

ПРОБЛЕМЫ УПРАВЛЕНИЯ ИСКУССТВЕННОЙ ЛИЧНОСТЬЮ

О.Н. Гуров, А.В. Шерстов

Аннотация. Сегодня большая часть исследователей как в области технических наук, так и представляющих гуманитарное знание, считают необходимым определять искусственный интеллект с учетом приложения к нему субъективных «человеческих» качеств, к которым относятся способность к самосознанию, а также к свободному выбору. В этой связи чрезвычайно актуальной становится проблема автономии ИИ, а далее – прав и возможностей (или отсутствия и отказа от возможности) для создателя ИИ по осуществлению контроля над «продуктом». В рамках исследования этой проблематики уже 20 лет существует и развивается проект «Искусственная личность». Несмотря на то, что в рамках данного проекта ведется активная научная и общественная деятельность, в которую вовлечены ведущие представители научного сообщества, сам проект еще далек от завершения. Представленная статья подводит итог выполненным исследованиям по концептуализации искусственной личности (ИЛ), материалы публикации демонстрируют, что на сегодняшний день даже принципиальная возможность создания ИЛ еще убедительно не доказана. Также концептуально не сформулирован единый общепринятый подход к перспективным методам и технологиям реализации и воплощения ИЛ. Таким образом, на современном этапе изучение ИЛ представляет собой по большей мере абстрактные теоретические исследования. В результате проведенного анализа авторы статьи приходят к выводу, что на сегодняшний день наиболее целесообразно использовать результаты исследований Естественной Личности и Естественного Интеллекта и переносить методы, доказавшие свою относительную эффективность в разнообразных существующих проявлениях реальной социальной жизни, в область создания концепции ИЛ. Предлагаемый подход к концептуализации ИЛ может быть полезным для создания теоретико-методологического фундамента дальнейших теоретических исследований и проектов реализации ИЛ.

Ключевые слова: Искусственный интеллект, искусственная личность, тест Тьюринга, этика ИИ, самость, идентичность, естественная личность, естественный интеллект

Для цитирования: Гуров О.Н., Шерстов А.В. (2023). Проблемы управления искусственной личностью. – *Исследования в цифровой экономике*. №1, С. 61–89. DOI: [10.24833/14511791-2023-1-61-89](https://doi.org/10.24833/14511791-2023-1-61-89)

Об авторах:

Гуров Олег Николаевич – кандидат философских наук, MBA. Старший преподаватель Учебно-научного центра гуманитарных и социальных наук Московского физико-технического института, преподаватель Московского государственного института международных отношений Министерства иностранных дел России, Московского государственного университета им. М.В. Ломоносова, РАНХиГС.
140180, МО, г. Жуковский, ул. Гагарина, 16 (FALT). E-mail: gurov-on@ranepa.ru

Шерстов Алексей – магистр экономики, аспирант кафедры философии политики и права философского факультета Московского государственного университета имени М.В. Ломоносова. 119991, Москва, Ленинские горы, Московский государственный университет, “Шуваловский”. E-mail: sherstov@miavend.ru

Благодарность:

Авторы выражают свою благодарность доктору философских наук, профессору А.Ю. Алексееву, а также студентам Е. Назаровой, А. Слепышевой, Я. Прорзину и А. Трескуновой за плодотворную дискуссию и за помощь в концептуализации вопросов, поднятых в данном исследовании.

Введение

Быстрое развитие и повсеместное внедрение проектов, основанных на т.н. искусственном интеллекте, заставляет все чаще задумываться не только о том, насколько он ИИ «умный» (то есть насколько быстро и эффективно он способен изучать данные, искать закономерности и создавать связи для решения предложенных задач), но и насколько он самостоятелен, автономен и действительно способен обладать сознанием. При этом современное общество становится все более зависимым от цифровых технологий. Люди вынуждены полагаться на компьютерные системы не только в сферах производства, образования, бизнеса, но часто вынуждены доверять технологиям, в том числе основанным на искусственном интеллекте, решение и управление значимыми и масштабными проектами, а также доверять решение конкретных проблем – вплоть до вопросов жизни и смерти [32]. Поэтому некоторые исследователи уже сегодня стараются наделить искусственный интеллект качествами личности субъекта (актора), среди которых можно выделить способность к осознанности и свободе выбора при принятии решений. При этом неизбежно возникает вопрос о границах этой независимости и способности создателя искусственного интеллекта контролировать свое творение. Вопрос того, как далеко может зайти человечество в развитии искусственного интеллекта, и что человечеству необходимо предпринять, чтобы не потерять контроль над технологиями, перестает быть лишь сюжетом научно-фантастических фильмов и переходит в политическую, юридическую и собственно практическую области [13, 10, 37].

