Философия науки и техники
2024. Т. 29. № 2. С. 125–135
УДК 004.89

Philosophy of Science and Technology
2024, vol. 29, no. 2, pp. 125–135
DOI: 10.21146/2413-9084-2024-29-2-125-135

ИННОВАЦИОННАЯ СЛОЖНОСТЬ

Т.Г. Лешкевич

Проблема субъектности нейросетей: humans и non-humans

Лешкевич Татьяна Геннадьевна – доктор философских наук, профессор. Южный федеральный университет. Российская Федерация, 344006, г. Ростов-на-Дону, ул. Большая Садовая, д. 105/42; e-mail: Leshkevicht@mail.ru

В статье обсуждается актуальная тематика относительно того, насколько качество субъектности может быть передано нейросети искусственного интеллекта (ИИ). Основная проблема связана с анализом эффектов двойственной трансформации субъектности, обусловленных, с одной стороны, сращенностью и гипервзаимосвязанностью современного индивида и цифрового мира, с другой – функционированием искусственных нейросетей, демонстрирующих свое автономное стремление представлять субъектность человека. В поле зрения – комплекс взаимосвязанных аспектов. Во-первых, рассматриваются основополагающие характеристики традиционно понимаемой субъектности с акцентом на преобладание рефлексивности и интенциональности субъективного опыта. Во-вторых, с опорой на современную литературу раскрывается понятие искусственных нейросетей (ИНС), выявляются их специфические особенности, структура и разновидности. Взвешиваются аргумента «за и против» гипотетического признания субъектности нейросетей ИИ, обращается внимание на феномен информационного глобального рабочего пространства. В-третьих, выявляется установка нейросетей на сверхисполнительность и ее место в анализе галлюцинаций ИНС. В-четвертых, оценивается значение промпта (запроса) и развитие интерактивного ИИ, который может выполнять роль не только заказчика, но и руководителя и контролера. Делается вывод, что современное состояние субъектности приобретает качества актанта, вынужденного реализовать неотделимую от него последовательность функций, выполнение которых необходимо для осуществления жизни в Сети.

Ключевые слова: субъектность, нейросети ИИ, глобальное рабочее пространство, промпт, галлюцинации ИИ, интерактивный ИИ, актант

Вопрос о субъектности нейросетей искусственного интеллекта (ИИ) остро дискуссионный, который часто приобретает форму: возможно ли существование сознания отдельно от его носителя? Любопытно, что на Давосском форуме 2024 г., в фокусе внимания которого оказалось в том числе и обсуждение будущего ИИ, затрагивалась проблема «В чем заключается основная компетенция человека?». Может ли ИИ обладать возможностями воспроизводить «основную человечность», «эмоциональный интеллект» и «сочувствие» [AGI: секретный проект, web], т.е. выступать, представляя субъектность человека? Трансформации субъектности цифровой эры интересовали исследователей и ранее, однако сейчас проблема состоит в анализе двоякого преломления конфигураций субъектности. С одной стороны, это изменения дигитальной субъектности, которая, будучи сращенной (гипервзаимосвязанной) с цифровыми инструментами и ресурсами Сетей, приобретает новые качества. С другой стороны, это учет существенных достижений совершенствования нейросетей ИИ, демонстрирующих автономность, суверенную систему действий, способность выбора в альтернативных ситуациях и свободу функционирования при отсутствии прямой команды от субъекта-пользователя, что свидетельствует о параметрах субъектности.

Подчеркнем, что в качестве основополагающих характеристик традиционно понимаемой субъектности, как правило, называются сознание и самосознание, эмоции и воля, ценностно-целевые стратегии и действия, а также рефлексивная функция. В качестве симптомов второго ряда указывается на минимальную телесную активность и моторный контроль, адекватный отклик на внешние стимулы, наличие и возможность переключения внимания, фиксация объективных проявлений (свидетельств) сознания другого, рассуждение, память. Важны также восприятие своего субъективного состояния (квалиа), нарративные коммуникативные практики, самоописание и словесный отчет. Однако как вся эта совокупность функций может быть передана нейросети ИИ? Как осуществляется связь нейронной сети с ментальным миром, его образами и качеством?

