СУБЪЕКТИВНЫЙ ОПЫТ, ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ПРОБЛЕМА МОДЕЛИРОВАНИЯ СМЫСЛОВ*

Д.Э. ГАСПАРЯН

Аннотация

В настоящей статье на примере задачи компьютерного моделирования смыслов (в рамках искусственного интеллекта – ИИ) и вытекающей из нее проблемы обоснования значений, показывается, что ценности являются одним из наиболее выраженных параметров разграничения субъективного и объективного, и в этом смысле составляют проблему для возможностей компьютерной формализации. Показывается, что, по-видимому, объяснительные функции наук, изучающих сознание и всю совокупность ментальных феноменов далеки от тех возможностей, которые есть у наук, изучающих природные факты. Предполагается, что это связано не с «дефектами» метода понимания, но со спецификой самого «предмета» – в данном случае, субъективного опыта (предположительно существенно связанного с природой ценностей). Равным образом это не отменяет полноценной эвристичности знания о субъективном опыте и сознании, но лишь указывает на необходимость применения иной, более релевантной методологии. В этом смысле применение лишь объяснительной (редуцирующей) стратегии в качестве научного метода не позволяет приблизиться к работе самого субъективного опыта и сознания, а не их психических коррелятов. Для демонстрации этой идеи предлагается привлечь проблему обоснования значений и показать существенную роль субъективной аффектации как основания для остановки бесконечного регресса, возникающего при чисто лингвистической процедуре означения.

Ключевые слова: субъективный опыт, искусственный интеллект, проблема обоснования значений, смысл, ценности, first-person access, валюативность, синтаксис, семантика, бесконечный регресс оснований, тотальный тест Тюринга, робот, программа, компьютеционализм, аффектация.

Гаспарян Диана Эдиковна – кандидат философских наук, доцент Школы философии факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики».

anaid6@yandex.ru

Цитирование: *ГАСПАРЯН Д.Э.* (2017) Субъективный опыт, искусственный интеллект и проблема моделирования смыслов // Философские науки. 2017. № 4. С. 98–109.

* Статья подготовлена в результате проведения исследования (№ проекта 16-01-0032) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2016–2017 гг. и с использованием средств субсидии на государственную поддержку ведущих университетов Российской Федерации в целях повышения их конкурентоспособности среди ведущих мировых научно-образовательных центров, выделенной НИУ ВШЭ.

Введение

Одной из важных проблем, поднятых в статье Светланы Климовой, стало обращение к осмыслению последствий развития электронной культуры или шире: цифровой цивилизации, ее значения с точки зрения трансформаций личности. Здесь есть предмет для глубокой философской рефлексии, позволяющей не только акцентировать внимание на аксиологически дискутируемых вопросах о перспективах развития искусственного интеллекта (ИИ), возможных последствиях для человека и для человечества этого процесса, но и рассмотреть данный вопрос с научной точки зрения, показать роль и значение субъективного опыта в исследовании философии сознания в целом.

К вопросу моделирования смыслов

На сегодняшний день существует немало способов доказательства феноменальной реальности и субстантивированности сознания. В настоящем тексте мы рассмотрим такой способ обоснования, который отталкивается от проблемы объективации значений-смыслов (в рамках проблемы бесконечного регресса) и апеллирует к аффективной (связанной с ценностью) природе сознания.

Проблема объективации смысла – это вопрос о том, каким образом различные символы (к примеру, знаки языка) могут приобрести смысл, избегая бесконечного регресса объяснений с точки зрения других требующих осмысления символов.