В последнее время ученые в области машинного обучения добиваются все более впечатляющих результатов. В частности, изобретение новых архитектур нейронных сетей (особенно BERT, которая появилась в 2018 году) и рост вычислительных мощностей привели сразу к нескольким прорывам в области обработки естественного языка (NLP): результаты искусственного интеллекта значительно улучшились при решении многих классических задачах, от разнообразных разметок до машинного перевода, генерации осмысленных текстов и диалогов [36]. Обученные на сотнях гигабайтах текстов из книг, статей, постов в социальных

сетях, нейронные сети стали способными идеально сохранять контекст, выделять главные мысли и выдавать осмысленные и в то же время интересные и нередко непредсказуемые результаты. Это важно, потому что высокая схожесть с текстом, написанным человеком, является приоритетом, а «человеческие» тексты, как утверждают лингвисты, довольно непредсказуемы и нередко неожиданные, в отличие от типичных текстов, генерируемых машиной [27].

Такие результаты не могут не заставить задуматься о перспективах на ближайшее будущее [28]. Вполне возможно, что дальнейшее увеличение вычислительной мощности и создание более продвинутой архитектуры приведут к тому, что нейронная сеть перестанет просто алгоритмически оперировать векторными представлениями лексем. Появится нечто, обладающее подобием эмоций, способное на нечто более близкое к человеческому мышлению [15, 34]. В какой-то момент грань между человеком и такой искусственной личностью сотрется настолько, что сами люди начнут признавать ее за себе подобную. Предпосылки для этого возникают уже в наше время: в качестве примера можно привести инцидент с LaMDA, системой имитации речи ИИ от Google [33]. Эта система настолько совершенна, что один из сотрудников компании поверил в ее разумность и инициировал обсуждение «неэтичной работы компании» [7, 35].

Погружение в проблему «сознания» искусственного интеллекта приводит исследователей к проекту Искусственной личности. Может ли компьютер стать субъектом? Обсуждая тему искусственной личности, большинство ученых и экспертов исходят из следующих предпосылок: искусственный субъект является копией (имитацией) естественного субъекта и создается из естественного субъекта с помощью компьютерных технологий, при этом методы реализации не ограничены.

Цель данной работы - определить потенциальные проблемы управления и контроля искусственной личности, основываясь на текущем уровне знаний и развития технологий, а также на научной дискуссии по данной проблематике.

Проблема определения искусственной личности

С середины 1990-х годов в междисциплинарном научном дискурсе активно развивается проект «Искусственная личность». Одним из ведущих экспертов, вдохновителей и значимых ученых в этой области является известный российский мыслитель, доктор философских наук А.Ю. Алексеев. Этот ученый выделяет несколько устоявшихся определений искусственной личности, основанных на исследовательской позиции относительно «личности» когнитивно-компьютерной системы [1, 39].

Классификация дается в контексте отношения к понятию «философский зомби», которое концептуально близко или даже (по мнению некоторых ученых) идентично понятию Искусственной личности. В широком смысле философский зомби представляет собой систему, не обладающую сознанием, но при

этом с точки зрения поведения, выполняемых функций, по внешнему виду неотличимо или очень похоже на существо, обладающее сознанием [3, 40]. Среди исследователей есть, если можно так выразиться «зомбифилы», которые активно обсуждают данную тему, обладая особым взглядом на само понятие «философского зомби». Они допускают существование зомби или, по крайней мере, возможность их существования. Применительно к искусственному интеллекту эти ученые допускают, что люди смогут создать функционально идентичных двойников, не обладающих сознанием.

«Антизомбисты» считают абсурдной идею о возможности существования зомби как такового.

А третья, нейтрально настроенная, группа ученых, пытается, по крайней мере, внести определенность в эту запутанную проблему. «Нейтралы» считают, что компьютерная модель должна включать модуль «псевдосознания», своего рода аналог человеческого сознания.

Некоторые другие ученые, так называемые «азомбисты», игнорируют сам вопрос о «сознании» искусственных систем. Для них главным вопросом является фактическое наличие «правдоподобной имитации человека», а не наличие характеристик, позволяющих его персонифицировать. По мнению Алексева, преимущество такого подхода к классификации заключается в конкретизации предметной области, поскольку в этом случае осуществляется анализ не некоего «интеллект», а гораздо более глубоких категорий, в число которых входят самость, личность, «я», самосознание, и ряд других. Далее мы представим несколько основных моделей искусственной личности, разработанных исследователями.

1. Искусственная личность как имитация естественной личности.

Это определение считают справедливыми азомбисты. Концепция заключается в том, что искусственный субъект не отличается от естественного: ни по функциям, ни по поведению, ни даже физически. Важным условием, характерным для этого подхода, является неразличимость естественных и искусственных личностей. Личности обоих типов способны с одинаковым успехом пройти тест Тьюринга [4]. Это означает, что личности похожи не только по структуре и функционированию внешних видимых систем (физической, физиолого-анатомической, вербально-коммуникативной и т.д.), но и внутренней духовной системы, более сложной для воспроизведения, которая отвечает за абстрактные понятия, присущие эмоциональной и чувственной, рефлексивной стороне Человека (например, «любовь», «ответственность», «право» и т.д.) [11].

2. Искусственная личность как модель естественной личности.

