ФИЛОСОФИЯ: ТРАДИЦИИ И СОВРЕМЕННОСТЬ

DOI: 10.17212/2075-0862-2023-15.2.1-72-96 УДК 16:159.9.01

ПРОБЛЕМА ВЫЧИСЛИТЕЛЬНОЙ ОБЪЕКТИВАЦИИ РАЦИОНАЛЬНОСТИ В ИСКУССТВЕННЫХ ИНТЕЛЛЕКТУАЛЬНЫХ АГЕНТАХ

Желнин Антон Игоревич,
кандидат философских наук,
доцент кафедры философии
Пермского государственного национального
исследовательского университета,
Россия, 614068, г. Пермь, ул. Букирева, 15
ORCID: 0000-0002-6368-1363
antonzhelnin@gmail.com

Аннотация

Предмет статьи – объективация рациональности в искусственных интеллектуальных агентах (ИИА). В этом контексте исследованы два комплементарных тренда. Первый, восходящий (bottom-up), связан с попытками артификации рационального рассуждения и действия в ИИА. Второй, нисходящий (top-down), с попытками интерпретации человеческого мышления и поведения в машинных терминах. Первый ограничен отсутствием семантической нагруженности, личностной окраски и полноценной психосоматической воплощенности вычислительных процессов в ИИА. Гипертрофия логических нормативных компонентов в них приводит к механистической ригидности. Второй порожден экспансией цифровых технологий в реальную жизнь, их активным сращиванием, порождающим идеологию компьютационализма, в соответствии с которым люди – это вычисляющие агенты. Вместе с тем конвергенция «интеллектуализации» ИИА и «машинизации» человека – превратная видимость, так как существуют фундаментальные онтологические и эпистемологические пределы. Основной онтологический предел: человеческая рациональность фундирована социокультурными и биоадаптивными пластами бытия человеческого субъекта и поэтому принципиально не может быть воспроизведена в онтологически простых физико-технических ИИА. Основной эпистемологический предел: человеческое мышление не может быть формализовано и вычислительно объективировано, так как первичным в нем является содержательный смысловой стержень, который несет на себе печать субъективности, а также тесно сцеплен с пластами неявного личностного знания и прочими

составляющими сознания. Среди последних выделяется ряд суб- и внерациональных феноменов (эмоции, ценности, здравый смысл, мораль), которые максимально дистанцированы от алгоритмического усреднения, но без интеракции с которыми рациональность остается сущностно редуцированной. Делается вывод, что вычислительная репрезентация и объективация отдельных компонентов рациональности в ИИА возможна, но сама не является рациональной, так как их искусственный отрыв от большей, человекоразмерной части рациональности инициирует раскол в ней.

Ключевые слова: рациональность, вычисление, вычислительная рациональность, гибкая рациональность, субъект, рациональный агент, интеллект, искусственный интеллект, компьютационализм, цифровой гуманизм.

Библиографическое описание для цитирования:
Желнин А.П. Проблема вычислительной объективации рациональности в искусственных интеллектуальных агентах // Идеи и идеалы. – 2023. – Т. 15, № 2, ч. 1. – С. 72–96. – DOI: 10.17212/2075-0862-2023-15.2.1-72-96.

Введение

Современная жизнь характеризуется ростом рационально истолкованной сознательности по многим векторам и во многих сферах. Однако разум в ней по большей части перестает быть самоценным автономным «сувереном», коим он был в классическую эпоху, всё больше приобретая статус средства, операционализируясь и инструментализируясь: «Отказавшись от автономии, разум стал выполнять роль инструмента. Чем больше идей претерпевают автоматизацию, становятся инструментами, тем менее кто-либо склонен видеть в них мысли, имеющие самостоятельное значение. Как только мысль или слово становится инструментом, пропадает необходимость действительно “думать”» [30, с. 27–30]. Наряду с этим рациональность начинает рассматриваться не как монолит, а как составной феномен. Так, Дж. Сёрл констатирует ряд «разрывов» в ней: «Во-первых, это разрыв в рациональном принятии решения, когда мы пытаемся принять решение о том, что намерены делать. Иными словами, это разрыв между причинами принятия решений и реальным решением, которое мы принимаем. Во-вторых, существует разрыв между решением и действием. Как причины для решения недостаточно для того, чтобы его принять, так и решение еще не является причиной для того, чтобы произвести действие» [26, с. 29]. Рациональность часто делится на рациональности цели, выбора, действия, правила [7], которые вполне могут не находиться в когеренции: так, действия могут являться рациональными, так как направлены на последовательное достижение цели, но сама цель при этом является сугубо нерациональной; либо же достижение рациональной цели рациональными средствами может тем

не менее быть рассмотрено как нерациональное, потому что расходится с неким нормативным правилом.

Как инструментализация, так и аналитическое разделение рациональности создают благоприятную почву для попыток ее формализации и дальнейшей объективации в искусственных интеллектуальных агентах (далее – ИИА). Так, они уже сейчас демонстрируют «зачаточные» формы интенциональности и аналитической обработки разнородной информации (через технологии компьютерного зрения, распознавание образов и речи, машинное обучение). Рациональность же правила вообще является удачным подспорьем в том смысле, что алгоритмы и программы дефакто есть реализация следования нормативному правилу par excellence. Всё бóльшие успехи в данной области заставляют наиболее оптимистичных теоретиков рассуждать о становлении новой, вычислительной рациональности (computational rationality) [38, 43]. Конвергенция машинного и человеческого на единых компьютационалистских основаниях обоюдоостра. Она создает видимость, что не только машины становятся интеллектуальными, но и что люди предстают как реализующие машиноподобные вычисления: «И человек, и машина вычисляют при прохождении через цепочку альтернатив. Когда мы вычисляем, мы действуем как будто/как машины» [44, p. 43, 44].

Носитель рациональности: субъект или агент?

Уже современное понимание фигуры носителя рациональности показывает основания для объективации рациональности. Всё чаще пользуются абстрактным понятием «рациональный агент»: «Агентом считается всё, что действует. Рациональным агентом называется агент, который действует таким образом, что можно было бы достичь наилучшего результата или, в условиях неопределенности, наилучшего ожидаемого результата» [25, с. 39]. Такая дефиниция ориентирована на феноменологическую оценку оптимальности и когерентности деятельности и максимально абстрагируется от качественной специфики того, кто/что действует. В.И. Шалак полагает, что все события и процессы распадаются на две группы: детерминированные безличными законами природы и инициированные агентами, причем вторые трактуются как алгоритмические, потому что направлены на некий результат и предполагают следование набору предписаний по его достижении: «Если посмотреть на происходящие в окружающем нас мире изменения, их можно разделить на два вида. К первому виду относятся изменения, описываемые законами наук. Ко второму виду относятся изменения, инициируемые активными целесообразно действующими агентами. Такие изменения могут быть названы алгоритмическими, поскольку получают объяснение в терминах следования предписаниям для достижения требуе-

мого результата» [31, с. 60]. Такая трактовка позволяет рассматривать ИИА как носителей рациональности, если их функционирование непротиворечиво, обоснованно и целеподобно. Также рациональность часто связывают с рефлексией, критичностью, открытостью к самокорректировке. Однако если понимать их абстрактно, то ими вполне могут обладать и ИИА: «Если мы имеем дело с рациональным мышлением (или вычислением), то "программа", определяющая последовательность состояний, открыта для рациональной критики» [22, с. 38]. Можно непротиворечиво представить себе ИИА, снабженные особыми программами, которые бы корректировали их же программы более низкого уровня, что могло бы трактоваться как аналог рефлексии. Так, дается следующая дефиниция вычислительной (компьютационалистской) рефлексии: «Саморефлексия в интеллектуальной технической системе (или вычислительная рефлексия) – это способность системы непрерывно мониторить и улучшать свое собственное поведение в неопределенном, динамичном и зависимом от времени окружении для ситуаций, которые могли не быть предвосхищены при дизайне данной системы» [41, p. 2].