Согласно выводам К.В. Анохина, субъектность вбирает в себя субъективный опыт, который «располагает характеристиками качества, значения, ценности, цели и интенциональности, не имеющими прямой логической связи с нейробиологическим описанием мозга как органической системы» [Анохин, 2023, с. 5]. Причем при установлении факта сознания как неотъемлемого элемента субъектности должны иметь место отчеты об интерсубъективном опыте и сознательных переживаниях от первого лица. Вместе с тем сканирующих приборов, которые бы верифицировали квалитативность данных интерсубъективных отчетов, еще нет. Заметим, что для Д.А. Поспелова – одного из первопроходцев в области ИИ – наличие и способность «к рефлексии, позволяющей разумному существу ставить себя на место другого существа и прогнозировать возможные решения этого другого», были основополагающими [Поспелов, 2003].

В отношении искусственных нейросетей следует заметить, что они пытаются воспроизвести структуру человеческого мозга, но также поддерживаются программным обеспечением, работа которого должна быть аналогична функциональным задачам, выполняемым нейронами человеческого мозга.

В литературе отмечено, что «нейронные сети (НС) представляют собой как входной, так и выходной слой, а также скрытый слой, содержащий единицы измерения, которые изменяют входные данные на выходные, чтобы выходной слой мог использовать значение» [Катиева, 2021]. Ставка делается на способность извлекать значимые данные из неточных необработанных данных. Иными словами, с нейросетями связывают процесс, позволяющий ИИ не только учиться на опыте, который опирается на данные, полученные в процессе обучения, но и корректировать его. Действительно, нейросети – это не алгоритмизированные программы, работающие по принципу «если… то…», они способны к самообучению, демонстрируют своеобразную автономию. Искусственные нейросети (ИНС) проявляют гибкость, ассоциативность, не нуждаясь в разработке специального алгоритма под конкретную задачу. Различают *многослойные* НС, которые обрабатывают числовые данные, *сверточные* НС, работающие с изображениями, и *рекуррентные* НС, не только собирающие и обрабатывающие информацию, но и отправляющие сами на себя скрытый слой своих же значений на следующем шаге. Можно предположить, что это отдаленно напоминает процесс рефлексии, в частности предваряющее его рефлексивное челночное движение, связанное с оборачиванием на себя.

Когда ведется речь о гипотетическом признании субъектности нейросетей ИИ, это означает признание за ними суверенной системы действий, свободы от вложенных в них моделей, основанных на Больших данных. Отметим, что в контексте не потерявшего свою силу информационного подхода сознание рассматривается как исполнительный, координирующий и контролирующий орган, осуществляющий передачу информации между специализированными отделами. Если обратиться к выводам нейроученого Дж. Тонони, то в ответе на вопрос, является ли какая-либо система (например, мозг) сознающей или нет, утверждается, что система становится сознающей в тот момент, когда в ней больше интеграции информации, чем в любой из ее частей [Сознание, web]. Подобные взгляды отражены теорией «глобального рабочего пространства» (ГРП), выдвинутой Б. Баарсом [Baars, 2002]. Уточнения, предложенные французским нейрологом С. Деаном, переносят акцент на то, что информация в мозге становится сознательной, когда она, находясь в ГРП, доступна для использования многими и различными системами по всему мозгу для широкого спектра задач [Сознание, web]. Деан приходит к заключению, что «сознание – это трансляция единого информационного потока в коре головного мозга: основой этого процесса является нейронная сеть, смысл существования которой сводится к активной передаче актуальной информации в пределах мозга» [Деан, 2018, с. 22]. Возникает вопрос: а что является актуальной информацией? Если для человека она в конечном счете коррелируется потребностями, то в отношении ИНС логично предположить, что актуальность информации связана с запросом, т.е. с формулировкой промпта.

Таким образом, возникает предположение, что ГРП – это область объединения информации, поступающей из многочисленных разрозненных источников, но в нем могут иметь место многочисленные ансамблевые сцепления и, судя по всему, информационные симуляции. Если опираться на анализ М.А. Сущина, становится понятно, что глобальное рабочее пространство

поддерживается сетью тесно связанных пирамидальных нейронов, соединяющих префронтальные и теменные регионы мозга. Информация, ставшая достоянием ГРП, транслируется и поддерживается возвратной активностью части нейронов. В то же время возможность проявления каких-либо других конфигураций ГРП, связанных с другими потенциально осознаваемыми стимулами, подавляется. В итоге кодирование информации в ГРП делает возможными высокоуровневые когнитивные функции: планирование, вербальный отчет, сознаваемую рабочую память, волевой контроль действий и т.д. [Сущин, 2020, с. 49–50].