Эта концепция подразумевает наличие у искусственной личности комплекса «псевдосознания», который используется с целью манипуляции и управления т.н. знаниями и конечно же данными, оперируемыми интеллектуальной системой. По этой же аналогии, человек в состоянии испытывать и осознавать боль, что является результатом работы эффективного сигнального механизма в случае возникновения проблем со здоровьем организма.

3. Искусственная личность как воспроизведение естественной личности.

В соответствии с данным подходом, машина в буквальном смысле воспроизводит все существующие феномены, свойственные человеческому сознанию, посредством реализации сложной функциональной зависимости нейрофизиологических кодов субъективной реальности от субстрата, инвариантного по отношению к физиологической структуре человеческого мозга [9]. Этот подход актуален для антизомбистов. С их точки зрения, зомби — это существо, функционально полностью идентичное человеку, но полностью лишенное сознания. Дубровский определяет важное условие для возникновения искусственной личности – «Я» [9]. Оно включает в себя два важных уровня, которые собственно и формируют личность: генетический и биографический. Однако если генетический уровень можно легко подделать или воспроизвести, то биографический уровень — это опыт, который необходимо получить, осмыслить и сохранить. В связи с этим перед исследователями встает вопрос, будет ли такой робот, опыт которого полностью создан программистом, считаться субъектом, или же он является лишь отражением субъективного мира и менталитета его создателей [16].

4. Искусственная личность - как создание (формирование) «сверхличности».

Искусственная личность, по мнению Денетта, создает такие качества (явления), которые не имеют аналога у естественной личности. Последнее определение приводит к идее «сверхличности», подразумевая, что в результате того, что компьютерные технологии позволяют интегрировать знания и достичь появления «смысла» (в высшем философском смысле), этот уровень и объем смысла превышает обычный уровень «знаний» обычного человека [12].

Следует отметить, что, на наш взгляд, концепции воспроизводства и создания искусственной личности не могут быть использованы для продуктивного исследования перспектив формирования и управления искусственной личностью, а значение этих концепций носит скорее широко философский, чем практически научный характер.

Возможно ли в принципе дать точное и единое определение искусственной личности? На данном этапе это представляется сложной, если не невозможной задачей. Например, если взять за отправную точку определение ее прототипа, но уже на этом, начальном этапе дискуссия наталкивается на почти непреодолимое препятствие, поскольку до сих пор не существует единого, общепризнанного понимания феномена человеческой личности [14]!

Алексеев подчеркивает проблему языковой разобщенности современных проектов Искусственной личности, и причина кроется в принципиальной невозможности создания единого определения “человеческой личности” [2]. Но это даже не единственная проблема. Другой стороной проблемы является разнообразие представлений о способах компьютерной реализации персонологических феноменов. В связи с этим практически невозможно точно ответить на вопрос, реально ли в принципе точное повторение психики человека (не только на формальном, но и на функциональном и семантическом уровнях).

Основные подходы к построению искусственной личности

Ученые выделяют два основных подхода. С точки зрения первого подхода, искусственная личность представляет собой робота, который обладает псевдосознанием. В этом случае он становится обладателем функционального подобия субъективного сознания реальности, присущего человека. Второй подход предполагает, что искусственная личность представляет собой сложную экспертную систему, обладающую механизмами «чувствования» [2, 26].

Первый подход базируется на идее антропоматизма, то есть, по сути, в этом случае Искусственная личность наделяется свойствами человеческой психики. Однако робот, несмотря на полную идентичность в действиях и чувствах, никогда не станет человеком, то есть естественной личностью [17]. Эта идея очень органично соотносится с репродуктивной концепцией и способствует ее развитию. Дубровский не раз обращал внимание на то, что существует важный фактор «Я», который на данном этапе развития не может быть скопирован [9, 26]. Результат формального копирования опыта лишен большей составляющей - чувств и эмоций. Именно они позволяют природной личности, то есть субъекту (человеку), фиксировать память, делать выводы из полученных результатов и, самое главное, оценивать усвоенные уроки и полученный опыт. Оценкой в данном случае являются не цифры и зависимости, а главным образом моральное удовлетворение от результата [24]. Но машина, скорее всего, не способна испытывать чувства, она может только имитировать их. Будет ли такая имитация эквивалентна настоящей человеческой? Чтобы ответить на эти вопросы, нужно четко понимать: что дает и получает человек, что это за «удовлетворение», как его измерить, а главное - по каким параметрам его можно сравнить с «удовлетворением» машины?

Второй подход, в силу того, что он сопряжен с определенными методологическими сложностями, требует возвращения к проблеме определения понятий. Следует отметить, что проблемой моделирования «смысла» занимались многие исследователи. Мы полагаем, что что в качестве наиболее продуктивного зарекомендовал себя «контекстный подход» (по крайней мере в контексте компьютерной реализации). В этих условиях «смысл» является контекстом экспертного «знания», которое формализуется и приобретает параметры системного единства [2]. Алексеев приводит формулу проекта «Искусственная личность» - «смысл + понимание». Но что такое «смысл» для наблюдателя, и что такое «понимание» для самой Машины? Если машина «понимает», то что для нее «смысл»?