Можно сказать, что сама процедура означения является довольно проблематичным процессом для моделирования. Это можно показать следующим образом. Если нам дано некоторое предложение, то значением этого предложения будет другое предложение, которое в свою очередь затребует еще одно предложение для прояснения своего собственного смысла, и так до бесконечности. Для того чтобы нечто *осмысленно*е было высказано, оно должно быть сформулировано в виде предложения, состоящего из слов, для понимания которых, мы должны сформулировать предложения, раскрывающие их значение. Тем самым мы вынуждены ввести новые слова, требующие экспликации в новых предложениях и т.д. и т.п. Если принять предложение за некое имя, то ясно, что каждое имя, обозначающее объект, само может стать объектом нового имени, обозначающего его смысл: n1 отсылает к n2; n2 отсылает к n3 и т.д. Когда мы высказываем что-то, мы при этом никогда не проговариваем непосредственный смысл того, о чем идет речь. Смысл того, о чем мы говорим, можно сделать разве что объектом следующего предложения, смысл которого также не проговаривается. В подобной ситуации имеет место своего рода бесконечное умножение того, что подразумевается.

Другой аспект этой трудности иллюстрирует известный аргумент «Китайская комната», введенный Дж. Сёрлом: мысленный экспери-

мент, в котором англоговорящий человек получает подробные инструкции на английском языке о том, как манипулировать знаками китайского языка для составления правильных ответов на вопросы, заданные на китайском [Searle 1991]. Сёрловский мысленный эксперимент показывает, что тот факт, что машина может вести с человеком вполне осмысленную беседу, вовсе не означает, что машина может «думать». Согласно аргументу Сёрла, программа будет успешно справляться с задачей внятной коммуникации, если только будет ставить в соответствие одним символам (вопросам) другие (ответы). Программа исключительно *синтаксична*, т.е. в ней принимается во внимание лишь начертание символов. *Значение их никак не раскрывается*. Человек, работающий в качестве такой программы, видит непонятные значки и ставит им в соответствие другие значки. При этом «китайские» собеседники полагают, что он ведет с ними вдумчивый разговор, на деле же он лишь механически соединяет загадочные иероглифы. Однако реальное человеческое понимание предполагает не только синтаксические правила, но и владение *значением слова,* т.е. по сути знание того как выглядит денотат. Это значит, что программа должна была бы обладать опытом «синего и зеленого цветов» – в этом случае ее «знание» терминов «синий» и «зеленый» было бы семантическим и подобным человеческому. Но поскольку на это машинный интеллект не способен, он не обладает реальным пониманием. Таким образом, даже если компьютер проходит тест Тьюринга, т.е. демонстрирует вербальное поведение, неотличимое от человеческого, из этого еще не следует, что он обладает подлинным интеллектом. Компьютерные программы работают исключительно с синтаксическими операциями и потому носят сугубо формальный характер.

Несмотря на эту критику, сторонники полноценной эмуляции ИИ утверждают, что следует несколько преобразовать условия эксперимента для того чтобы доказать возможность абсолютного уподобления машинного интеллекта человеческому. Более современная версия этой дискуссии вращается вокруг уточненных аргументов и, соответственно, контраргументов. Мы уже поняли, что, согласно широко распространенному в теории познания направлению «компьютационализма» («computationalism»), познание есть всего лишь разновидность вычислительной работы. Но вычисление, как показали критики компьютационализма (в частности, Сёрл), является лишь формальной манипуляцией с символами согласно правилам, которые основаны на формах символов, а не их значениях. Соответственно, возникает вопрос: как эти символы подключены к вещам, с которыми они соотносятся? Это соотнесение не может происходить сугубо внешним (для сознания) образом, поскольку поиск значений слов в словаре того или иного языка, которого никто не понимает, приведет к бесконечному регрессу. Как раз в этом месте проявится

тот аспект проблемы (наиболее буквальный) наделения символов значениями, который мы упомянули выше. Чтобы его избежать, нужно будет «субъективировать» процесс означения, а именно замкнуть его на такой частный опыт, который имел бы предел регрессирования. В основе этого тезиса лежит наблюдение, что люди учат значения слов посредством причинной связи между впечатлением и объектом, которому соответствует символ, и мы понимаем слово «вода», потому что мы имели жизненный опыт с водой [Rapaport 2006].