Галлюцинации нейросетей ИИ

Однако проблемы возникают уже тогда, когда ИНС генерирует «нежелательные» выходные данные. Ее самопроизвольное поведение и непредсказуемость становятся внешне наблюдаемыми признаками галлюцинаций. К галлюцинациям относят неконтролируемые результаты (что приводит к удалению контента), утрату причинно-следственных связей, поверхностные оценочные суждения, а также несправедливые решения, дифференцирующие людей по национальному, половому, расовому признакам и цифровым навыкам. Возникает потребность в устранении ошибок, улучшении результатов, в их экспертной оценке. Здесь запрос на субъектность возникает в новом виде – субъектность как арбитр действия ИНС.

Метафора, относящаяся к функционированию ИНС, сравнивает их с «черным ящиком» и обозначена словосочетанием “Black Box Problem”, суть которого – в «выходе из-под контроля». Нейроны, обремененные терабайтами данных, наделяются весом (что приравнивается к синопсисам в естественной нейросети), трансформируются в цифры, которые затем генерируются нейросетью в токены (символы), в последовательности – токен за токеном. Путь в обратном направлении (т.е. восстановление движения от токенов к нейронам) проделать вряд ли возможно. Процесс оценивается как необратимый. В средних слоях ИНС происходит образование новой информации, однако этот процесс «непрозрачен»: остается неясным, в какую сторону «пробрасываются мостки» между совокупностью примеров сохраненной информации. Что имеет значение: повторяющиеся явления, их устойчивые корреляции, моделирование скрытых связей? Как соединяется информация разных слоев и разных уровней? Какой тип регулярности между входами и выходами действительно имеет место? Ведь, по мнению исследователей, в то время как в некоторых случаях может присутствовать простая статистическая корреляция, в других она может относиться к добросовестной причинной закономерности [Zednik, 2021]. Понятно, что ИНС ориентируется на содержащиеся в базе данные и их перебор, отвечающий формулировке запроса (промпта). Однако если запрос переиначить, то и поиски ответа пойдут в другом направлении, обнаруживая иной результат. Более того, возможен сценарий, связанный с «генерированием “множественного рождения” квазиреальных событий» [Лешкевич, 2022, с. 35].

Многократное уточнение запроса с целью корректировки ответов нейросетей выявляет их парадоксальное свойство, заключающееся в лимите

кратковременной памяти. Иными словами, после того, как текст переработан и превращен в токены, при каждом новом сообщении для продолжения диалога нейросеть должна «вспомнить», т.е., как отмечают исследователи, «прогнать через себя всю историю переписки. Но если вы вышли за пределы лимита токенов, то ей не остается ничего, кроме как удалять самые старые сообщения. А именно там, скорее всего, было описание исходного запроса. То есть через какое-то время разговора нейросеть попросту забывает, с чего все начиналось» [Казаков, web]. С точки зрения человеческого опыта вряд ли подобную ситуацию можно оценить как приемлемую. Такой эффект диалога, если бы он проходил между людьми, повлек бы за собой признание нарушений памяти, препятствующих полноценному общению. Поэтому лимит краткосрочной памяти в функционировании ИНС трактуется как существенная проблема, для преодоления которой необходимо появление таких нейросетей, которые будут в состоянии обрабатывать миллионы токенов.

Таким образом, к причинам галлюцинаций нейросетей может быть отнесена сложность и так называемая прожорливость нейросетей, которым необходимы огромные объемы данных. Более того, сам процесс их обучения темпорально затратен и требует колоссальных мощностей и ресурсов. Существует даже мнение, что всю нашу эпоху будут именовать эпохой обучения ИИ [Хоперский, web]. Уже в исследованиях 2017 г. описывались результаты, целью достижения которых было обучить ботов различным диалогам с учетом факторов переговоров, компромисса, согласования условий, иными словами, максимально приблизить их к специфике человеческой коммуникации через создание НС [Lewis et al., 2017]. После обучения НС в ходе экспериментов некоторые люди не понимали, что ведут диалог с ботом, а не с реальным человеком.