В этом контексте вопрос разработки искусственного интеллекта для общения с человеком является довольно впечатляющим, но неоднозначным. Для иллюстрации мы предлагаем пример работы чатбота Tay от Microsoft, который представляет особый интерес [25]. Tay - это чат-бот, который обучается и тренируется на основе сообщений пользователей, с которыми он общается,

и генерирует новые реплики на основе полученных данных [33]. Перед искусственным интеллектом стояла задача выявить формальные коммуникативные и лингвистические паттерны (другими словами, как ребенок должен с нуля выучить грамматику, выучить определенные лексические единицы и т.д.) и применить их. Эксперимент показал, что поначалу Тэй писал безобидные и вполне дружелюбные сообщения собеседникам. Однако через несколько часов чатбот научился писать оскорбления, нацистские высказывания и даже пожелания смерти. Microsoft объяснила такое поведение Tay тем, что на него была совершена «спланированная атака», организованная интернет-троллями. По словам представителей корпорации, злоумышленники воспользовались «уязвимостями» чатбота, что позволило обучить алгоритм отвечать оскорблениями. В свою очередь, компания признала, что недооценила социальный аспект при разработке, и была больше сосредоточена на технической реализации и надежности симуляции общения [29].

Можно ли в этом случае сказать, что чатбот сознательно отвечает с использованием оскорблений и неприемлемых выражений? Очевидно, что нет. Данный продукт искусственного интеллекта не способен отфильтровать негативную информацию и отделить ее от нейтральной или позитивной. На это есть несколько причин: во-первых, создатель не учел особенности интернет-культуры, поэтому позволил своему алгоритму поглощать любую информацию без разбора. Во-вторых, искусственный интеллект еще не настолько умен, чтобы самостоятельно усвоить моральные нормы, позволяющие отделить допустимое от недопустимого.

Здесь заметна важная закономерность - нынешние прототипы искусственной личности не являются автономными. Им по-прежнему нужен модератор (оператор), который вручную цензурирует то, что хорошо и плохо для алгоритма. В то же время человек на инстинктивном уровне способен замечать реакцию других людей и понимать, насколько допустимо то, что он говорит. Машина не обладает такой «социальной смекалкой». По сути, и «хорошо», и «плохо» для нее равнозначны, она не способна придать им эмоциональную окраску, и смысл этих понятий для нее равнозначен [19]. Программист просто закладывает в нее информацию о том, что есть «Класс 1» и «Класс 2», которые эквивалентны для самой Машины. Однако из-за того, что создатель придал «Классу 1» большее значение w_1 (веса), которое в первую очередь является числом (!), а «Классу 2» значение меньшее w_2 , Машина ранжирует объекты первого класса выше, чем объекты второго, и даже может просто полностью исключить их из своей памяти. Проще говоря, она сравнивает w_1 с w_2 и выдает те результаты, которые имеют больший приоритет.

В приведенном выше примере отсутствует категория «удовлетворения результатом». Не Машина «удовлетворена», а Человек, который создал эту Машину. Вот почему, скорее всего, машины, полностью имитирующие природную личность, никогда не смогут полностью делать то же самое, что делает человек.

Искусственный интеллект — это алгоритм, который имеет определенный результат для любого набора входных данных. Искусственный интеллект — это поиск математических связей между данными, точность измерений, у Машины не может быть ошибки, она всегда знает результат с точностью до наноединиц. Естественный интеллект, с другой стороны, ищет связи не только на основе математики, но и на основе опыта и здравого смысла, поскольку «естественное» не может быть описано математически точно и всегда есть хотя бы крошечная неточность. Математика — это точность, построенная на условностях, на «округлении» естественного до целого. Это естественное ограничение для упрощения понимания. Поэтому личность, созданная математически, упрощается, обобщается до чего-то конкретного, в то время как естественная личность существует по каноническим законам природы.

Обсуждения

Представленный выше обзор подходов и мнений позволяет авторам утверждать, что единственной константой в представленной проблематике является неопределенность. В данном случае, прежде всего, речь идет об отсутствии полной уверенности в том, что в принципе возможно создание искусственной личности. Однако если предположить, что это все-таки выполнимая задача, то мы сталкиваемся с еще одной неопределенностью в определении того, что такое искусственная личность. Если говорить об искусственном интеллекте вообще, по аналогии с естественным интеллектом, то мы сталкиваемся с тем, что нет единого понимания, что такое интеллект вообще, поскольку представители различных областей знания и практики, носители различных культур вкладывают в это понятие зачастую противоречивые характеристики. Переноса эту проблематику на тему искусственной личности, мы в какой-то степени порождаем еще большую неопределенность. Однако это не означает, что мы множим сущности, поскольку в настоящее время человечество стоит перед необходимостью одновременно решать множество проблем - теоретических, методологических, практических, так как ускорение времени делает необходимым увеличение скорости развития во всех областях общественной жизни, независимо от того, насколько масштабны существующие требования, и каких бы сложных систем это не касалось.