Если теперь, как это и продолжают упорно делать сторонники компьютеционализма, попробовать смоделировать и этот аспект, то придется сказать следующее. Для того чтобы компьютер понимал значения символов, которыми он манипулирует, он должен быть оснащен сенсорной аппаратурой, например, камерой – именно так он в действительности сможет установить значения объектов, представленных символами. Чтобы заслужить ярлык «думающий», машина должна пройти расширенный (Тотальный) тест Тьюринга (робот Тьюринга), т.е. уметь соотносить символы с вещами; для этого она должна быть соединена с внешним миром. Но будут ли эти символы иметь смысл, а не просто соотноситься с объектом – на этот вопрос нельзя ответить ни с точки зрения мысленного эксперимента при участии робота Тьюринга, ни с точки зрения когнитивной науки.

Валюативный предел регресса значений: значимо то, что ценно

В большинстве теорий, рассматривающих проблему объективации значения, можно выделить три способа, посредством которых может быть обнаружен смысл какого-либо действия. Один из этих способов вводит существование такой причины действия, которая может быть обоснована принципиально с субъективной точки зрения. В данном случае речь идет о качественном характере обоснования – *валюативная*, или, как мы покажем далее, *аффективная валентность действия*, которая помогает остановить бесконечный регресс оснований.

Речь идет о двух возможных значениях каузальности. Это 1) референтное обоснование (значение содержания) и 2) обоснование причин действий для себя (осмысленность содержания/аффективность).

Традиционная референтная теория значения предполагает наличие внешнего мира – знак имеет значение тогда, когда он отсылает к некому объекту вовне системы знаков[8]. Известно, что подобное обоснование символа проблематично в любых теориях и подходах конструктивистского или корреляционистского типа, в которых «адаптивные действия в мире следует понимать в контексте активно сконструированных, а не пассивно отражающих реальность». Одним из возможных решений проблемы обоснования символов, стала теория, предложенная Харнадом, согласно которой обоснование символа происходит в ходе сенсомоторной активности «как распознавание и

выделение объектов, событий и положений дел, с точки зрения их сенсорных проекций» [Harnad 1994].

Общей чертой многих теорий обоснования, в том числе теории Харнада, является понимание осмысленности (обоснованности) действия. Обоснованным (осмысленным) является то действие, которое каузально эффективно: обработанная информация должна иметь определенные средства воздействия на внешний мир. Если система восприятия изолирована от окружающей среды и каузально замкнута, она неизбежно находится в условиях отсутствия причинной эффективности. Соответственно, для создания теории сознания, в которой смысл и причины действий являются обоснованными, необходимо постулировать существование внешнего мира. Это уточнение важно потому, что на сегодняшний день возлагаются большие надежды на оснащение программ сенсорно-моторными датчиками, обеспечивающими процедуры локального распознавания объектов «в мире». Если компьютер сможет получить сенсорный доступ к реальным объектам в мире, то он получит доступ к смыслу, а значит и к пониманию. Ряд исследователей полагает, что если научить компьютер (Робот) работе с индексикальными ситуациями, предполагающими ответы типа: «вижу перед собой камень», «навстречу мне идет Иван», то можно будет считать, что машина научилась точно так же соотносить символы с объектами (при посредничестве смысла), как это делает человек, а значит, машина научилась понимать и думать.

Итак, для того чтобы значение было обоснованным, должно быть соединение с окружающим миром. Такую связь с внешним иногда пытаются толковать в сугубо биологических терминах: внешний мир, в котором живые организмы развиваются и с которым они взаимодействуют, фундирует наличие причин, определяемых в терминах биологической полезности. При этом биологический тип обоснования зависит от ряда допущений, в частности от того, что редупликация вида – это «успех», что выживание – это «хорошо». Таким образом, здесь появляется минимальное понятие «валюативности», а значит ценностной интерпретации обоснования значения [Soares 2015]. С одной стороны, подобная биологизаторская интерпретация обоснования значений удобна в случае моделирования интеллекта, так как может срабатывать технический критерий оптимизации (оптимальности выбора, оптимальности соотношения элементов и степеней возможностей и т.д.). Однако используемый при такой трактовке вид «ценностей» (лучше назвать их целями) мы не можем оценивать в терминах собственно человеческих валюативных систем. Это связано с тем, что биологическая программа не может адекватно объяснить нормативные причины возникновения начальных действий. Так, биологический тип обоснования говорит нам лишь о том, что существует некая эволюционная история того, как мы появились, чтобы вести

