Философия науки и техники
2023. Т. 28. № 1. С. 34–47
УДК 004.89

Philosophy of Science and Technology
2023, vol. 28, no. 1, pp. 34–47
DOI: 10.21146/2413-9084-2023-28-1-34-47

НАУКА, ТЕХНИКА, ОБЩЕСТВО

Т.Г. Лешкевич

Парадокс доверия к искусственному интеллекту и его обоснование

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

Статья посвящена анализу доверия к цифровым интеллектуальным системам, которые, активно участвуя в трансформации социальных практик и процессах принятия решений, становятся «новыми» посредниками нашего существования. Однако, поскольку технологии искусственного интеллекта (ИИ) могут столкнуться с неполадками, основная проблема связана с анализом парадокса доверия к ИИ, включающего в себя как позицию доверительного отношения к онлайновым системам, серверам и программному обеспечению, так и факты сбоев, демонстрируемых искусственным интеллектом. Целью статьи является изучение парадокса доверия к системе ИИ с учетом сопоставления аргументов «pro & contra». Это предполагает, во-первых, выявление сути и специфики прокси-культуры как культуры доверия. Во-вторых, рассмотрение непрозрачности ИИ, усиливающей потребность в алгоритмической ответственности. В качестве теоретической основы выступают работы отечественных и зарубежных исследователей. Методологическая стратегия опирается на компаративистский анализ и сопоставление социально-гуманитарного понимания доверия и специфики доверия к ИИ. В основе выводов лежит заключение, что существование парадокса сопровождается малопонятным компромиссом, суть которого в игнорировании всей полноты рисков в пользу удобства пользования быстрыми интеллектуальными системами. Алгоритмическая ответственность, направленная на сокращение ненадежности и угроз в использовании ИИ, вступает в противоречие с обязательностью, обеспеченной запрограммированным кодом. Приоритеты технологического развития состоят в выработке стандартов алгоритмической прозрачности, обеспечивающих раскрытие информации, относящейся к последствиям принятых алгоритмических решений.

Ключевые слова: искусственный интеллект, прокси-культура, доверие, алгоритмическая ответственность, непрозрачность алгоритмов

В ситуации XXI в. на фоне четко артикулированной проблемы неконтролируемых последствий развития искусственного интеллекта (ИИ), именуемой как «Black Box Problem» («проблема выхода из-под контроля»), распространена альтернативная практика доверия интеллектуальным системам, онлайн-платформам и веб-сайтам, принимающим запросы и действующим от имени пользователя. Такой тип взаимодействий, инициированных от имени «кого-то», в англоязычных источниках назван прокси-культурой [Floridi, 2015]. Устройства, становясь умными и «сверхумными», возлагают на себя функции быть референтом человеческих апелляций. Но поскольку имеют место как позиция, свидетельствующая о безусловном доверии к ИИ, так и позиция, фиксирующая беспокойство по поводу неконтролируемого или даже злонамеренного использования ИИ [Пашенцев, 2019], возникает парадокс доверия к ИИ. Информация, поступающая из различных источников, работает на обоснование как одной, так и второй стороны обозначенного парадокса, содержащего в себе либо возможность испытывать состояние предельно доверительного отношения к ИИ, либо ужасаться фактами ошибок интеллектуальных систем, оказавшись своеобразной жертвой цифрового алгоритма. Причем бытийствование данного парадокса сопровождается малопонятным компромиссом, суть которого в игнорировании всей полноты рисков в пользу удобства пользования быстрыми интеллектуальными системами. Международная группа ученых собрала внушительную подборку сбоев в работе алгоритмов [Lehman et al., 2020], показав, что не случайно цифровую эпоху называют эпохой программ, которые могут столкнуться с неполадками. Поэтому основная проблема связана со столкновением двух противоположных тенденций: с одной стороны, это субъекты, доверяющие онлайновым системам, серверам и программному обеспечению так, как они доверяли бы самим себе, с другой – глитч, который демонстрирует искусственный интеллект, свидетельствуя о своих неполадках. Анализ как зарубежных, так и отечественных исследований, представленных именами: Л. Флориди, Дж. Кларк, Н. Диакопулос, А. Мамфорд, Дж. Зарски, Дж. Зедник, М. Уилсон, М. Бруссард, а также С. Гарбук, А. Латыпова, Т. Лешкевич, Л. Манович, Т. Мартыненко, Д. Добринская, Е. Пашенцев, А. Пинчук, Д. Тихомиров и др., показывает, что выявление аргументов, выставленных в «защиту» как одной, так и другой стороны, актуально в силу того, что позволит продвинуться в осмыслении траектории совершенствования ИИ. Более того, парадоксальность как таковая принята за свойство и инструмент исследовательского мышления, пробуждающего творческую мысль в направлении поиска ответа. Можно попытаться измерить частотность ошибок ИИ или же постараться определить прослойку людей, избегающих обращения к ИИ, однако всё это будет работать на фиксацию данного парадокса.