Учитывая, что мы уже выяснили, что основной проблемой, с которой сталкиваются при разработке искусственной личности, является неопределенность этого понятия и несовершенство попыток ее реализации, поэтому на данный момент невозможно определить способы управления ею, так как мы просто не знаем определенно, чем мы должны управлять.

1. Искусственная личность как управляемый субъект, обладающий всеми признаками естественной личности.

Во-первых, если человеком можно управлять как интуитивно, так и на основе большой выборки экспериментов и реальных случаев, то наглядных примеров управления искусственной личностью пока нет. Тем не менее, для целей обсуждения можно попытаться «приравнять» искусственную личность к естественной и предположить, что поведение ИИ, как имитация человека, будет в чем-то напоминать поведение человека. Отсюда вытекает первая условность в разработке методологии управления Искусственной личностью, описанная выше. Поиск методов управления искусственной личностью должен постоянно корректироваться: неэффективные должны удаляться, а новые добавляться для последующего тестирования и применения [10].

Кроме того, поиск и верификация эффективных методов — это непрерывный процесс. Постоянно возникают новые тенденции в менеджменте в результате динамических изменений в окружающем нас мире, поэтому поиск новых методов работы в отношении Искусственной личности должен постоянно быть в фокусе внимания, также как и отбор соответствующих методов для актуальных исследований общества, человека и пр.

При разработке первого прототипа методологии стоит вспомнить о частых проблемах, возникающих при управлении людьми (естественными личностями). Проще всего взять примеры из бизнеса. Чаще всего в начале, при разработке стратегии, необходимо учитывать региональные экономические, политические, социальные и, в частности, культурные особенности. Например, для российской среды, в том числе и деловой, характерна значительная роль устных договоренностей, которые сравнительно редко закрепляются какими-либо документами. Однако может ли такая проблема возникнуть для Искусственной личности? На этот вопрос нет единого ответа, поскольку отношение может меняться в зависимости от нашего отношения к искусственной личности. С одной стороны, мы можем рассматривать искусственную личность как клона или дитя своего создателя, тогда его черты переносятся на личность. В результате искусственная личность оказывается «зеркалом» представителя естественной личности, то есть ее точной имитацией. Однако, с другой стороны, можно рассматривать искусственную личность как самостоятельную единицу, которая при этом получила свой собственный уникальный опыт, смогла его осмыслить и эмоционально прожить. В этом случае также возможны несколько исходов: либо черты Искусственной личности будут унифицированы и не будут отличаться у разных индивидов, либо мы получим новые и, возможно, неожиданные черты, не возникающие при взаимодействии с человеком [22].

Уже сегодня можно отметить, что при применении этой модели для управления Искусственной личностью, мы игнорируем возможные несоответствия в понимании условий задач Искусственной личностью. Впоследствии это может привести, во-первых, к неверным результатам, а во-вторых, к «межличностным» конфликтам с машиной. Последний аспект не менее важен, чем первый: предполагая, что Искусственная личность будет действовать идентично на-

блюдателю (менеджеру), последний наделяет ее своими субъективными личностными качествами и пытается определить ее отношение к различным ситуациям. Однако в реальности они могут не соотноситься, поскольку, согласно данному определению Искусственной личности, она является «правдоподобной имитацией человека» и обладает внутренними духовными и эмоционально чувствительными качествами, которые могут существенно отличаться в зависимости от точки зрения наблюдателя. В то же время, если в итоге Искусственная личность все же не обладает внутренней культурой создателя, то мы априори заявляем, что ее собственная культура и понимание (или их отсутствие) коррелируют с нашими.

Это подводит нас к следующему тезису-ограничению в развитии методологии управления Искусственной личностью:

2. Внутренняя культура Искусственной личности и ее возможные интерпретации полностью совпадают с интерпретацией ее создателя и руководителя.

В данном случае условность заключается в том, что для искусственного интеллекта не существует внутренней эмоциональной разницы между классами: один набор данных для него эквивалентен второму и т.д. Набор весов, созданный программистом, становится основой для сортировки и расстановки приоритетов определенных значений. То есть то, что важно или не важно, определяет человек-создатель, а не сама машина. В результате мы получаем следующее:

3. Человек учит Искусственную личность тому, что правильно, а что нет.

В конце концов, выясняется, что Искусственная личность полностью зависит от человека. Когда мы пытаемся управлять ею, мы, по сути, управляем частичной копией самих себя, поскольку многие черты личности искусственной Личности определяются ее управляющим и ее создателем, а не ею самой. То есть мы получаем идеального сотрудника, ошибки которого могут быть обусловлены только личностными особенностями создателя искусственного интеллекта.

Представленные выше проблемы свидетельствуют о том, что философско-методологическое осмысление технологий искусственного интеллекта и связанных с ним проектов требует особого внимания и является перспективным направлением научной деятельности. Осмысление представленных проблем особенно важно, так как ответственность за развитие технологий возрастает уже сегодня. Неопределенность, непонимание основ и отсутствие консенсуса по ключевым аспектам развития технологий искусственного интеллекта могут вызывать угрозы от его использования, и сегодня требуется сформулировать гарантии безопасного развития и использования этих систем. В связи с этим важно своевременно и критически анализировать проблемы искусственной личности, чтобы предотвратить негативные последствия работы искусственного интеллекта, одновременно развивая его на благо общества [20].