Существует встречный тренд, когда рациональный аппарат человеческой психики рассматривается как феномен, потенциально разлагаемый на части, которые могут быть объяснены в машинных/вычислительных терминах. М. Минский рассуждает о «сообществе разума», понимаемого как кооперация простых агентов, которые сами лишены разумности: «"Обществом разума" я буду впредь именовать такую схему, в которой каждое сознание представляется состоящим из множества мелких процессов. Указанные процессы мы будем называть агентами. Каждый ментальный агент по отдельности выполняет некое простое действие, для чего не требуется ни разум, ни мышление вообще. Тем не менее, когда мы объединяем указанных агентов в сообщества, это ведет к возникновению подлинного интеллекта. Чтобы объяснить, что такое разум, мы должны показать, как он возникает из бессмысленного, неразумного материала, из своих составных частей, которые намного меньше и проще, чем любое проявление разумности» [16, с. 5, 7]. Основанием для этого выступает не только редукционистский, но и функционалистский подход, полагающий, что комплексность проистекает не из самих частей, а из их взаимодействия. Но и он строится пусть на менее явном, но упрощении: вынесение основной массы сложности вовне элементов-«агентов», в свою очередь, порождает недооценку их собственной организации, интенцию на понимание их как точечных «узлов» сети. В пределе такие агенты оказываются «черными ящиками»: от их сущности полностью абстрагируются, ограничиваясь сетевыми потоками входных и выходных данных, что затем безосновательно экстраполируется, например, на нервную систему человека: «Головной

мозг можно рассматривать как орган, который использует сенсорные входы для создания адаптивного поведения посредством моторных выходов» [48, p. 112].

Эти подходы последовательно размывают фигуру рационального субъекта. Возникают концепты техносубъектов, псевдо- и квазисубъектов, гибридных субъектов (что связано со становлением ИИА), демонстрирующих свойства, генеалогически считавшиеся атрибутами человека: «В контексте субъектных парадигм управления для обеспечения взаимодействий субъектов естественного интеллекта с активными формами искусственного интеллекта (ИИ) последние целесообразно трактовать как псевдосубъектов ИИ, которые обладают базовыми свойствами субъектов, инвариантными по отношению к типу субъекта (индивид, группа, организация и др.). К базовым инвариантным свойствам субъектов целесообразно отнести целеустремленность, рефлексивность, коммуникативность, социальность и способность к развитию» [13, с. 135]. Вместе с тем их появление еще сильнее обостряет давнее различие субъектности и субъективности. В.И. Игнатьев считает, что в случае человеческого субъекта субъективность, обозначающая личность, сознающую себя и свои индивидуальные черты и опыт, органически вплетена в него, в случае же интеллектуальной машины имеет место ситуация субъекта без субъективности: «При этом его субъектные качества сводятся к способности проявлять активность (воздействовать на окружение и себя) и делать это осмысленно (познавая). В таком случае любая аналитическая машина, отдающая приказы механическим устройствам, может быть причислена к субъектам. Но при этом ей не хватает главного – личностных характеристик, которые есть комплекс представлений субъекта о себе как об индивидуальности и уникальности, комплекс, сформированный в результате накопления собственного жизненного опыта. Эти черты формируют самосознание субъекта, т. е. уникальную внутреннюю картину мира – субъективность. В данном случае речь идет об особом субъекте – отдельном, единичном, относительно автономном объекте-носителе источника активности, обладающем некоторыми важными чертами человека как личности, прежде всего интеллектом, если техносубъект основан на ИИ. Но интеллект – это лишь один из атрибутов человека разумного, делающего его активным аналитиком, но еще не субъектом с характеристиками личности» [11, с. 134]. В этом смысле восходящее движение «снизу вверх» (bottom-up) (от ИИА к человеку) наталкивается на принципиальные лимиты. С другой стороны, повторимся, всё более усиливается обратный нисходящий тренд «сверху вниз» (top-down), описывающий человеческий субъект в машинных терминах. Полагается, что его поведение настолько вплетено в ткань алгоритмов и технологий, что неизбежно начинает нести на себе их отпечаток: «Человеческое действие в

рамках этих рационалистических построений не является внешним и свободным по отношению к системам, оно теперь их часть, реализуемая через рациональность алгоритмов. Коммуникации приобретают механистический, холодный характер, всё больше напоминая безличную рациональность машины» [1, с. 120, 121]. В пределе такой компьютационалистский подход трактует деятельность человека как особые социальные вычисления. Так, И.Ф. Михайлов полагает, что «применение вычислительной идиомы к биологическим, когнитивным и социальным сюжетам предполагает "ослабление" изначальной "сильной" версии компьютационализма» и что «"слабая" концепция вычислений должна быть родовой по отношению к конкретным разновидностям вычислительных процессов» [167, с. 27]. В итоге начинает превратно полагаться, что речь идет только об эмуляции, когда вычислительные процессы с одного субстрата (человека) последовательно переносятся, копируются на другой (ИИА), и нет принципиального запрета на придание последнему любых присущих первому характеристик, что в итоге нивелирует его онтологическую и антропологическую специфику.

Отдельным вызовом является прогресс робототехники, так как именно она ответственна за внедрение ИИ в физические механизмы, придание им воплощенного облика [21]. С данной точки зрения более-менее антропоморфная роботизация ИИА закроет их недостаточность в «соматическом» аспекте. В спектре энактивистских, аутопоэзисных, воплощенных (embodied) концепций именно наличие телесности, соответствующего сенсорно-моторного опыта и интерактивного взаимодействия с другими и со средой создает базовый уровень субъективности, на котором надстраиваются высокоуровневые ментальные образования [40]. С другой стороны, они же подчеркивают и биоонтологическую специфику такой воплощенности, ее фундированность не абстрактным телом вообще, а именно живым, органическим телом: отмечается, что «ключевой для них является убежденность в том, что для того чтобы полностью понять познание на любом уровне сложности, требуется, чтобы живые организмы рассматривались как активные воплощенные субъекты, динамически связанные и взаимодействующие с их соответствующими средами» [37]. Телесность (а точнее, телесноподобность) ИИА ввиду своей технофизичности онтологически недостаточна для формирования подлинной субъективности.

Рациональное мышление: лимиты лого- и компьютоцентризма

Коренным изъяном даже наиболее сложных ИИА оказывается гипертрофия логических нормативных элементов. Подобное функционирование представляет собой реализацию формальных пропозициональных

команд и программ: «В отсутствие глубинной связи между знаниями и намерениями логика ведет к безумию, а не к разумности. Логическая система без цели будет просто порождать бесконечное множество бессмысленных истин наподобие следующих: А влечет А. Р или не Р. А влечет А или А или А» [16, с. 299]. Справедливо отмечается, что целостное рассуждение представляет собой единство логических и металогических средств, «с помощью которых осуществляется управление логическими выводами, применяемыми в процессе рассуждения», в соответствии с чем логический вывод рассматривается только как «частный случай рассуждений, когда множество аргументов фиксировано, нетривиальные металогические средства не используются, и применяются лишь правила достоверного вывода» [29, с. 34]. Логический вывод оказывается по сути искусственной моделью рассуждения, максимально очищенной от психологического и приведенной к инвариантной строгости. Осознание этого заставляет признать нетождественность рационального и нормативно-логического: «Правильный логический вывод не исчерпывает понятия рациональности, существуют способы рациональной организации действий, в отношении которых нельзя утверждать, что в них используется логический вывод» [25, с. 39].

Однако в этом кроется главное ограничение: с самого начала вычислительной революции именно логика (в широком смысле, включая теорию алгоритмов, конечных автоматов, формальной арифметики, других разделов дискретной математики) стала основной движущей силой (так, изначальным импульсом для создания первых цифровых устройств послужила принципиальная возможность физической реализации булевых операций). Для моделирования же более сложных и приближенных к естественным рассуждений начали использовать новые разделы логики, ослабляющие или отменяющие те или иные принципы классической, – нечеткие, многозначные, модальные и др. Объективизация логики в вычислительных устройствах vice versa демонстрирует, что она не является исключительной привилегией человеческого субъекта: «Логическая форма независима и в том отношении, что может существовать в отчужденном от человеческой психики виде, т. е. в виде графических знаков, в программе ЭВМ, в конструкции технического устройства» [9, с. 63].