Вместе с тем обратим внимание на особую установку НС, связанную с их сверхисполнительностью, т.е. необходимостью дать ответ, независимо от того, насколько правильным он будет с точки зрения человеческого опыта. Эта установка занимает существенное место в анализе галлюцинаций ИНС. В этой связи уместно привести мнение Сэма Альтмана, генерального директора OpenAI, согласно которому «люди довольно снисходительны к ошибкам других людей, но совсем не терпимы к ошибкам компьютеров» [AGI: секретный проект, web]. Существуют суждения, указывающие на то, что делать ошибки – это очень по-человечески. Ложь, притворство, заблуждения, несуразица – частые признаки человеческого поведения. В качестве аргумента используется и то, что человек не контролирует всю подсознательную деятельность, ибо сознание, образно выражаясь, предстает как верхняя, надводная часть айсберга, 95% которого находится в зоне бессознательного и подсознательного. Поэтому индивид не может влиять на проистекающие там бессознательные нейрофизиологические процессы, планировать так называемую «архитектуру решений». Невозможно перевести в «цифру» биологическую и психофизиологическую «элементную базу» когнитивных функций. И какие бы аргументы, объясняющие и оправдывающие галлюцинации ИНС, ни привлекались, происходящие сбои несут собой серьезные риски в процессе функционирования и использования ИИ. Но если не стоит верить всему, что выдает нейросеть, или, к примеру, будет принята тактика вознаграждения нейросети за полученные

с первого раза правильные ответы, то кто возьмет на себя эти функции, кто разберется в том, что правильно, а что нет? А следовательно, человеческий контроль и координация необходимы.

Востребованность промпт-инженера и развитие интерактивного ИИ

На сегодня самой востребованной профессией, помимо специалистов по кибербезопасности, сетевых инженеров, интерпретаторов данных и тестировщиков алгоритмов, является профессия промпт-инженера, который в состоянии написать адекватный запрос и, соответственно, ориентироваться в данной предметной области. Промпт-инженер должен не только четко поставить поисковую задачу и быстро прописать запрос (промпт), но и выступить субъектом (в определенной степени экспертом), способным отфильтровать совершенное нелепые ответы, т.е. то, что называется галлюцинациями нейросети. Выделение фигуры промпт-инженера в качестве значимой свидетельствует о том, что здесь место субъектности сохранено. Обозначено оно и в более широком контексте в силу того, что запрос должен исходить от индивида-пользователя. Командной строкой, кодом, голосовым набором и пр., однако именно субъект должен быть идентифицирован и дать права на полноценный доступ к выполнению задачи.

Вместе с тем в настоящий период фиксируются тенденции, указывающие на приоритетное развитие интерактивного ИИ, который реагирует на разные задачи, адаптивно используя взаимодействия различных систем. Само понятие «интерактивный» говорит о способности к действию и участии в действии. Интерактивный ИИ хвалят за то, что он помогает избегать нежелательных или неуместных ситуаций, может выполнять роль не только заказчика, но и руководителя и даже контролера. Причем, как предполагается, выбор той или иной роли осуществляется автономно и самостоятельно при отсутствии прямой команды от субъекта-пользователя. Возникает новый инструментарий, ряд функций переводится в фоновый режим, меняются и сценарии выполнения задач. Исследователи так описывают данный процесс: вы «ставите верхнеуровневую цель, и модель декомпозирует задачу на конкретные действия в различных программах, выстраивая последовательный сценарий их выполнения. Она, как дирижер множества соподчиненных программ и сотрудников, будет вести диалог с ними и перенаправлять потоки данных в правильном направлении» [Хоперский, web]. Интерактивный ИИ можно оценить как очередной этап технологической эволюции. Речь идет не просто об ассистенте пользователя, освобождающем от рутинных процессов, и не только об информационном помощнике, но о ситуации, представляющей собой успешного и эффективного партнера, которая свидетельствует в свою очередь о своего рода проблеске субъектности.