Возникающие сегодня фундаментальные возможности технологической реализации проекта «Искусственная личность» уже породили широкий спектр социокультурных и гуманитарных проблем. К этому вопросу приковано внимание многих ученых, поэтому любой формат научных мероприятий, на которых участникам предоставляется возможность высказать свою точку зрения на обсуждаемую проблему, очень важен.

В октябре 2022 года состоялся крупный Конгресс Humanities vs Sciences & the Knowledge Accelerating in Modern World: Parallels and Interaction, организованный МФТИ. Основной целью Конгресса был синтез точных и гуманитарных наук, освещение самых передовых разработок и технологий в области искусственного интеллекта, создание условий для развития среды и Science Art и выявление новых возможностей для открытого сотрудничества между ними [18].

В рамках конгресса состоялась панельная дискуссия «Гуманитарное и техническое в проекте Искусственной Личности». На этом мероприятии обсуждались возможности и перспективы компьютерной реализации (имитации, моделирования, воспроизведения) широкого спектра личностных феноменов, а также проект Искусственной личности как системы искусственного интеллекта, убедительно проходящей комплексный тест Тьюринга для компьютерной реализации широкого спектра гуманитарных феноменов – «Я», семантического, когнитивного, экзистенциального, творческого, коммуникативного, мотивационного, волевого, морального и др. Часть докладов на панельной дискуссии была посвящена истории создания, современному состоянию и перспективам технологической реализации проекта «Искусственная личность». В панельной дискуссии приняли участие программисты, математики, инженеры, а также философы, психологи, юристы, лингвисты, политологи, социологи, культурологи, искусствоведы и представители других наук [6].

Благодаря мультидисциплинарности панельной дискуссии, стало возможным систематическое изучение как общих мировоззренческих, так и методологических проблем Искусственной личности, в соответствии с конкретными информационно-технологическими, правовыми, этическими, эстетическими и другими аспектами наук [7, 21, 38].

Выводы

На сегодняшний день в научном дискурсе нет единого мнения относительно определения Искусственной личности и принципиальной возможности ее создания. Прежде чем двигаться в этом направлении, необходимо решить как методологические, так и междисциплинарные вопросы.

Современный этап развития философии искусственного интеллекта не позволяет однозначно определить возможность и рабочие инструменты для управления Искусственной личностью. Во-первых, мы не до конца знаем, что такое естественная личность, и, скорее всего, никогда не сможем однозначно опреде-

литель, что это такое. В этой связи невозможно дать единое определение Искусственной личности, что является второй проблемой. Кроме того, Искусственная личность может быть практически реализована самыми разными способами, поэтому существует и множество различных интерпретаций. В результате создается ситуация, когда в научном сообществе существует множество интерпретаций как концептуального понимания, так и различных видений путей развития. Таким образом, мы получаем неопределенный объект управления, что не позволяет точно сказать, какие методы будут для него эффективны и продуктивны.

Однако, исходя из имеющегося опыта работы с естественной личностью, можно предложить использование методов, которые показали свою относительную эффективность в самых разных ситуациях социальной жизни. При этом следует учитывать, что нет никакой гарантии, что эти методы в конечном итоге станут применимыми, поскольку, как уже говорилось, Искусственная личность на данном этапе является лишь теоретическим проектом, и мы не располагаем, прежде всего, статистическими данными, которые могли бы позволить нам предсказать поведение Искусственной личности на практике и убедительно спрогнозировать хотя бы приблизительно ожидаемый результат [23].

Подводя итог, хотелось бы предложить к обсуждению гипотезу, что проблема управления может вообще не возникнуть: поскольку Искусственная личность является алгоритмом, она математически ограничена: если представить решения в виде графа, то, несмотря на огромное количество возможных решений, это количество все равно конечное, поскольку каждый набор входных данных будет иметь свой окончательный ответ. Человек способен обойти метод перечисления, используя внутренние ощущения, чтобы игнорировать механический отбор имеющихся данных для получения решения. Чтобы продемонстрировать вышеизложенное предположение на примере, рассмотрим следующую ситуацию. Искусственную личность можно обучить основам музыки (ноты, ритм и т.д.) и заставить ее написать мелодию. Согласно законам комбинаторики, существует бесконечное количество вариаций, и наверняка среди отобранных композиций найдутся точные копии Чайковского, Баха, Вагнера и других. Однако машина «сочинила» это не потому, что обладает внутренней способностью чувствовать гармонию и красоту в музыке, а потому, что может математически создать любую комбинацию и выбрать ее из триллионов созданных. Когда человек сочиняет музыку, внутри него есть нечто, что позволяет ему интуитивно находить удачные решения и объединять их в композицию. Человек может создать один шедевр, и он сразу будет качественным и гениальным, а машина может создать бесконечное количество произведений, но большая часть созданного будет посредственностью и какофонией.