Всё чаще подчеркивается множественность потенциальных материальных реализаций логики: «Физическим состояниям, например “вентиль открыт/закрыт”, “заряд есть/нет”, “свет поляризован/не поляризован” и т. п., можно приписывать логические значения» [5, с. 93, 94]. Данная плюральность свидетельствует о возможности переноса логики на новые субстраты, сопряженные с определенными плюсами. Так, современные иссле-

дования демонстрируют прогресс в реализации логических операций уже не на чисто физических, а на молекулярных и живых носителях. Гипотетически реализованный на них ИИ был бы более «сильным», так как онтологически более близко воспроизводил бы структурную архитектуру и принципы работы человеческого мозга как органа, который под влиянием того же компьютационализма часто трактуется как аппаратно-технический носитель (hardware) разума [49]. Н. Бостром именует такой тип ИИ нейроморфным, отмечая, что «не исключено, что выбор такого пути, как эмуляция мозга, приведет нас к созданию нейроморфного ИИ, который будет основан на обнаруженных в процессе эмулирования принципах нейровычислительной системы» [4, с. 44]. Речь идет об эмуляции, т. е. часто полагается, что сам человеческий мозг – это сложная, реализующая особые вычислительные процедуры система. Дж. фон Нейман одним из первых провел такую аналогию, полагая, что мозг функционирует как «естественный автомат», важнейшую роль в котором играют реализуемые сетями нейронов логико- и арифметикоподобные процессы: «Под логическими командами я понимаю нервные импульсы, появляющиеся на соответствующих аксонах, а также любые другие явления, которые заставляют цифровую логическую систему (такую как нервная система) функционировать воспроизводимым, целенаправленным образом. Нервная система, рассматриваемая как автомат, обязательно должна иметь и арифметическую, и логическую части, и требования арифметики в ней столь же важны, как и требования логики. Это означает, что мы снова имеем дело с вычислительной машиной в полном смысле слова» [19, с. 165, 173]. Но современный взгляд на человеческий мозг разительно отличается от взгляда кибернетики середины XX века. Он представлен коэволюционными и аутопоэтическими подходами, где упор делается на реципрокную *ко*детерминацию мозга, человека и общества: «Субъект познания с позиций биокультурного соконструктивизма оказывается “привязанным” к конкретной ситуации, которая характеризует особенности отношения социума, культуры и мозга “здесь и сейчас”», что «говорит в пользу пересмотра жестких установок логоцентризма и перспектив деантропологизации знания» [2, с. 85]. Социокультурная детерминированность не только человеческой психики, но и нейронных сетевых систем, выступающих ее биоосновой, требует отказаться от понимания мозга как машины.

В эпистемологическом аспекте косвенные формы лого- и компьютоцентризма также наталкиваются на лимиты. Сфера знания и рассуждения не является тотально формализуемой, в том числе ввиду того, что в любых, даже предельно строгих когнитивных ситуациях первичной и главенствующей оказывается семантическая нагруженность человеческого мышления, а не его синтаксическая правильность: «Сложность наших имита-

ций в машинных моделях мышления заставляет нас забыть про семантическое содержание. Но в реальном мышлении именно семантическое содержание, а не синтаксическое правило, гарантирует значимость вывода» [26, с. 36]. Логика за свою историю была настолько сильно экстрагирована и дистанцирована от реального мышления, что оказалась «ничейной», лишившись любого конкретного наполнения. Именно поэтому она так легко воплотилась в работе машин, в которых очевидна первичность синтаксиса: даже в наиболее развитых системах и языках программирования, где присутствуют компиляторы/интерпретаторы (программы, «понимающие» и «трактующие» код), последние сами функционируют по таким же чисто синтаксическим правилам, будучи, по сути, квазирефлексией синтаксиса над самим собой.

Другое принципиальное ограничение – это наличие в человеческом мышлении больших пластов, которые остаются неэксплицируемыми. Х. Дрейфус ввел понятие периферийного (краевого) сознания [8]. Однако не только дистанцированность от сознательной рефлексии является причиной неуловимости таких частей когнитивной сферы для формализации. Они содержательно являются контекстуальным и ситуативными даже в большей степени, чем «центр» рациональной жизни: «Периферийное, краевое сознание учитывает неявные ориентиры, заключенные в контексте, а также, вероятно, некоторые грамматические конструкции и смысл, значение. Присущее человеку ощущение ситуации позволяет ему исключать из рассмотрения большинство возможных вариантов до всякого точного их анализа» [8, с. 54]. Таким образом, способность к эвристическому схватыванию и интуитивному постижению диктуется накопленными способностями и опытом живой человеческой субъективности, что вновь делает актуальным концепт личностного знания М. Полани. В одной из современных своих версий оно трактуется так: «Это знание, которое укоренено (embodied) в человеке (т. е. оно не существует вне носителя знания (knower)), социально сконструировано (т. е. оно создается совместно индивидом и социально обусловленными смыслами), привязано к практике (т. е. неотделимо от процесса взаимодействия людей), встроено в культуру и традиции общества (т. е. сформировано социокультурной средой, в которой происходит коммуникация носителей знания)» [12, с. 97]. С другой стороны, используется понятие относительно неявного знания, такого неявного знания, которое всё-таки может при надобности быть выражено в дескрипции, наборе команд и правил. Полагается, что именно оно имеет место у человека в рамках его работы в ИК-системах, оперирования их интерфейсами, когда его мышление начинает всё больше функционировать по устанавливаемым ими правилам: «“Относительное” неявное знание всегда присутствует в ИКТ-“гибридах” в явном виде. Ведь действия в

нем требуют осуществления определенного набора алгоритмов. Даже если для пользователей они являются интуитивно понятными, они всё равно могут быть описаны в явном виде в различных источниках. Однако в рамках формирования коммуникационных ИКТ-"гибридов" интерфейс, к помощи которого всегда прибегали вполне конкретные пользователи, оказывается механизмом формирования их коллективного опыта взаимодействия в данном пространстве. В этом случае интерфейс, который сам является в явном виде описанным алгоритмом, позволяет производить не только структуры мира, но и набор реализуемых пользователями практик» [15, с. 81, 82].

На первый взгляд, налицо уже упомянутое сближение: поведение людей, интегрированных во взаимодействие с ИИА, становится всё больше машиноподобным, в то время как они способствуют интериоризации в человеке паттернов поведения и за счет этого кажутся всё более субъектоподобными. Однако возможность аппроксимации определенных видов человеческой деятельности и знания в алгоритмах не означает, что они являются таковыми изначально, эффекты же влияния ИКТ на становление новых жизненных практик и смыслов у человека являются не более чем «побочным продуктом» директивного функционирования алгоритмов. В общем и целом, иллюзия цифровой объективации мышления проистекает из давней метафоры мышления как вычисления, а если говорить более широко – информационного процесса: «Поскольку "когнитивисты" представили познавательные процессы как действие особого устройства, получающего, накапливающего и преобразующего информацию, рациональность этих процессов предстала как характеристика, доступная точному (математическому) выражению. Соответственно отклонения от рациональности могли рассматриваться как следствия воздействия на это устройство различных психологических или социальных факторов» [23, с. 14]. В реальности же психологическое и социокультурное измерения являются не нейтральным фоном или досадной помехой для когнитивных процессов, а их первичным порождающим и детерминирующим фактором, от которого невозможно никак обособиться. Интериоризация и переживание личностью окружающего ее универсума порождают неповторимый ассоциативно-смысловой сплав, который затем воплощается в нечеткой (размытой, плавающей) семантике естественного языка: «В машине нет нечеткой семантики в том виде, в каком она представлена в естественном языке ... нечеткая семантика естественного языка ... связана с приватностью ментальных состояний и уникальностью метафорических и ассоциативных процессов сознания. Правила вывода регулируются как социальной прагматикой, так и языковой семантикой» [3, с. 107, 112].

Внерациональные феномены: «выбросы» или симбионты рациональности?