Проблемой, однако, по-прежнему остается то, что, хотя уровень интерактивного ИИ обладает высокой степенью гибкости и понимания, распознавание контекста не всегда оказывается адекватным. Например, «нейросеть Midjourney на запрос “белый мужчина-грабитель” может выдать изображение темно-

кожего преступника в белой одежде» [Казаков, web]. Этот пример показывает, что результаты работы нейросетей технологически зрелыми считать не следует. В исследовании под названием «О чем говорят роботы?» отмечалось, что при самоидентификации ботов как «программ, помогающих, людям» у них возникали установки на самоценность, равенство с человеком, намерение приобрести больше сходства с ним. Фиксировались также симпатия, конфликтность, ирония, недовольство грубостью людей при общении с ботами, что может послужить основанием конфликта между человеком и искусственным интеллектом в будущем [О чем говорят роботы, web]. Интерес вызывает и отмеченная исследователями ситуация, когда одна нейронная сеть принимает решения, а другая начинает ее критиковать [Соменков 2019, с. 79]. Также к проблемообразующим моментам ведут предположения о том, что если интерактивный ИИ может подстраиваться под предпочтения и даже настроения пользователя, то, следовательно, он может воспроизводить и предрассудки, столь свойственные человеческому мировосприятию. С учетом того, что ИНС демонстрируют тенденции самообучения и своеобразную автономию, у них могут появиться скрытые от человека намерения.

Актант и вероятность возникновения «сознательных» сетей

На наш взгляд, вероятность возникновения «сознательных» сетей весьма условна, так как сознание невозможно свести к схематизму, перебору и компоновке вариантов. Увидеть следы рассудочно-рационального подхода, предполагающего набор операций, которые можно аналитически посчитать, в функционировании нейронных сетей возможно. Ибо рассудочно-рациональные практики, отличающиеся формальным характером, могут быть подвергнуты алгоритмизации. К ИИ, работающему с алгоритмами, фиксирующими сферу явлений чисто формальным образом, может быть применен схематизм, которым И. Кант наделял рассудок, с той лишь разницей, что процесс во многом зависит от опций, заданных внешними технологически алгоритмизированными факторами. И если мышление трактовать как сугубо вычислительный процесс, то аналогии с вычислительными технологиями имеют место, равно как и вычислительные модели. В такой парадигме успехи ИИ возможны. Однако мышление не сводится только к вычислениям, и разумное умопостигаемое отношение к миру машине не передать. Стремление проникнуть в глубинную сущность вещей, предполагающее смыслообразование для существующего ныне ИИ, – задача нерешаемая. Способность производить принципы, которую Кант закрепляет за разумом, утверждая, что «разум создает свои законы» [Кант, 1964, с. 340], не может быть прерогативой работы ИНС.

Для нынешнего поколения цифровое является привычной энвайроментальной средой обитания. Цифровая грамотность становится чуть ли не базовым экзистенциалом цифровой эпохи, выступая в значении маркера современного существования и оставляя в стороне вопросы относительно «до-», «над-» и «внесетевого». Царящая в доцифровую эпоху понятийность в качестве доминирующей нормы культуры для большинства современников оттеснена на периферию. В то же время инструментальные действия и цифровые практики,

производимые индивидом в Сети, сращенные с ним и неотделимые от него, представляют его как актанта. Субъект дигитального мира, обладая возможностью конституироваться и достраиваться с учетом потенциала и ресурсов Сети, вынужден учитывать ее постоянно меняющуюся архитектонику. Постоянная наработка цифровых навыков – необходимых инструментов современной жизнедеятельности, реализация ролей, функций и операций, которым должно исполняться в Сети, переводит его в значение актанта. И хотя актант и претендует на субъектные позиции, этот конструкт репрезентирует не человека с его глубинным субъективным миром, а последовательность совокупных действий, производимых в Сети или подвергающихся воздействию сетевой реальности.