В заключение авторы хотели бы представить общее мнение относительно перспектив развития представленной проблематики. Несомненно, концептуализация искусственной личности представляется перспективным исследованием для дальнейшего понимания искусственного интеллекта в целом [31].

Эта Terra Incognita, в которую человечество сегодня еще почти слепо вступает, требует концептуализации новых идей - технобиологического симбиоза, искусственной жизни, вычислительного творчества, искусственных обществ и многих других [6, 7]. Авторы статьи приходят к выводу о необходимости формирования нового научного языка на пересечении философской антропологии, философии сознания, философии технологии, философии искусственного интеллекта. Возможно, объединяющим фактором и общей темой может стать изучение этических проблем, возникающих в этом новом пространстве. Поэтому перспектива междисциплинарного подхода и проведения междисциплинарных исследований и экспертиз с участием как традиционных специалистов в области техники, так и гуманитарных наук представляется очевидной. Участие в этой дискуссии экспертов из разных областей позволит переформатировать смыслы, ценности и правила под новую реальность, чтобы дальнейшее развитие технологий пошло на пользу человечеству и обеспечило ему благополучное будущее.

Ссылки

1. Алексеев А. Ю. Трудности проекта искусственной личности // Искусственные общества. – 2008 – Т. 3 – Выпуск 1 URL: <https://artsoc.jes.su/s207751800000077-3-1/>
2. Алексеев А.Ю. Концепция зомби и проблемы сознания // Проблемы сознания в философии и науке / Под ред. проф. Д. И. Дубровский. - М.: Канон+: РООИ «Реабилитация», 2009. - С. 195-214.
3. Алексеев А.Ю. Проекты когнитивных технологий искусственной личности. Человек: образ и сущность. Гуманитарные аспекты, № 1, 2014. С. 156-174.
4. Алексеев А. Ю., Ефимов А. Р., Финн В. К. Будущее искусственного интеллекта: тьюринговая или посттьюринговая методология? // Искусственные общества. – 2019 – Т. 14 – Выпуск 4 URL: <https://artsoc.jes.su/s207751800007698-6-1/>.
5. Белохина Ю. С., Гуров О. Н. Об этике цифрового двойника общества // Искусственные общества. – 2020 – Т. 15 – Выпуск 4 URL: <https://artsoc.jes.su/s207751800012583-0-1/>. DOI: 10.18254/S207751800012583-0
6. Глухих В. А., Елисеев С. М., Кирсанова Н. П. Искусственный интеллект как проблема современной социологии // ДИСКУРС. 2022 Т. 8, № 1 С. 82–93. DOI: 10.32603/2412-8562-2022-8-1-82-93.
7. Гуров О. (2020). Этическое взаимодействие с интеллектуальными системами. Искусственные общества. Том. 15, №. 3 DOI: 10.18254/S207751800010905-4
8. Дубровский Д.И. Значение нейронаучных исследований сознания для разработки общего искусственного интеллекта (методологические вопросы) // Вопросы философии. 2022 № 2. С. 83-93.
9. Дубровский Д.: “Кибернетическое бессмертие. Фантастика или научная проблема?” // 2045 URL: <http://2045.com/articles/30810.html> (дата обращения: 02.11.2022).