себя определенным образом; но мы не можем предоставить никаких ценностных оснований для самого возникновения. Главная причина, которая здесь указывается – биологическая полезность – не является самоценной и в этом смысле не может приостановить бесконечный регресс оснований. Мы вынуждены искать причины, выходящие за пределы биологической полезности, которые могли бы сообщить ей (самой биологической полезности) некую ценность (смысл).

Обнаруживая ценности во внешнем мире, мы также вновь сталкиваемся с проблемой бесконечного регресса, на этот раз самих ценностей. Основания одних ценностей отсылают к другим, так что в итоге реально ценной оказывается только сама утилитарная (рациональная в силу своей оптимальности) мотивация для достижения того или иного блага. Однако рациональная утилитарная мотивация всегда оказывается промежуточным звеном и в этом смысле не может остановить бесконечный регресс. Отсюда можно сделать предварительное предположение, что обоснование причин действий, во избежание регресса оснований должно предполагать замкнутость. Эта замкнутость должна вводиться как принципиально непрозрачное (по меньшей мере, для внешнего наблюдения и объяснения) основание для действия, которое представляется *кому-то* значимым. Если нельзя ввести объективную (доступную для наблюдения в перспективе от третьего лица) оптимальность выбора, то ее необходимо сделать объективно непрозрачной. По сути, это возвращает нас к понятию субъективности – традиционно понимаемой в философии с точки зрения автокаузации.

Если теперь вернуться к вопросу обоснования значений, то для формализации этой проблемы, возможно, придется ввести некое *валюативное*, своего рода *аффективное*, и в этом значении *смысловое* содержание, – то, что, по сути, называется «субъективной значимостью». В данном случае можно предположить, что именно качественный характер аффективных реакций останавливает бесконечный регресс оснований (значений). Это связано с тем, что объективная оптимальность, в действительности, не может ничего объяснить. Всегда можно поставить вопрос, почему мы должны ценить определенные результаты, например, наше собственное выживание. В конечном счете, мы должны быть в состоянии остановить серию вопросов, в которых выясняется, почему мы должны что-то делать. Но очевидно, что причина, связанная со следствием логически, никогда не будет последней в ряду возможных других причин. Соответственно, можно утверждать, что именно аффективная (а не рационально оптимальная) валентность ожидаемого результата действий обеспечивает остановку регресса «почему-вопросов». Это связано также с тем, что мы не можем не верить, что определенные аффективные состояния предпочтительнее других. Аффективная валентность также является необходимой для

построения концептуальной основы таких понятий, как «хороший» и «плохой», «свой» и «чужой», «мирный» и «враждебный». Для того чтобы эти понятия имели значение в рамках контекстов, в которых они применяются, требуется опыт аффективного восприятия. Под определение аффективного восприятия может подойти и понятие «желания» и понятие «эмоции», т.е. в общем и целом то, что в традиционной философии и отчасти психологии связывается с областью *воли*. В свою очередь обоснование действия как процесс определения значения фактически заканчивается в момент, когда упирается в некое ценностное, эмоционально схватываемое понимание. Эмоциональное или аффективное состояние знаменует предел объяснения в силу своей непосредственной самопонятности.

Таким образом, ценность, которая имеет аффективную валентность как часть своего собственного содержания, имеет особый смысловой статус. Он возникает тогда, когда есть способность переживать аффективные реакции. Если не принять их во внимание, то обосновать то или иное значение, избегая регресса, невозможно. Ценностный тип обоснования отличается от типа обоснования, который просто определяет символы для референтного обозначения чего-то внешнего по отношению к системе. Но ценностный тип осмысленности предполагает субъекта, а именно точку зрения от первого лица. Как правило, мы непосредственно знаем, *что* мы ценим и в этом смысле желаем. В свою очередь, наше желание (знание не равно желанию) можно использовать не только как предпосылку в рассуждениях, без каких-либо дополнительных обоснований, но и как предел регрессирования в случае каузального объяснения. Аффективная валентность имеет способность останавливать регресс. Но аффективность как самодостаточность может переживаться только как приватный, непосредственный и прозрачный (для перспективы от первого лица) опыт.