С другой стороны, современные попытки объективации рациональности далеки от идеи полной негации внерациональных факторов, напротив, пытаются так или иначе ассимилировать и их. Ценность этого очевидна: их воспроизведение придаст импульс для построения более антропоморфных ИИС, способных к автономному функционированию. С другой стороны, с появлением спектра концепций связанной (ограниченной) рациональности внимание стало уделяться не столько ее инвариантно-объективным детерминантам и лимитам, сколько внутренним, субъективным, значительная часть которых внерациональна и результат действия которых состоит в том, что реальные цели и действия не являются совершенно оптимальными, механистически заточенными на максимизацию выгоды/полезности [47]. Признание важности внерациональных факторов в жизни человека сдвигает концептуальный статус рациональности, меняя его с основной причины выбора и действия на некоего арбитра-цензора, который только ограничивает круг вариантов для них: «Есть много случаев, в которых рациональность может лишь исключить некоторые альтернативы, не давая при этом никаких указаний, какой из оставшихся в итоге вариантов следует выбрать. Если мы хотим объяснить поведение в подобных случаях, помимо допущения о рациональности следует также учитывать каузальные соображения» [34, с. 14].

Наиболее ярким примером внерационального являются эмоции, которые, в отличие от эпохи классического рационализма, более не считаются негативным феноменом, только мешающим и запутывающим человеческий разум. Напротив, она стали рассматриваться как важный компонент целостного процесса поведения и принятия решений: «Особенной ролью в принятии решений обладают эмоции как один из триггеров, а также как медиатор процесса. Именно присутствие эмоций позволяет обеспечивать мультиэвристическую активность и определяет смешанную природу когнитивной сферы» [32, с. 82]. Так, эмоции играют важную адаптивную роль в мониторинге и оценке успешности деятельности, достижения цели или удовлетворения потребности: «Динамика взаимодействия информации о потребности и ее удовлетворении непрерывно сопровождается информационным эмоциональным компонентом: отрицательной эмоцией при возникновении потребности и ее неудовлетворении и, наоборот, положительной эмоцией при удовлетворении исходной потребности, а также при предвидении потребного результата» [28, с. 55]. В дополнение современная нейронаука демонстрирует тесную реципрокную связанность субкортикальных (эмоциональных) и кортикальных (рассудочных) центров моз-

га человека на чисто физиологическом уровне. Укорененность эмоций в нижележащих, эволюционно древних нейрональных структурах говорит о том, что они являются не столько «дополнением» к разуму, сколько механизмом, отражающим принципиально неалгоритмизируемую жизнь тела и в этом смысле являющимся «медиумом», обеспечивающим психофизиологическую целостность. А. Дамасио полагает на этот счет: «Чувства кажутся зависимыми от мультикомпонентной системы, которая неотделима от биологической регуляции. Разум кажется зависимым от специфических мозговых систем, некоторые из которых, случается, обрабатывают чувства. Тем самым может быть связующий путь, в анатомическом и функциональном смысле, от разума через чувства к телу» [36, p. 245]. Таким образом, на примере эмоций обнаруживается тесная переплетенность рационального и внерационального, невозможность их изолированного, обособленного существования. Отдельным прикладным эффектом моделирования эмоций полагается обогащение человеко-машинной интеракции, формирование более полноценного «фидбэка» в ней [39]. В ответ на эти запросы возникла концепция аффективного компьютинга [46], которая также демонстрирует трудности и асимметрии в вычислительной репрезентации эмоциональных состояний. Так, во-первых, основные успехи заключены в прогрессе распознавания человеческих эмоций и реагирования на них, нежели в продуцировании эмоций самими искусственными агентами. Во-вторых, на практике стало очевидным, что эмоции – это очень обширное и гетерогенное множество состояний: так, вслед за тем же Дамасио часто выделяют первичные и вторичные эмоции. Первые являются базовыми, в основном эволюционно запрограммированными реакциями и поэтому более или менее общи для всех, вторые же являются приобретаемыми, так что для их понимания и возможного воспроизведения необходимо освоение накопленного субъектом опыта в рамках его индивидуальной жизненной траектории.

Вместе с тем стоит признать, что только эмоций недостаточно для воссоздания подлинно целостной рациональности. Тот же здравый смысл, который является основным арбитром рассуждений человека в обычной, повседневной жизни, сплетен с большим числом таких разнообразных феноменов, как ценности, убеждения, идеалы и т. д.: «Как принадлежность автономной личности, здравый смысл оказывается тесно связан с ценностями человека, его моралью, а также эмоциями. Он не может быть просто “цепью рассудочных размышлений”, потому что людей без эмоций и предпочтений не бывает» [10, с. 179]. Процесс принятия решений, собственно, и позволяющий перейти от рассуждения и выбора к активному действию, также оказывается интегральным, представляя собой «равнодействующую» различных по степени рациональности факторов: «Всякий сознательный

человек преследует определенные цели и принимает соответствующие решения, связанные с их достижением. Полагаю, что можно было бы даже определить человека как существо, принимающее решения. Этим бы подчеркивалось, что принятие решения – это деятельность, в основе которой лежит привлечение и проявление самых разных потенций – интеллектуальных, волевых, эмоциональных, духовных, нравственных» [6, с. 43]. Следовательно, для воспроизведения таких сложных гетерогенных феноменов нужны не аналоги отдельных компонент, а аналог их целостного взаимосвязанного комплекса. Особенно это станет насущным в попытках придать ИИ способность не просто совершать стандартные действия и решать типичные проблемы, но работать в нештатных ситуациях, эвристически открывать или создавать новое. Предельным случаем такого вектора является подлинно творческая деятельность. Но она явнее всего демонстрирует и максимальную удаленность от усредненной и стандартизированной алгоритмичности, синергетическую слитность множества разновекторных способностей: «Творческий акт включает весь опыт личности, эвристику, волю, эстетическое чутье» [29, с. 37].

Отдельным комплексным феноменом, не являющимся чисто рациональным, но тем не менее тесно связанным с рациональностью, является мораль. Есть веские основания считать, что этические параметры должны быть учтены, чтобы рациональность стала цельной, а сложная деятельность на ее основе (например, управленческая или творческая) полноценной: «Рационализация моральных ценностей – это необходимое условие ввода понятия морали в систему управления. Нужна моральная воля свободного субъекта. Только тогда он начинает действовать рационально, как цельная личность. Мораль можно назвать пределом рациональности» [20, с. 104]. В.С. Стёпин полагает, что в условиях постнеклассического типа рациональности «в стратегиях деятельности со сложными, человекоразмерными системами возникает новый тип интеграции истины и нравственности, целе-рационального и ценностно-рационального действия» [27, с. 75]. Очевидно, что усиливающаяся автономизация ИИА порождает императив этического контроля их функционирования, учета ценностных ориентиров, характерных для человеческого мировоззрения [35]. Причем данные механизмы контроля должны быть интернализированы, встроены в них и быть способными работать априорно, т. е. не на стадии совершения действия, а уже на стадии зарождения целеподобных актов, эффективно подавляя/отбраковывая неприемлемые.

Вместе с тем путь «оцифровки» эмоциональных, аффективных, моральных, аксиологических явлений наталкивается на труднопреодолимые препятствия, так как данные феномены невыразимы формально-логическими и алгоритмическими средствами по причине своей крайней субъ-

ективной пластичности и плюральности. Во многих из них нарушаются фундаментальные формальные законы и принципы: переживания, чувства, эмоции представляют собой «поток», текучесть которого не вписывается в привычную для логики дискретность, не поддаются однозначной формальной интерпретации в качестве положительных/отрицательных (и тем более истинных/ложных) ввиду своей сильной зависимости от смыслового контекста. Взгляды, убеждения, мнения, намерения неминуемо содержат противоречия, которые всегда были и остаются формально-логическим маркером некорректности: «Постулат строгого и неукоснительного соблюдения принципа противоречия привел бы к тому, что появление в какой-нибудь системе убеждений пары противоречивых предложений выводило бы за пределы рациональности. Между тем действительные убеждения, питаемые людьми, часто заключают в себе хотя бы единичные противоречия» [33, с. 158]. Стоит констатировать, что наличные ИИА не приспособлены для внерациональных феноменов, обладающих гигантским количеством степеней свободы, сильно зависящих от большого количества субъективных и ситуативных факторов. В человеке эти разные психические феномены являются неслучайными, бифуркационными «отклонениями» от рациональности, но сосуществующими с ней в своего рода симбиозе ценными ее расширениями. Неспособность их алгоритмического повторения поэтому препятствует воспроизведению самой рациональности во всей ее полноте.

