

Хофштадтер, Деннет, 2003 – *Хофштадер Д., Деннет Д.* Глаз разума. Самара: Бахрах-М, 2003. 432 с.

Consciousness as an antinomy (antinomy of the concept of consciousness and the philosophy of artificial intelligence)

Julia S. Morkina

Institute of Philosophy, Russian Academy of Sciences. 12/1 Goncharnaya Str., 109240, Moscow, Russian Federation; e-mail: morkina21@mail.ru

The paper shows that the reasons for the contradictions existing in the philosophy of artificial intelligence (AI) are rooted in the lack of general philosophical foundations reflection. Among such foundations are the concepts of consciousness and “other consciousness”. Various approaches to these problems have been analyzed. A logical and methodological reconstruction of the of different approaches to the problem controversy of AI and consciousness is investigated. The critique of thought experiments used in the philosophy of consciousness and AI is given. A significant methodological difference between such experiments and Galileo’s experiment is demonstrated. Galileo’s experiment is a classic case of a successful thought experiment that gave an objective result. In the framework of conceptualizing and correlating the concepts of consciousness, “I” and the Other epistemological gaps inevitably emerge caused by qualitative differences in phenomenological manifestations and ways of conceptualizing the spheres of reality that need to be correlated with each other. The paper shows that an epistemological gap is inevitably present between the external (including verbal) behavior of the Other and its interpretation by the observer as manifestations of certain internal contents of the consciousness of the Other. Moreover, phenomenological manifestations for a person of his own consciousness and manifestations of the consciousness of the Other (including a verbal description of contents of consciousness) are not theoretically correlated. The antinomy of the concept of consciousness is shown as impossibility

of concluding its presence or absence in the system (“I”, Other, AI) when using purely philosophical (theoretical) methods of reasoning. The concept of consciousness as well as ideas about the consciousness of the Other, however, are necessary for practical activity and communication, that’s why a person uses practical ways of conceptualizing consciousness including intuitive perception, empathy, attribution, etc.

Keywords: consciousness, artificial intelligence, philosophical foundations, meaning-setting, analogizing apperception, thought experiment, Turing test, antinomy

References

Alekseev, A.Yu. *Kompleksnyi test T’yuringa: filosofsko-metodologicheskie i sotsiokul’turnye aspekty* [Complex Turing Test: Philosophical, Methodological and Socio-Cultural Aspects]. Moscow: IInteLL Publ., 2013. 304 pp. (In Russian)

Hofstadter, D., Dennett. D. *Glaz razuma* [The Mind’s I]. Samara: Bakhrakh-M Publ., 2003. 432 pp. (In Russian)

Husserl, E. *Logicheskie issledovaniya. Kartezianskie razmyshleniya. Krizis evropeiskikh nauk i transtsendental’naya fenomenologiya. Krizis evropeiskogo chelovechestva i filosofii. Filosofiya kak strogaya nauka* [Logical Investigations. Cartesian Meditation. The Crisis of European Sciences and Transcendental Phenomenology. The Crisis of European Humanity and Philosophy. Philosophy as a Strict Science]. Minsk: Harvest Publ.; Moscow: AST Publ., 2000. 752 pp. (In Russian)

Kant, I. *Kritika chistogo razuma* [Critique of Pure Reason]. Simferopol: Renome Publ., 2003. 464 pp. (In Russian)

Mach, E. *Analiz oshchushchenii i otnoshenie fizicheskogo k psikhicheskomu* [The Analysis of Sensations and the Relation of the Physical to the Psychical]. Moscow: LENAND Publ., 2021. 288 pp. (In Russian)

Mainzer, K. “Issleduya slozhnost’: ot iskusstvennoi zhizni i iskusstvennogo intellekta k kiberfizicheskim sistemam” [Exploring Complexity: from Artificial Life and Artificial Intelligence to Cyberphysical Systems], *Innovatsionnaya slozhnost’* [Innovative Complexity], ed. by E.N. Knyazeva. Saint Petersburg: Aleteiya Publ., 2016, pp. 469–508. (In Russian)

Smirnova, N.M. “Transtsendental’naya intersub’ektivnost’ i problema ‘chuzhikh soznanii’” [Transcendental Intersubjectivity and the Problem of “Alien Consciousnesses”], *Intersub’ektivnost’ v nauke i filosofii* [Intersubjectivity in Science and Philosophy], ed. by N.M. Smirnova. Moscow: “Kanon+” ROOI “Reabilitatsiya” Publ., 2014, pp. 183–272. (In Russian)

Turing, A.M. “Vychislitel’nye mashiny i razum” [Computing Machinery and Intelligence], in: D. Hofstadter, D. Dennett, *Glaz razuma* [The Mind’s I]. Samara: Bakhrakh-M Publ., 2003, pp. 47–59. (In Russian)