Под прозрачностью в первую очередь понимается опыт обладания «квалиа» – особой, не сводимой ни к чему иному сущности непосредственного переживания некоторого состояния. Как мы уже говорили выше, можно попытаться обосновать, что именно сфера ценностей в первую очередь подходит под определение в терминах «квалиа», которое принципиально сохраняет свою перспективу от первого лица, и, будучи внутренним измерением, не может быть объяснено внешним каузальным способом. Для ценностей существенным оказывается то, *кто* их ценит (личный опыт), так как при его отсутствии они попросту перестают существовать. Это связано с тем, что квалитативному состоянию присущ особый субъективный эпистемический доступ. Поскольку при кажущейся интуитивной ясности понятия «субъективность» оно все-таки довольно неопределенно, прояснить его можно с помощью таких характеристик, как приватность, привилегирован-

ный доступ или перспектива от первого лица. Группа этих концептов призвана пояснить идею того, что о своих сознательных актах человек знает прямо и непосредственно. Например, я не могу сомневаться в том, что испытываю боль или вижу желтое пятно на стене. Любое сомнение в этом случае может означать новое состояние сознания, но не может отменить старое, ибо его истинность дана наглядно, только как сам факт осознания. Положение непосредственности можно пояснить еще проще – свои феноменальные состояния мы не дедуцируем и не индуцируем, мы вообще их ниоткуда не выводим, и потому знание о них не есть результат некой сложной интеллектуальной работы, не есть продукт вывода. Избыточным выглядело бы умозаключение вида: «Всякий укол сопровождается ощущением боли; я укололся, следовательно, я испытываю боль». Кроме того, непосредственность моих переживаний не может стать объектом наблюдения из перспективы третьего лица – только я могу наблюдать свою боль, ее нельзя наблюдать со стороны так, как можно наблюдать мое тело, которое эту боль испытывает.

Соответственно, только сам субъект знает, как собирается реагировать на определенные впечатления. Строго говоря, для этого ему не нужно выполнять процедуру референрования от одного значения к другому. Остановка процедуры означения происходит в тот момент, когда некий смысл входит в состояние непосредственного переживания. Одной из простейших форм такого вхождения является оценка события, имеющего коррелят в виде некой аффектации – эмоции, желания, переживания и пр. Качественный характер аффективных реакций на переживания может считаться более фундаментальным, чем производство любых других референций, для которых требуется в пределе бесконечная референция.

Искусственный Интеллект и моделирование смыслов

Рассмотренную вкратце тему можно толковать в соответствии с высказанной выше идеей о невозможности внешнего объяснения внутреннего измерения субъекта. Именно в связи с ценностной составляющей процедуры означения речь идет о том, чтобы не объяснять, но понимать некоторые события, в частности, определенные действия. Можно сказать, что именно эмоции означают конец обоснования.

Но эту тему можно также рассматривать в связи с возможностями моделирования смыслов. Если компьютерное моделирование значений упирается в сформулированную выше проблему регресса, то прояснится ли ситуация при оснащении компьютера сенсорно-моторными компетенциями? Ответить на этот вопрос можно попробовать с помощью следующего мысленного эксперимента. Допустим, в некий глобальный компьютер можно «зашить» все возможные ходы и комбинации – несомненно, это возможно, и такие машины суще-