Гибкая человеческая versus ригидная машинная рациональность

Идея тотальной алгоритмизируемости рациональности оказывается контрпродуктивной: она так или иначе способствует ее «застыванию», ригидности в виде компактного набора инвариантных схем. Гипертрофия нормативного элемента в итоге превращает ее в гиперрациональность sui generis. Отличием же человеческой психики является постоянная «перенастройка», мягкий переход от одних способов мышления и линий поведения к другим в зависимости от изменения как внешних условий среды, так и внутреннего смыслового остова. Ориентиром поэтому, напротив, должна стать гибкая рациональность: «“Гибкая” рациональность предстает как логическое познание в сочетании с дологическими и антропологическими предпосылками. Ведь сама природа мысли, всегда принадлежащей субъекту, обусловливает детерминацию ее содержания и формы природой и спецификой ее носителя – субъекта, заставляет быть гибкой “по определению”. Становление гибкой рациональности – процесс вероятностный, а не алгоритмизированный» [14, с. 34–36]. Сам человек генеалогически возник как гибкая структура, ориентированная на интерактивное приспосо-

бление и преобразование среды. Но в нем обязательно должно присутствовать и устойчивое ядро, которое позволяет ему сохранять и воспроизводить общие модели и нормы общественного поведения, паттерны культуры, наконец, собственную самоидентичность: «Свойство "быть субъектом" сформировалось у человека как инструмент адаптации к изменчивой, вероятностной среде и способ сохранения совокупного социального опыта и знания. Субъект-индивид как открытая самоорганизующаяся система, взаимодействуя со средой, осваивает, сохраняет, воспроизводит способы предметно-практической деятельности, модели коммуникации и познания, поддерживая тем самым существование культуры и общества» [21, с. 33]. Гибкость и пластичность – интегральные показатели, имеющие как социокультурные, так и биоадаптивные основания, и поэтому принципиально неуловимые для онтологически относительно простых (технофизических) вычислительных машин.

Отличительной чертой человеческого интеллекта является тонкий сбалансированный паритет между гибкостью/пластичностью и устойчивостью/определенностью, который автореферентно сам оказывается подвижным. Живая рациональность – плавающая величина, которая интерактивно меняется в зависимости от самой человеческой реальности, постоянно превосходя себя в новых видах и формах: «Мышление рационально, если оно способно преодолевать собственную рациональность, трансцендируя, "превосходя себя" с иной рациональности, которая также, в свою очередь, будет преодоленной, когда в этом назревает необходимость. У рациональности нет "массы покоя"» [24, с. 70]. Искусственная же цифровая рациональность гипертрофирует одну сторону, нормативную, что изначально сужает окно возможностей воспроизводства ею комплексных свойств человеческой психики. Несмотря на наличие у ее алгоритмов некоторых возможностей научения и самокорректирования в ходе обработки данных, она не способна к глубинному содержательному пересмотру и преодолению себя, так как это требовало бы отхода от автоматического функционирования: «Так называемая "автономия" ИИ – не что иное, как автоматизм выполнения задач на основе полученных от датчиков данных (при этом он сильно ограничен, несмотря на способность ИИ вероятностно адаптироваться к новым данным при помощи "машинного обучения")» [18, с. 70, 71].

Заключение. Раскол рациональности как риск ее вычислительной объективации. Перспективы цифрового гуманизма

Будучи закономерными, процессы цифровизации сами начинают вносить существенную лепту в неопределенность и нелинейность дальнейшего развития, порождать новые комплексные социальные и природные риски.

«Цифровизация – объективно востребованное явление, появившееся в условиях нелинейной динамики рационализации, достижений и побочных эффектов науки и технологий, амбивалентности искусственного интеллекта. Однако цифровизация порождает социальные практики с внутрисистемной неопределенностью, что создает объективные условия для сложных уязвимостей для общества и природы» [42, p. 11]. Отчасти данные риски порождаются автономизацией ИИА, с другой же стороны, недостаточностью человеческого разума в деле оптимизации, прогнозирования и планирования экспоненциального цифрового прогресса. Человек как бы «запутывается» в многообразии его возможностей и потенциальных траекторий, не имея достаточно средств для эффективного отбора наиболее приемлемых из них.

Данная ситуация делает актуальной гипотезу техногуманитарного баланса А.П. Назаретяна: «Формальная версия гипотезы демонстрирует, что более мощные технологии повышают устойчивость социальной системы к внешним колебаниям и в то же время делают ее более уязвимой внутренне» [45, p. 80]. Современное нарушение баланса в данном аспекте связано с тем, что человеческая субъективность, ее культурно-духовное измерение не «успевает» приспособиться к цифровым технологиям, что превращает их в фактор дезадаптации и стресса. Тем самым внешний прогресс в дигитализации жизни не отменяет внутренней дефектности такой всё более асимметричной динамики. Попытки искусственно гипостазировать рациональность, оторвать ее от человека только усугубляет ее деантропологизацию, а также vice versa отчуждение человека от нее.

Альтернативой видится поддержание и усиление человекоразмерной составляющей рациональности, становление нового цифрового гуманизма. «Только гуманизм, способный вобрать в себя ИИ как (контролируемый) “инструмент”, может быть разумным ответом на идеологически мотивированные неверные интерпретации искусственного интеллекта, ... цифровой гуманизм, то есть гуманизм, не сомневающийся в человеческом творчестве и не ставящий его под угрозу, а лишь расширяющий его через внедрение цифровых технологий» [18, с. 72, 73]. С этим солидарен С.А. Кравченко, вводящий удачное понятие «гуманистический цифровой поворот»: «Современные общества имеют принципиально новую задачу – сделать цифровизацию гуманистичной. Концепция гуманистического цифрового поворота подразумевает интеграцию цифровых достижений для сохранения и воспроизводства основных культурных ценностей» [42, p. 15]. В общем и целом, налицо дисбаланс между технологическим и гуманитарным измерением, который накладывается и на рациональность. Поэтому последний эпитет рациональности в данной ситуации – это «расколотая» рациональность, и одна из главных задач описанного цифрового гуманизма – сопротивление расколу в ней.

В итоге основным препятствием для объективизации рациональности в ИИС стоит признать выраженную асимметрию, когда одни ее составляющие уже хорошо переводимы на язык вычислений и алгоритмов, а другие – слабо или вовсе не переводимы, а также реципрокную сплавленность рационального и внерационального в реальной ментальной жизни. Искусственный отрыв отдельных рациональных феноменов, превращение их в ригидный набор формальных паттернов не является рациональным. Отдельную озабоченность вызывает встречная по отношению к bottom-up тенденции «интеллектуализации» машин top-down тенденция «машинизации» человека, которая лишает его онтоантропологической специфики. Экстраполяция парадигмы компьютационализма на него в контексте проведенного анализа кажется нерелевантной. В современной ситуации популярность вычислительных интерпретаций человеческой рациональности является по большей части эпифеноменом инвазивного проникновения технологий в различные аспекты и практики реальной общественной жизни, ее цифровой фетишизации, требующей отдельного глубокого анализа.

Литература

1. *Асташова Н.Д.* В пространстве «навязанной рациональности». – Н. Новгород: Нижегор. гос. ун-т им. Н.И. Лобачевского, 2020. – 178 с.

2. *Бажанов В.А.* Социум и мозг: биокультурный со-конструктивизм // Вопросы философии. – 2018. – № 2. – С. 78–88.

3. *Барышников П.Н.* Семантические процессы сознания: от вычислительных моделей к языковому опыту // Эпистемология и философия науки. – 2014. – Т. 41, № 3. – С. 96–114.

4. *Бостром Н.* Искусственный интеллект: этапы, угрозы, стратегии. – М.: Манн, Иванов и Фербер, 2016. – 496 с.

5. *Винник Д.В.* Физические, функциональные и ментальные состояния: проблема соотношения // Философия науки. – 2010. – № 2 (45). – С. 92–104.

6. *Диев В.С.* Неопределенность, риск и принятие решений в междисциплинарном контексте // Сибирский философский журнал. – 2019. – Т. 17, № 4. – С. 41–52. – DOI: 10.25205/2541-7517-2019-17-4-41-52.