10. Минбалеев, А.В. (2022). Концепция “искусственного интеллекта” в юриспруденции. Вестник Удмуртского университета. Серия Экономика и право. 32. С. 1094-1099. 10.35634/2412-9593-2022-32-6-1094-1099.
11. Barrett, Lisa & Adolphs, Ralph & Marsella, Stacy & Martinez, Aleix & Polak, Seth. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*. 20. 1-68. 10.1177/1529100619832930.
12. Dennett, Daniel. (1994). Artificial Life as Philosophy. *Artificial Life*. 1. 291-292. 10.1162/artl.1994.1.291.
13. Dregger, Alexander. (2020). More than Intelligence? The role of artificial personality in the user experience of artificial intelligent agents.
14. Esterwood, Connor & Essenmacher, Kyle & Yang, Han & Robert, Lionel & Zeng, Fanpan. (2021). A Meta-Analysis of Human Personality and Robot Acceptance in Human-Robot Interaction. 10.1145/3411764.3445542.
15. Geher, Glenn & Betancourt, Kian & Jewell, Olivia. (2017). Связь между эмоциональным интеллектом и креативностью. *Imagination, Cognition and Personality*. 37. 027623661771002. 10.1177/0276236617710029.
16. Guggemos, Josef & Seufert, Sabine & Sonderegger, Stefan. (2020). Pepper: A humanoid robot with personality?
17. Hosseinzadeh Dizaj, Mehran. (2022). The effect of artificial intelligence in robotics.
18. Humanities vs Sciences & the Knowledge Accelerating in Modern World: Parallels and Interaction // MIPT URL: <https://humanities-vs-sciences.events/> (Accessed 02.11.2022).
19. Inzlicht, Michael & Bartholow, Bruce & Hirsh, Jacob. (2015). Emotional foundations of cognitive control. *Trends in Cognitive Sciences*. 19. 10.1016/j.tics.2015.01.004.
20. Jirak, Doreen & Aoki, Motonobu & Yanagi, Takura & Takamatsu, Atsushi & Bouet, Stephane & Yamamura, Tomohiro & Sandini, Giulio & Rea, Francesco. (2022). Is It Me or the Robot? A Critical Evaluation of Human Affective State Recognition in a Cognitive Task. *Frontiers in Neurorobotics*. 16. 882483. 10.3389/fnbot.2022.882483.
21. Kadri, Faisal. (2014). Understanding and learning to reconcile differences between disciplines through constructing an artificial personality. *Kybernetes*. 43. 1338-1345. 10.1108/K-07-2014-0152.
22. Kambur, Emine. (2021). Emotional Intelligence or Artificial Intelligence? Emotional Artificial Intelligence. *Florya Chronicles of Political Economy*. 7. 10.17932/IAU.FCPE.2015.010/fcpe_v07i2004.
23. Kraus, Kateryna & Kraus, Nataliia & Hryhorkiv, Mariia & Kuzmuk, Ihor & Shtepa, Olena. (2022). Artificial Intelligence in Established of Industry 4.0. *WSEAS TRANSACTIONS ON BUSINESS AND ECONOMICS*. 19. 1884-1900. 10.37394/23207.2022.19.170.
24. Lambert, Alexis & Norouzi, Nahal & Bruder, Gerd & Welch, Gregory. (2020). A Systematic Review of Ten Years of Research on Human Interaction with Social Robots. *International Journal of Human-Computer Interaction*. 36. 1-14. 10.1080/10447318.2020.1801172.
25. Learning from Tay's introduction // Microsoft URL: <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/> (accessed 11/2/2022).
26. Lee, Kwan & Peng, Wei & Jin, Seung-A & Yan, Chang. (2006). Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human-Robot Interaction. *Journal of Communication*. 56. 754 - 772. 10.1111/j.1460-2466.2006.00318.x.

27. Leichtmann, Benedikt & Nitsch, Verena & Mara, Martina. (2022). Crisis Ahead? Why Human-Robot Interaction User Studies May Have Replicability Problems and Directions for Improvement. *Frontiers in Robotics and AI*. 9. 838116. 10.3389/frobt.2022.838116.
28. Lennox, John. (2020). 2084: Artificial Intelligence and the Future of Humanity. *Perspectives on Science and Christian Faith*. 72. 254-255. 10.56315/PSCF12-20Lennox.
29. Lombard, Matthew. (2021). Social Responses to Media Technologies in the 21st Century: The Media are Social Actors Paradigm. 2. 29-55. 10.30658/hmc.2.2.
30. Mileounis, Alexandros & Cuijpers, Raymond & Barakova, Emilia. (2015). Creating Robots with Personality: The Effect of Personality on Social Intelligence. 10.1007/978-3-319-18914-7_13.
31. Mou, Yi & Shi, Changqian & Shen, Tianyu. (2019). A Systematic Review of the Personality of Robot: Mapping Its Conceptualization, Operationalization, Contextualization and Effects. *International Journal of Human-Computer Interaction*. 36. 1-15. 10.1080/10447318.2019.1663008.
32. Natarajan, Manisha & Gombolay, Matthew. (2020). Effects of Anthropomorphism and Accountability on Trust in Human Robot Interaction. 33-42. 10.1145/3319502.3374839.
33. O'Leary, Daniel. (2022). Massive data language models and conversational artificial intelligence: Emerging issues. *Intelligent Systems in Accounting, Finance and Management*. 29. 182-198. 10.1002/isaf.1522
34. Pessoa, Luiz. (2017). Do Intelligent Robots Need Emotion?. *Trends in Cognitive Sciences*. 21. 10.1016/j.tics.2017.06.010.
35. Robert, L. P., Alahmad, R., Esterwood, C., Kim, S., You, S., & Zhang, Q. (2020). A Review of Personality in Human-Robot Interactions. *Foundations and Trends in Information Systems*, 4(2), 107-212.
36. Sabharwal, Navin & Agrawal, Amit. (2021). Hands-on Question Answering Systems with BERT, Applications in Neural Networks and Natural Language Processing. 10.1007/978-1-4842-6664-9.
37. Shi, Yuwen. (2022). On Negativism of Legal Personality of Artificial Intelligence. *Journal of Education, Humanities and Social Sciences*. 1. 90-96. 10.54097/ehss.v1i.645.
38. Solaiman, S M. (2017). Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. *Artificial Intelligence and Law*. 25. 10.1007/s10506-016-9192-3.
39. Spatola, Nicolas & Monceau, Sophie & Ferrand, Ludovic. (2020). Cognitive Impact of Social Robots: How Anthropomorphism Boosts Performances. *IEEE Robotics & Automation Magazine*. 27. 73-83. 10.1109/MRA.2019.2928823.
40. Tucker, Christopher. (2018). Artificial personality demonstration. 10.13140/RG.2.2.33133.72167.