ствуют. Посмотрим теперь, что такая машина может делать со своим сверхзнанием. В первую очередь она может наиболее оптимально совершать выбор игровых действий, и антропоморфному игроку обыграть ее практически невозможно. Зададимся теперь вопросом, что служит запускающим механизмом для начала игры. Фактически человек не только закладывает в нее программу оптимизации ходов, но и дает команду: «Играй белыми против черных», или наоборот. Программист закладывает в программу не только все оптимальные ходы, но и *цель* – следует выигрывать, а не проигрывать. Но два этих мотива: 1) играть белыми против черных и 2) правильно – выигрывать, а не проигрывать, привносит человек как свои частные человеческие (локальные) желания и цели. Может ли компьютер сам «хотеть» выиграть, или может ли он сам разобраться против кого ему играть – против белых или против черных? Ведь для машинного интеллекта выигрыш белых и выигрыш черных абсолютно *эквивалентны.* Он просто знает, что надо делать для выигрыша одних и проигрыша других, но он не знает *за кого он хочет* играть. Доступна ли ему аффектация в виде желания начала игры, выигрыша, выбора в пользу цвета фигур и, главное, момента окончания игры? Реально подобная программа будет работать, только если на нее будет влиять локальный заказчик (считающий, например, что выигрывать хорошо, а не плохо). Но в этом смысле она выступает антропоморфным заказчиком, так как подчиняется его валюативным командам, оказывающим на нее метавоздействие. Можно предположить, что если на подобную программу не сможет влиять никакое антропоморфное целеполагание, то она не сможет работать. Если для программы все позиции равны, она не сможет начать партию, ибо для этого она должна сделать выбор: выиграть партию белыми, а не проиграть ее черными. Но для этого у нее должны быть основания для выбора, которые не являются частью правил игры (это не позиции, не комбинации и не стратегии). Однако в первую очередь программа оснащена именно этими компетенциями. Основания для выбора есть у такого игрока, который руководствуется аффективными основаниями (будь то программист или пользователь программы).

Если перейти на язык программирования, то можно сказать, что ценностный выбор будет «выбиваться» из алгоритма сугубо синтаксических связей. Речь идет о своего рода произвольном смысле, который невозможно прописать в номенклатурной системе, но который легко производится (понимается) творческим актом сознания. Тогда можно выдвинуть гипотезу, что машины не способны отличать ценностные выборы от «ценностно нейтральных» – прописанных алгоритмом в соответствии с оптимизацией локальных и глобальных результатов работы программы. Как бы ни были богаты имитационные ресурсы программ, в задачах, где нужно осуществить ценностный выбор

(принять решение – выиграть или проиграть) приходится допускать своего рода ломку алгоритмической парадигмы. Фактически здесь происходит остановка символьного ряда, поэтому выполнение подобных нестандартных задач осложняется. Даже если предположить, что в компьютер будут заложены шаги, имитирующие аффекты, нарушение правил и прочие девиантные способы реагирования, само отличие ценностной ситуации от стандартной (прописанной в алгоритме) представляет собой трудность.

Соединяя эту иллюстрацию с идеей регресса означений, можно также сказать, что остановка в процедуре обоснования, которая, собственно, и считается моментом понимания, как правило, является произвольной, в буквальном смысле моментом волевого произвола. Логическая формализация подобной остановки, как мы уже говорили, не имеет имманентных причин, и выливается в проблему бесконечного регресса. Вместе с тем, если допустить основание, отличающееся автореференциальной природой, т.е. перформативное и прозрачное, то бесконечный ряд означений удастся остановить. Таким основанием оказывается ценностный параметр, имеющий корреляцию в области субъективных аффектаций. В свою очередь компьютерная формализация данного измерения не может быть буквальной – ее можно предполагать у программы, совершающей определенные действия, но ее нельзя внешним образом алгоритмизировать. В противном случае это означало бы нарушение свойства приватности.