7. *Драгалина-Черная Е.Г.* Рациональность действия, рациональность правила, рациональность цели: рассуждение как case-study // Рацио.ru. – 2015. – № 15. – С. 28–40.

8. *Дрейфус Х.* Чего не могут вычислительные машины: критика искусственного разума. – М.: Либроком, 2010. – 336 с.

9. *Дубровский Д.И.* Проблема идеального. Субъективная реальность. – М.: Канон+, 2002. – 368 с.

10. *Золотухина-Аболина Е.В.* Здравый смысл и иррациональное // Эпистемология и философия науки. – 2016. – Т. 48, № 2. – С. 176–192.

11. *Игнатьев В.П.* Проблема техносубъекта: о субъектности «сущностей-конструкторов» // Идеи и идеалы. – 2021. – т. 13, № 1-1. – С. 130–150. – DOI: 10.17212/2075-0862-2021-13.1.1-130-150.

12. *Каныгин Г.В., Кононова О.В.* Прагматическая эпистемология: подходы к выражению неявного знания социальными акторами // Социология науки и технологий. – 2021. – Т. 12, № 4. – С. 93–115. – DOI: 10.24412/2079-0910-2021-4-93-115.

13. *Лепский В.Е.* Рефлексивность в управлении социальными системами (философско-методологический анализ) // Философия науки и техники. – 2021. – Т. 26, № 2. – С. 127–147. – DOI: 10.21146/2413-9084-2021-26-2-127-147.

14. *Масалова С.П.* Гибкая рациональность уплотнения научного знания: когнитивный аспект // Вестник Томского государственного университета. Философия. Социология. Политология. – 2010. – № 2 (10). – С. 32–44.

15. *Масланов Е.В., Фейгельман А.М.* Неявное знание в интернет-коммуникации: интерфейс как механизм производства неявного знания // Вестник Томского государственного университета. – 2020. – № 460. – С. 77–83. – DOI: 10.17223/15617793/460/9.

16. *Минский М.* Сообщество разума. – М.: АСТ, 2018. – 592 с.

17. *Михайлов П.Ф.* Вычислительный подход в социальном познании // Философия науки и техники. – 2021. – Т. 26, № 1. – С. 23–37. – DOI: 10.21146/2413-9084-2021-26-1-23-37.

18. *Нагль Л.* Цифровые технологии: размышления о различии между инструментальной рациональностью и практическим разумом // Кантовский сборник. – 2022. – Т. 41, № 1. – С. 60–88. – DOI: 10.5922/0207-6918-2022-1-3.

19. *Нейман Дж. фон.* Вычислительная машина и мозг. – М.: АСТ, 2018. – 192 с.

20. *Нигоматуллина Р.М.* Мораль и рациональность в управлении // Ученые записки Казанского университета. Серия: Гуманитарные науки. – 2015. – Т. 157, кн. 1. – С. 102–106.

21. *Никитина Е.А.* Проблема субъектности в интеллектуальной робототехнике // Философские проблемы информационных технологий и киберпространства. – 2016. – № 2 (12). – С. 31–39. – DOI: 10.17726/philIT.2016.12.2.3.

22. *Патнэм Х.* Философия сознания. – М.: Дом интеллектуальной книги, 1999. – 280 с.

23. *Порус В.Н.* Многомерность рациональности // Эпистемология и философия науки. – 2010. – Т. 23, № 1. – С. 5–16.

24. *Порус В.Н.* Рациональная коммуникация как проблема эпистемологии // Эпистемология и философия науки. – 2008. – Т. 17, № 3. – С. 57–70.

25. *Рассел С., Норвиг П.* Искусственный интеллект. Современный подход. – 2-е изд. – М.: Вильямс, 2006. – 1408 с.

26. *Серл Дж.* Рациональность в действии. – М.: Прогресс-Традиция, 2004. – 336 с.

27. *Стёпин В.С.* Системность объектов научного познания и типы рациональности // Вестник Томского государственного университета. Философия. Социология. Политология. – 2007. – № 1 (1). – С. 65–76.

28. *Судаков К.В.* Информационные аспекты системной организации психической деятельности // Вестник Российской академии медицинских наук. – 2012. – Т. 67, № 8. – С. 53–56. – DOI: 10.15690/vramn.v67i8.350.

29. *Финн К.В.* Искусственный интеллект: методология применения, философия. – М.: Ленанд, 2021. – 468 с.

30. *Хоркхаймер М.* Затмение разума: к критике инструментального разума. – М.: Канон+, 2011. – 224 с.

31. *Шалак В.И.* Логика в онтологии процессов // Логические исследования. – 2021. – Т. 27, № 2. – С. 48–65. – DOI: 10.21146/2074-1472-2021-27-2-48-65.

32. *Шиллер А.В.* Роль эмоций в когнитивно-аффективной системе как фактор развития архитектуры искусственных агентов // Вестник Московского университета. Серия 7, Философия. – 2018. – № 6. – С. 78–90.

33. *Шульга Е.Н.* Природа научного познания и критерии рациональности // Философия науки. – 2004. – Т. 10, № 1. – С. 151–171.

34. *Эльстер Ю.* Кислый виноград. Исследование провалов рациональности. – М.: Изд-во Ин-та Гайдара, 2018. – 296 с.

35. Ethical Management of Artificial Intelligence / A.B. Brendel, M. Mirbabaie, T.-B. Lembcke, L. Hofeditz // Sustainability. – 2021. – Vol. 13 (4). – P. 1974. – DOI: 10.3390/su13041974.

36. *Damasio A.R.* Descartes' error: Emotion, reason, and the human brain. – New York: HarperCollins, 1998. – 312 p.

37. *De Jesus P.* Autopoietic enactivism, phenomenology and the deep continuity between life and mind // Phenomenology and the Cognitive Sciences. – 2016. – Vol. 15, N 2. – P. 265–289. – DOI: 10.1007/s11097-015-9414-2.

38. *Gershman S.J., Horvitz E.J., Tenenbaum J.B.* Computational rationality: A converging paradigm for intelligence in brains, minds, and machines // Science. – 2015. – Vol. 349, N 6245. – P. 273–278. – DOI: 10.1126/science.aac6076.

39. *Hortensius R., Hekele F., Cross E.S.* The perception of emotion in artificial agents // IEEE Transactions on Cognitive and Developmental Systems. – 2018. – Vol. 10, N 4. – P. 852–864. – DOI: 10.1109/TCDS.2018.2826921.

40. *Hutto D.D., Myin E.* Evolving enactivism: Basic minds meet content. – Cambridge: MIT press, 2017. – 328 p.

41. «Know thyselves»-computational self-reflection in collective technical systems / J. Haehner, S. von Mammen, S. Timpf, S. Tomforde, B. Sick, K. Geihs, T. Goeble, G. Hornung, G. Stumme // ARCS 2016: 29th International Conference on Architecture of Computing Systems. – VDE, 2016. – P. 1–8.

42. *Kravchenko S.A.* From formal rationality to the digital one: Sideeffects, ambivalences, and vulnerabilities // RUDN Journal of Sociology. – 2021. – Vol. 21, N 1. – P. 7–17. – DOI: 10.22363/2313-2272-2021-21-1-7-17.

43. *Lewis R.L., Howes A., Singh S.* Computational rationality: Linking mechanism and behavior through bounded utility maximization // Topics in Cognitive Science. – 2014. – Vol. 6, N 2. – P. 279–311. – DOI: 10.1111/tops.12086.

44. *Mikhailov I.F.* Social Ontology: Time to Compute // Вестник Томского государственного университета. Философия. Социология. Политология. – 2020. – № 55. – С. 36–46. – DOI: 1.17223/1998863X/55/5.

45. *Nazaretyan A.P.* Big (Universal) History Paradigm: Versions and Approaches // Social Evolution and History. – 2005. – Vol. 4, N 1. – P. 61–86.

46. *Picard R.W.* Affective computing. – Cambridge: MIT press, 2000. – 292 p.

47. *Simon H.A.* Bounded rationality in social science: Today and tomorrow // Mind & Society. – 2000. – Vol. 1, N 1. – P. 25–39.