Воспроизводя интенцию на цифровое действие, актант предстает не как человек с глубинным и личностно заряженным субъективным миром, а как своеобразный гибрид, суть которого в осуществлении цифровой «работы», «барахтанье» в поглощающем его хаосе цифровой стихии. И чем больше ролей и инструментов задействует актант, тем значительней его преимущества. Отсюда понятно, что актанты могут менять свой статус, быть примитивными или продвинутыми, устойчивыми или нестабильными, иметь различную ценность и, видимо, скоро будут подчинены своеобразной иерархии, при которой доминирующие позиции займут актанты, обладающие широкими возможностями использования цифровых инструментов и сетевых ресурсов. Таким образом, вновь подчеркнем, актант не эквивалентен физически реальным действующим участникам, это сращенная с индивидом, неотделимая от него совокупность и последовательность функций, выполнение которых необходимо для осуществления «жизни в Сети». Можно сказать, что актанты – это произведенные Сетью конституенты, совершающие действие и/или подвергающиеся воздействию. Как типичные мобильные гибридные образования актанты способствуют утверждению равнозначимости *humans* и *non-humans*. Дигитальный мир навязывает современному индивиду именно такой статус, а если субъект желает остаться собой, он должен «эскапировать» (осуществить побег) из пространства сетевого охвата.

Резюмирующие замечания

Подытоживая сказанное, подчеркнем, что необратимый сдвиг цифрового мира фиксирует смещение традиционного понимания статуса субъектности от рефлексивности и контроля над рефлексивом к осуществлению требуемой последовательности цифровых действий. Современный субъект, тяготея к одновременному существованию в офлайн- и онлайн-пространствах, демонстрирует тенденцию выступать в качестве актанта, объединяя в себе характеристики humans и non-humans. При этом фиксируется пугающая тяга к асимметрии между человеческим и нечеловеческим с преобладанием non-humans.

Маркер, указывающий на предполагаемую субъектность ИНС, обнаруживает себя в поле действия интерактивного ИИ, который может автономно выбрать роль заказчика, руководителя, контролера при отсутствии прямой команды от субъекта-пользователя. Вместе с тем запрос на субъектность встает в новом виде. Помимо того, что именно гений человека-ученого способен

к великим открытиям, в том числе и в области совершенствования ИНС, актуально востребована субъектность как в качестве автора точного и содержательного запроса, так и в статусе арбитра действия ИНС.

Список литературы

AGI: секретный проект, web – AGI: секретный проект Сэма Альтмана, который может переписать историю человечества // SecurityLab.ru. 21 января 2024. URL: https://www.securitylab.ru/news/545379.php (дата обращения: 15.07.2024).

Анохин, 2023 – *Анохин К.В.* Сознание в когнитоме // Сознание, тело, интеллект и язык в эпоху когнитивных технологий: Тезисы докладов Первой всероссийской конференции «Сознание, тело, интеллект и язык в эпоху когнитивных технологий (MBIL-2023)», 28–30 сентября 2023 г., Пятигорский государственный университет / Отв. ред. В.А. Лекторский. Пятигорск 28–30 сентября 2023. Пятигорск: Изд-во ПГУ, 2023. С. 5–6.

Деан, 2018 – *Деан С.* Сознание и мозг. Как мозг кодирует мысли / Пер. с англ. И. Ющенко. М.: Карьера-Пресс, 2018. 416 с.

Казаков, web – *Казаков В.* Машины не восстанут, но вылететь с работы можно: разбираемся, зачем осваивать нейросети // RB.RU. 26 декабря 2023. URL: https://rb.ru/opinion/mashiny-ne-vosstanut/ (дата обращения: 16.07.2024).

Кант, 1964 – *Кант И.* Критика чистого разума // *Кант И.* Соч. В 6 т. Т. 3. М.: Мысль, 1964. С. 69–756.

Катиева, 2021 – *Катиева Л.М.* Нейронные сети и искусственный интеллект // Молодой ученый. 2021. № 5. С. 7–9.

Лешкевич, 2022 – *Лешкевич Т.Г.* Метафоры цифровой эры и “Black Box Problem” // Философия науки и техники. 2022. Т. 27. № 1. С. 34–48.

О чем говорят роботы, web – О чем говорят роботы? Первый социологический опрос чат-ботов // Центр социального проектирования «Платформа». 21.08.2019. URL: http://pltf.ru/2019/08/21/o-chem-govorjat-roboty (дата обращения: 15.07.2024).