Заключение

Подводя краткий итог сказанному, можно попытаться выделить определенные практические результаты исследований, посвященных философским вопросам, связанным с созданием искусственного интеллекта (ИИ) и прочих последствий развития цифровой цивилизации. Философская рефлексия может помочь яснее представлять границы научной эффективности подобных разработок. В первую очередь такая рефлексия должна оказаться полезной в понимании того, какие из текущих научных или междисциплинарных проблем могут быть успешно решены в обозримом или более отдаленном будущем, а какие проблемы представляют принципиально неразрешимые трудности. Оценка подобных перспектив отчетливо связана со сложившейся дилеммой, под знаком которой, вероятно, будет и дальше развиваться современная человеческая мысль. Речь идет о противостоянии между натуралистскими и трансценденталистскими программами. Фантастически блестящие результаты развития науки в наши дни не могут не поражать воображение: успехи нейроинженерии и робототехники таковы, что вера ученых в то, что вскоре нам удастся творить себе подобных, и этот прорыв уже не за горами, кажется очень воодушевляющей. На самом деле то, чего мы можем ждать

и на что надеяться в своих предвосхищениях будущего человечества, можно попытаться понять уже сегодня. Для этого нужно разобраться в правоте одной из двух противостоящих друг другу установок – натурализма или трансцендентализма. В зависимости от того, какая из них убедит нас больше, соответствующее представление о том, как далеко сможет простираться интеллектуальная мощь человечества, мы и получим. Но для решения этого вопроса, в свою очередь, предстоит понять, существуют ли какие-то принципиальные трудности на пути эпистемологии, претендующей на всеохватное понимание.

ИСТОЧНИКИ

Harnad 1994 – *Harnad S.* The symbol grounding problem // Physica. 1994. P. 100–140.

Rapaport 2006 – *Rapaport W.* How Helen Keller Used Syntactic Semantics to Escape from a Chinese Room // Minds and Machines. 2006. 16 (4). P. 381–436.

Searle 1991 – *Searle J.* Minds, Brains, and Programs // Rosental D. (ed.) The Nature of Mind. – N. Y., 1991. P. 20–24.

[4] Soares 2015 – *Soares N.* The value learning problem // Technical Report. 2015. Vol. 4, pp. 65-74.

SUBJECTIVE EXPERIENCE, ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF SENSE MODELING*

D.E. GASPARYAN

Summary

This article explores the symbol of grounding problem in the framework of artificial intelligence, the AI. The author explicitly demonstrates that values are one of the most pronounced parameters of differentiation the subjective perspective from the objective. The research applies a problem of computer formalization of the meaning. As a result the article shows that explanatory opportunities of science are far from the opportunities that supposed to be used for study the mental facts. It is assumed that this is not due to flawless of the method, but because of the peculiarity of the «subject». In this case, subjective experience is presumably significantly associated with nature of values. In this sense, the use of the reducing strategies as the scientific method does not allow to approach the first-person experience and consciousness, but only their mental correlates. To demonstrate this idea the author involve the notion of values to show a significant role of subjective affectation, as grounds for stopping the infinite regress of meanings that occurs when a purely linguistic procedure is applied.

* The article was wrote within the framework of the Academic Fund Program at the National Research University *Higher School of Economics* (HSE) in 2016–2017 (grant № 16-01-0032) and supported within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

Keywords: subjective experience, artificial intelligence, the symbol grounding problem, meaning, values, first-person access, syntax, semantics, infinite regress of reasons, the total Turing test, robot, program, computationalism, affectation.

Gasparyan Diana – Ph.D. in Philosophy, Associate Professor at the School of Philosophy, Faculty of Humanities, National Research University *Higher School of Economics*, Moscow.
anaid6@yandex.ru

Citation: *GASPARYAN D.E.* (2017) Subjective Experience, Artificial Intelligence and the Problem of Sense Modeling. In: *Philosophical Sciences*. 2017. Vol. 4, pp. 98-109.

REFERENCES

Harnad S. (1994) 'Grounding symbols in the analog world with neural nets'. In: *Physica*. 1994, pp. 100-140.

Rapaport W. (2006) 'How Helen Keller Used Syntactic Semantics to Escape from a Chinese Room'. In: *Minds and Machines*. 2006. 16 (4), pp. 381-436.

Searle J. (1991) 'Minds, Brains, and Programs'. In: Rosental D. (ed.) *The Nature of Mind*. New York, 1991, pp. 20-24.

Soares N. (2015) 'The value learning problem'. In: *Technical Report*. 2015. 4. Machine Intelligence Research Institute, pp. 65-74.