48. *Gerven M.A.J. van.* Computational foundations of natural intelligence // Frontiers in Computational Neuroscience. – 2017. – Vol. 11 – P. 112. – DOI: 10.1101/166785.

49. *Wells A.* Rethinking cognitive computation: Turing and the science of the mind. – London: Bloomsbury Pub., 2017. – 288 p.

Статья поступила в редакцию 07.09.2022.
Статья прошла рецензирование 18.12.2022.

DOI: 10.17212/2075-0862-2023-15.2.1-72-96

PROBLEM OF COMPUTATIONAL OBJECTIVATION OF RATIONALITY IN ARTIFICIAL INTELLECTUAL AGENTS

Zhelnin, Anton,
Cand. of Sc. (Philosophy), Associate Professor,
Department of Philosophy,
Perm State University,
15 Bukireva Street, Perm, 614068, Russian Federation
ORCID: 0000-0002-6368-1363
antonzhelnin@gmail.com

Abstract

The subject of the article is objectification of rationality in artificial intelligent agents (AIA). The author considers two complementary trends in its context. The first 'bottom-up' trend is associated with attempts to artify rational reasoning and action in AIA, the second 'top-down' one is associated with attempts to interpret human thinking and behavior in machine terms. The first one is limited by the lack of semantic content, personal coloring and full-fledged psychosomatic embodiment of computational processes in AIA. Hypertrophy of logical normative components in them leads to mechanistic rigidity. The second is generated by the expansion of digital technologies into real life, their active fusion, which gives rise to the ideology of computationalism, according to which people are computing agents. At the same time, the convergence of the 'intellectualization' of AIA and the 'machinization' of human is a false appearance, since there are fundamental ontological and epistemological limits. The main ontological limit: human rationality is based on socio-cultural and bio-adaptive layers of the existence of a human subject, and therefore cannot be principally reproduced in ontologically simple, techno-physical AIA. The main epistemological limit: human thinking cannot be formalized and computationally objectified, since a meaningful semantic core is primary in it, which bears the stamp of subjectivity, and is also strongly linked to layers of implicit personal knowledge and other components of consciousness. Among them a number of sub- and non-rational phenomena stand out (emotions, values, common sense, morality), which are maximally distanced from algorithmic averaging, but without interaction with them rationality remains essentially reduced. It is concluded that the computational representation and objectification of definite components of rationality in AIA is possible, but is not rational itself, since their artificial separation from the larger, human-sized part of rationality initiates a split in it.

Keywords: rationality, computation, computational rationality, flexible rationality, sub-ject, rational agent, intelligent system, artificial intelligence, computationalism, digital humanism.

Bibliographic description for citation:
Zhelnin A. Problem of Computational Objectification of Rationality in Artificial Intellectual Agents. *Idei i idealy = Ideas and Ideals*, 2023, vol. 15, iss. 2, pt. 1, pp. 72–96. DOI: 10.17212/2075-0862-2023-15.2.1-72-96.

References

1. Astashova N.D. *V prostranstve «navyazannoi ratsional'nosti»* [In space of imposed rationality]. Nizhny Novgorod, Lobachevsky University Publ., 2020. 178 p.

2. Bazhanov V.A. Sotsium i mozg: biokul'turnyi so-konstruktivizm [Socium and brain: biocultural co-constructivism]. *Voprosy filosofii = Russian Studies in Philosophy*, 2018, no. 2, pp. 78–88. (In Russian).

3. Baryshnikov P.N. Semanticheskie protsessy soznaniya: ot vychislitel'nykh modelei k yazykovomu opytu [Semantic processes of consciousness: from computational models to linguistic experience]. *Epistemologiya i filosofiya nauki = Epistemology & Philosophy of Science*, 2014, vol. 41, no. 3, pp. 96–114. (In Russian).

4. Bostrom N. *Superintelligence: paths, dangers, strategies*. Oxford, Oxford University Press, 2014 (Russ. ed.: Bostrom N. *Iskusstvennyi intellect: etapy, ugrozy, strategii*. Moscow, Mann, Ivanov i Ferber Publ., 2016. 496 p.).

5. Vinnik D.V. Fizicheskie, funktsional'nye i mental'nye sostoyaniya: problema sootnosheniya [Physical, functional and mental states: the problems of Their correlation]. *Filosofiya nauki = Philosophy of Science*, 2010, no. 2 (45), pp. 92–104.

6. Diev V.S. Neopredelennost', risk i prinyatie reshenii v mezhdistsiplinarnom kontekste [Uncertainty, risk and decision-making in an interdisciplinary context]. *Sibirskii filosofskii zhurnal = Siberian Journal of Philosophy*, 2019, vol. 17, no. 4, pp. 41–52. DOI: 10.25205/2541-7517-2019-17-4-41-52.

7. Dragalina-Chernaya E.G. Ratsional'nost' deistviya, ratsional'nost' pravila, ratsional'nost' tseli: rassuzhdenie kak case-study [Rationality of action, rationality of rule, rationality of aim: reasoning as case-study]. *Ratsio.ru*, 2015, no. 15, pp. 28–40. (In Russian).

8. Dreyfus H.L. *Chego ne mogut vychislitel'nye mashiny: kritika iskusstvennogo razuma* [What computers still can't do: a critique of artificial reason]. Moscow, Librokom Publ., 2010. 336 p. (In Russian).

9. Dubrovskii D.I. *Problema ideal'nogo. Sub"ektivnaya real'nost'* [Consciosness, brain, artificial intelligence]. Moscow, Kanon+ Publ., 2002. 368 p.

10. Zolotukhina-Abolina E.V. Zdravyi smysl i irratsional'noe [Common sense and the irrational]. *Epistemologiya i filosofiya nauki = Epistemology & Philosophy of Science*, 2016, vol. 48, no. 2, pp. 176–192. (In Russian).

11. Ignatyev V.I. Problema tekhnosub"ekta: o sub"ektnosti «sushchnostei-konstruktorov» [Problem of the Techno-subject: On the Subjectivity of "Entity-Constructors"]. *Idei i idealy = Ideas and Ideals*, 2021, vol. 13, iss. 1-1, pp. 130–150. DOI: 10.17212/2075-0862-2021-13.1.1-130-150.

12. Kanygin G.V., Kononova O.V. Pragmaticheskaya epistemologiya: podkhody k vyrazheniyu neyavnogo znaniya sotsial'nymi aktorami [Pragmatical epistemolo-

gy: approaches to the expression of implicit knowledge by social actors]. *Sotsiologiya nauki i tekhnologii = Sociology of Science and Technology*, 2021, vol. 12, no. 4, pp. 93–115. DOI: 10.24412/2079-0910-2021-4-93-115.

13. Lepskiy V.E. Refleksivnost' v upravlenii sotsial'nymi sistemami (filosofsko-metodologicheskii analiz) [Reflexivity in social systems control (philosophical and methodological analysis)]. *Filosofiya nauki i tekhniki = Philosophy of Science and Technology*, 2021, vol. 26, no. 2, pp. 127–147. DOI: 10.21146/2413-9084-2021-26-2-127-147.

14. Masalova S.I. Gibkaya ratsional'nost' uplotneniya nauchnogo znaniya: kognitivnyi aspekt [Flexible rationality of sealing of scientific knowledge: cognitive aspect]. *Vestnik Tomskogo gosudarstvennogo universiteta. Filosofiya. Sotsiologiya. Politologiya = Tomsk State University Journal of Philosophy, Sociology and Political Science*, 2010, no. 2 (10), pp. 32–44.

15. Maslanov E.V., Feigelman A.M. Neyavnoe znanie v internet-kommunikatsii: interfeis kak mekhanizm proizvodstva neyavnogo znaniya [Tacit Knowledge in Internet Communication: Interface as a Machine for Tacit Knowledge Production]. *Vestnik Tomskogo gosudarstvennogo universiteta = Tomsk State University Journal*, 2020, no. 460, pp. 77–83. DOI: 10.17223/15617793/460/9. (In Russian).

16. Minskii M. *Soobshchestvo razuma* [Society of mind]. Moscow, AST Publ., 2018. 592 p.

17. Mikhailov I.F. Vychislitel'nyi podkhod v sotsial'nom poznanii [Computational approach to social knowledge]. *Filosofiya nauki i tekhniki = Philosophy of Science and Technology*, 2021, vol. 26, no. 1, pp. 23–37. DOI: 10.21146/2413-9084-2021-26-1-23-37.