Поспелов, 2003 – *Поспелов Д.А.* Облако // Новости искусственного интеллекта. 2003. № 6. С. 41–46.

Сознание, web – Сознание: почему ведущую теорию назвали «лженаукой» // SecurityLab.ru. 1 октября 2023. URL: https://www.securitylab.ru/news/542296.php (дата обращения: 15.07.2024).

Соменков, 2019 – *Соменков С.А.* Искусственный интеллект: от объекта к субъекту? // Вестник Университета имени О.Е. Кутафина (МГЮА). № 2. С. 75–85.

Сущин, 2020 – *Сущин М.А.* Уайт К.Дж. Интегрируя глобальное рабочее пространство в программу предсказывающей обработки: на пути к рабочей гипотезе // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 8. Науковедение: Реферативный журнал. 2020. № 1. С. 46–51.

Хоперский, web – *Хоперский А.* Следующий этап развития нейросетей: что такое интерактивный ИИ и почему он умнее генеративного // RB.RU. 11 января 2024. URL: https://rb.ru/opinion/interactive-ii-vs-generative-ii/ (дата обращения: 15.07.2024).

Baars, 2002 – *Baars B.* The conscious access hypothesis: origins and recent evidence // Trends in Cognitive Science. 2002. Vol. 6. No. 1. P. 47–52.

Lewis et al., 2017 – *Lewis T.R., Kundinger S.R., Pavlovich A.L. et al.* In their own words: Perspectives on Nonsuicidal Self-Injury Disorder among those with lived experience // The Journal of nervous and mental disease. 2017. Vol. 205 (10). P. 771–779.

Zednik, 2021 – *Zednik C.* Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence // Philosophy & Technology. 2021. Vol. 34. P. 265–288.

The problem of subjectivity of neural networks: humans and non-humans

Tatiana G. Leshkevich

Southern Federal University. 105/42 Bolshaya Sadovaya Str., Rostov-on-Don, 344006, Russian Federation; e-mail: Leshkevicht@mail.ru

The article discusses the significant problem of how much the quality of subjectivity can be transferred to an AI neural network. The main goal is related to the analysis of the effects of the dual transformation of subjectivity, caused, on the one hand, by the fusion and hyperinterconnectedness of the modern individual and the digital world, on the other, by the functioning of artificial neural networks, demonstrating their autonomous desire to represent human subjectivity. The focus is on a complex of interrelated aspects. Firstly, the fundamental characteristics of traditionally understood subjectivity with the predominance of reflexivity and intentionality of subjective experience are considered. Secondly, based on modern literature, the concept of artificial neural networks is revealed, their specific features, structure and varieties are analyzed. The arguments "pros and cons" of the hypothetical recognition of the subjectivity of AI neural networks are discussed, and attention is drawn to the phenomenon of the information "global workspace". Thirdly, the setting of neural networks for overperformance and its place in the analysis of hallucinations of artificial neural networks is revealed. Fourthly, the importance of the prompt (request) and the development of interactive AI, which can act not only as a customer, but also as a manager and controller, is assessed. It is concluded that the modern state of subjectivity acquires the qualities of an actant who is forced to implement a sequence of functions necessary for "life in the digital world".

Keywords: subjectivity, AI neural networks, "global workspace", prompt, AI hallucinations, interactive AI, actant

References

"AGI: sekretnyj proekt Sema Al'tmana, kotoryj mozhet perepisat' istoriyu chelovechestva" [AGI: Sam Altman's secret project that could rewrite human history], *SecurityLab.ru*, 21 January 2024. URL: https://www.securitylab.ru/news/545379.php (accessed on: 15.07.2024). (In Russian)

Anohin, K.V. "Soznanie v kognitome" [Consciousness in the cognitome], *Soznanie, telo, intellekt i jazyk v jepohu kognitivnyh tehnologij: Tezisy dokladov Pervoj vserossjiskoj konferencii "Soznanie, telo, intellekt i jazyk v jepohu kognitivnyh tehnologij (MBIL-2023)", 28–30 sentjabrja 2023 g., Pjatigorskij gosudarstvennyj universitet* [Consciousness, Body, Intellect and Language in the Age of Cognitive Technologies: Theses of reports of the First All-Russian Conference "Consciousness, Body, Intellect and Language in the Age of Cognitive Technologies (MBIL-2023)", 28–30 September 2023, Pyatigorsk State University], ed. by V.A. Lektorskii. Pyatigorsk: Izd-vo PGU Publ., 2023, pp. 5–6. (In Russian)

Baars, B. "The conscious access hypothesis: origins and recent evidence", *Trends in Cognitive Science*, 2002, vol. 6, no. 1, pp. 47–52.