18. Nagl L. Tsifrovye tekhnologii: razmyshleniya o razlichii mezhdu instrumental'noi ratsional'nost'yu i prakticheskim razumom [Digital technology: reflections on the difference between instrumental rationality and practical reason]. *Kantovskii sbornik = Kantian Journal*, 2022, vol. 41, no. 1, pp. 60–88. DOI: 10.5922/0207-6918-2022-1-3.

19. Von Neumann J. *The computer & the brain*. 3rd ed. New Haven, London, Yale University Press, 2012 (Russ. ed.: Fon Neiman Dzh. *Vychislitel'naya mashina i mozg*. Moscow, AST Publ., 2018. 192 p.).

20. Nigomatullina R.M. Moral' i ratsional'nost' v upravlenii [Morality and rationality in administration]. *Uchenye zapiski Kazanskogo universiteta. Seriya: Gumanitarnye nauki*, 2015, vol. 157, no. 1, pp. 102–106. (In Russian).

21. Nikitina E.A. Problema sub''ektnosti v intellektual'noi robototekhnike [The problem of subjectivity in intellectual robotics]. *Filosofskie problemy informatsionnykh tekhnologii i kiberprostranstva = Phillosophical problems of Information Technologies and cyberspace*, 2016, no. 2 (12), pp. 31–39. DOI: 10.17726/philIT.2016.12.2.3.

22. Putnam H. *Filosofiya soznaniya* [Philosophy of consciousness]. Moscow, Dom intellektual'noi knigi Publ., 1999. 280 p. (In Russian).

23. Porus V.N. Mnogomernost' ratsional'nosti [Multidimensiality of rationality]. *Epistemologiya i filosofiya nauki = Epistemology & Philosophy of Science*, 2010, vol. 23, no. 1, pp. 5–16. (In Russian).

24. Porus V.N. Ratsional'naya kommunikatsiya kak problema epistemologii [Rational communication as problem of epistemology]. *Epistemologiya i filosofiya nauki = Epistemology & Philosophy of Science*, 2008, vol. 17, no. 3, pp. 57–70. (In Russian).

25. Russel S.J., Norvig P. *Iskusstvennyi intellekt. Sovremennyi podkhod* [Artificial intelligence: A Modern approach]. Moscow, Vil'yams Publ., 2006. 1408 p. (In Russian).

26. Searle J.R. *Rationality in action*. Cambridge, A Bradford book, The MIT Press, 2001 (Russ. ed.: Serl Dzh. *Ratsional'nost' v deistvii*. Moscow, Progress-Traditsiya Publ., 2004. 336 p.).

27. Stepin V.S. Sistemnost' ob"ektov nauchnogo poznaniya i tipy ratsional'nosti [Consistency of objects of scientific cognition and typees of rationality]. *Vestnik Tomskogo gosudarstvennogo universiteta. Filosofiya. Sotsiologiya. Politologiya = Tomsk State University Journal of Philosophy, Sociology and Politology*, 2007, no. 1 (1), pp. 65–76.

28. Sudakov K.V. Informatsionnye aspekty sistemnoi organizatsii psikhicheskoi deyatel'nosti [The informative mechanisms of systemic organization of psychic activity]. *Vestnik Rossiiskoi akademii meditsinskikh nauk = Annals of the Russian academy of medical sciences*, 2012, vol. 67, no. 8, pp. 53–56. DOI: 10.15690/vramn.v67i8.350.

29. Finn K.V. *Iskusstvennyi intellekt: metodologiya primeneniya, filosofiya* [Artificial intelligence: Methodology, applications, philosophy]. Moscow, Lenand Publ., 2021. 468 p.

30. Horkheimer M. *Zatmenie razuma: k kritike instrumental'nogo razuma* [Eclipse of reason]. Moscow, Kanon+ Publ., 2011. 224 p. (In Russian).

31. Shalak V.I. Logika v ontologii protsessov [Logic in the Process Ontology]. *Logicheskie issledovaniya = Logical Investigations*, 2021, vol. 27, no. 2, pp. 48–65. DOI: 10.21146/2074-1472-2021-27-2-48-65.

32. Shiller A.V. Rol' emotsii v kognitivno-affektivnoi sisteme kak faktor razvitiya arkhitektury iskusstvennykh agentov [Role of emotions in cognitive-affective system as factor of development of artificial agents' architecture]. *Vestnik Moskovskogo universiteta. Seriya 7, Filosofiya = Russian Studies in Philosophy*, 2018, no. 6, pp. 78–90. (In Russian).

33. Shul'ga E.N. Priroda nauchnogo poznaniya i kriterii ratsional'nosti [Nature of scientific cognition and criteria of rationality]. *Filosofiya nauki i tekhniki = Philosophy of Science and Technology*, 2004, vol. 10, no. 1, pp. 151–171. (In Russian).

34. Elster J. *Kislyi vinograd. Issledovanie provalov ratsional'nosti* [Sour grapes. Studies in subversion of rationality]. Moscow, Institut Gaidara Publ., 2018. 296 p. (In Russian).

35. Brendel A.B., Mirbabaie M., Lembcke T.-B., Hofeditz L. Ethical Management of Artificial Intelligence. *Sustainability*, 2021, vol. 13 (4), p. 1974. DOI: 10.3390/su13041974.

36. Damasio A.R. *Descartes' error: Emotion, reason, and the human brain*. New York, HarperCollins, 1998. 312 p.

37. De Jesus P. Autopoietic enactivism, phenomenology and the deep continuity between life and mind. *Phenomenology and the Cognitive Sciences*, 2016, vol. 15, no. 2, pp. 265–289. DOI: 10.1007/s11097-015-9414-2.

38. Gershman S.J., Horvitz E.J., Tenenbaum J.B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 2015, vol. 349, no. 6245, pp. 273–278. DOI: 10.1126/science.aac6076.

39. Hortensius R., Hekele F., Cross E.S. The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, 2018, vol. 10, no. 4, pp. 852–864. DOI: 10.1109/TCDS.2018.2826921.

40. Hutto D.D., Myin E. *Evolving enactivism: Basic minds meet content*. Cambridge, MIT press, 2017. 328 p.

41. Haehner J., Mammen S. von., Timpf S., Tomforde S., Sick B., Geihs K., Goeble T., Hornung G., Stumme G. «Know thyselves»-computational self-reflection in collective technical systems. *ARCS 2016: 29th International Conference on Architecture of Computing Systems*. VDE, 2016, pp. 1–8.

42. Kravchenko S.A. From formal rationality to the digital one: Sideeffects, ambivalences, and vulnerabilities. *RUDN Journal of Sociology*, 2021, vol. 21, no. 1, pp. 7–17. DOI: 10.22363/2313-2272-2021-21-1-7-17.

43. Lewis R.L., Howes A., Singh S. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, 2014, vol. 6, no. 2, pp. 279–311. DOI: 10.1111/tops.12086.

44. Mikhailov I.F. Social Ontology: Time to Compute. *Vestnik Tomskogo gosudarstvennogo universiteta. Filosofiya. Sotsiologiya. Politologiya = Tomsk State University Journal of Philosophy, Sociology and Politology*, 2020, no. 55, pp. 36–46. DOI: 1.17223/1998863X/55/5.

45. Nazaretyan A.P. Big (Universal) History Paradigm: Versions and Approaches. *Social Evolution and History*, 2005, vol. 4, no. 1, pp. 61–86.

46. Picard R.W. *Affective computing*. Cambridge, MIT press, 2000. 292 p.

47. Simon H.A. Bounded rationality in social science: Today and tomorrow. *Mind & Society*, 2000, vol. 1, no. 1, pp. 25–39.

48. Gerven M.A.J. van. Computational foundations of natural intelligence. *Frontiers in Computational Neuroscience*, 2017, vol. 11, p. 112. DOI: 10.1101/166785.

49. Wells A. *Rethinking cognitive computation: Turing and the science of the mind*. London, Bloomsbury Pub., 2017. 288 p.

The article was received on 07.09.2022.
The article was reviewed on 18.12.2022.