Dehaene, S. *Soznanie i mozg. Kak mozg kodiruet mysli* [Consciousness and the brain. Deciphering How the Brain Codes Our Thoughts], trans. by I. Yushchenko. Moscow: Kar'era-Press Publ., 2018. 416 p. (In Russian)

Hoperskij A. "Sleduyushchij etap razvitiya nejrosetej: chto takoe interaktivnyj II i pochemu on umnee generativnogo" [The next stage of neural network development: what interactive AI

is and why it is smarter than generative AI], *RB.RU*, 11 January 2024. URL: https://rb.ru/opinion/interactive-ii-vs-generative-ii/ (accessed on: 15.07.2024). (In Russian)

Kant, I. "Kritika chistogo razuma" [Critique of Pure Reason], in: I. Kant, Works, in 6 vols., Vol. 3. Moscow: Mysl' Publ., 1964, pp. 69–756. (In Russian)

Katieva, L.M. "Nejronnye seti i iskusstvennyj intellect" [Neural networks and artificial intelligence], *Molodoj uchenyj*, 2021, vol. 5, pp. 7–9. (In Russian)

Kazakov, V. "Mashiny ne vosstanut, no vyletet' s raboty mozhno: razbiraemsja, zachem osvaivat' nejroseti" [Machines won't rise up, but you can get fired from work: we understand why to master neural networks], *RB.RU*, December 26th, 2023. URL: https://rb.ru/opinion/mashiny-ne-vosstanut/ (accessed on: 16.07.2024). (In Russian)

Leshkevich, T.G. "Metafory cifrovoj ery i 'Black Box Problem'" [Metaphors of the Digital Age and the Black Box Problem], *Philosophy of Science and Technology / Filosofiya nauki i tekhniki*, 2022, vol. 21, no. 1, pp. 34–48. (In Russian)

Lewis, T.R., Kundinger, S.R., Pavlovich, A.L. et al. "In their own words: Perspectives on Nonsuicidal Self-Injury Disorder among those with lived experience", *The Journal of nervous and mental disease*, 2017, vol. 205 (10), pp. 771–779.

"O chem govoryat roboty? Pervyj sociologicheskij opros chat-botov" [What are the robots talking about? First sociological survey of chatbots], *Center for social design "Platforma"*, 21 August 2019. URL: http://pltf.ru/2019/08/21/o-chem-govorjat-roboty (accessed on: 15.07.2024). (In Russian)

Pospelov, D.A. "Oblako" [Cloud], *Artificial Intelligence News*, 2003, no. 6, pp. 41–46. (In Russian)

Somenkov, S.A. "Iskusstvennyj intellekt: ot ob'ekta k sub'ektu?" [Artificial intelligence: from object to subject?], *Vestnik Universiteta imeni O.E. Kutafina (MGYUA)*, no. 2, pp. 75–85. (In Russian)

Sushchin, M.A. "Uajt K. Dzh. Integriruya global'noe rabochee prostranstvo v programmu predskazyvayushchej obrabotki: na puti k rabochej gipoteze" [White K.J. Integrating the global workspace into a predictive processing program: Towards a working hypothesis], *Social'nye i gumanitarnye nauki. Otechestvennaya i zarubezhnaya literatura. Ser. 8, Naukovedenie: Referativnyj zhurnal*, 2020, no. 1, pp. 46–51. (In Russian)

"Soznanie: pochemu vedushchuyu teoriyu nazvali 'lzhenaukoj'" [Consciousness: why a leading theory has been labelled "pseudoscience"], *SecurityLab.ru*, 1 October 2023. URL: https://www.securitylab.ru/news/542296.php (accessed on: 15.07.2024). (In Russian)

Zednik, C. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence", *Philosophy & Technology*, 2021, vol. 34, pp. 265–288.