Анализ парадокса доверия к системе ИИ, являющийся основной целью статьи, показывает, что данный тип доверия предполагает принятие работающего технологического принципа, включающего программный код, и не может быть отождествлен с доверием, основанным на межличностном общении. Вместе с тем основа идеи доверия, признанная в том числе и в современном социо-гуманитарном дискурсе, связана с делегированием кому-либо тех или иных функций, или согласием и готовностью следовать тем или иным

«правилам игры», решениям и регламентациям на добровольной основе. Важно подчеркнуть, что феномен доверия выступает в статусе исходного метаотношения, выполняя универсальную регуляторную роль, т.е. индивид доверяет или не доверяет людям, институциям, организациям, техническим системам и пр. вплоть до прогноза погоды. Расхожей фразой является оценочный вывод о «дефиците доверия», бытует и противоположная метафора «абсолютного доверия», известно и то, что доверие нужно завоевывать и его легко потерять. Различные уровни доверия идентифицируются в связи с осознанием уязвимости того или иного типа взаимодействий и их негарантированных исходов. Согласно выводам исследователей, под доверием к системе искусственного интеллекта подразумевается степень уверенности пользователя или другого заинтересованного лица в том, что ИИ будет выполнять свои функции так, как это предполагалось. Так, «для пользователей систем ИИ первостепенное значение имеют функциональные возможности этих систем, для разработчиков и поставщиков – характеристики конкурентоспособности, для третьих лиц – характеристики безопасности и т.д.» [Гарбук, 2020]. В современных условиях, когда цифровизация проникла во все сферы жизненного мира, обеспечение доверия к системам ИИ является ключевым фактором и главным требованием технологической эволюции.

Феномен «прокси»-культуры и ее функции

При рассмотрении «прокси»-культуры следует отметить, что этимологически термин «прокси» (proxy) восходит к англоязычному понятию «procuracy» – «действовать по доверенности», и, по сути, представляет собой сокращение от этого термина. «Прокси», означая «законные действия, предпринятые от имени кого-то», обнаруживает себя в контексте общественно-политической сферы или юридической практики. В рамках международно-политического дискурса термин «прокси» употребляется, указывая на различные модификации прокси-войн и прокси-конфликтов как опосредованных войн и конфликтов, которые ведутся не самими акторами процесса, а, так сказать, «чужими руками» [Mumford, 2013]. В информационном обществе феномен «прокси», как отмечают исследователи, вызывает ассоциацию с онлайн системами (вебсайтами, платформами и серверами), которые принимают запросы на некоторые услуги и передают их в другую систему (например, в Интернет) [Floridi, 2015]. Тем самым идея доверия, разрешающая «действие от лица некоего субъекта», оказывается в основе такого взаимодействия. Прокси-сервер делает возможным и обеспечивает определенную операцию благодаря тому, что представляет некий искомый объект и включает в себя набор интеракций, обеспечивающих действия от имени субъекта. При этом физический контакт с реальным представителем становится ненужным. Проблема решается посредством ИИ с его технологическими возможностями, а доверие к ИИ выступает как своего рода доверие к разработчикам софта. Так, нас не волнует, что мы не знаем маршрута, если у нас есть доступ к Google Maps. Пяти звезд на Amazon достаточно, чтобы убедить нас в качестве товара, даже если мы никогда не видели его сами. Отмеченный звездочками статус «бестселлера» значим

для формирования нашего выбора. К примерам информационных «прокси» относятся также выбор отеля или ресторана, когда мы всецело полагаемся на сервер TripAdvisor, выбор путешествия, возникновение дружеских связей на соответствующих платформах без того, чтобы встретиться с реальным человеком и пр. Иными словами, мы доверяем выборам, предложенным информационными системами, что в свою очередь свидетельствует об отчужденной форме доверия, где главным актом является возможность действовать «от имени» или вместо своего референта. А доступ и сбой становятся фронтирами, которые либо поддерживают, либо прерывают данный процесс.

Наш современник, делегируя сервисам возможность действовать от своего имени, оказывается носителем прокси-культуры, которая в условиях цифровой детерминации подчиняет его анонимной коллективной идентичности. Цифровизация, пронизывая все виды взаимодействий, создает слои социальности, в которой нет нужды в физически присутствующем человеке. Вход в дигитальный мир связан с цифровыми компетенциями, расширение опыта означает умение ориентироваться в заданных алгоритмах, а адекватный ответ вызовам времени состоит в том, чтобы «нарастить цифровые мускулы». Ситуация конвергенции субъекта и ИИ, предполагающая взаимопроникновение способностей человека и ресурсов Сети, эту тенденцию усиливает. Вместо физического взаимодействия с означаемым субъект взаимодействует с миром посредством прокси, т.е. полагаясь на программные объекты и алгоритмы. Таким образом, ИИ, будучи технологическим артефактом, обретает статус естественного и постоянного партнера. Прикладные приложения ИИ, имея своей главной задачей максимизацию общественной пользы, охватывают разнообразные сферы жизненного мира. Это и услуги населению, и безопасность, и торговля, и управление. Особое значение для передачи культурной эстафеты приобретает цифровая библиотека, цифровой музей, кинотеатр, концертный зал, научная организация и пр. Осуществляемая при помощи ИИ оцифровка культурного наследия обладает эффектом его популяризации и меморизации и полагается на активы и резервуары цифрового наследия, характеризуемого отсутствием материально-вещественной формы [Leshkevich, Motozhanets, 2022]. И поскольку цифровые системы и технологические приложения становятся формообразующим ингредиентом современности, то, выступая связующим основанием между субъектом и полюсом его потребностей, они выполняют посредническую функцию.

Помимо функции опосредования и по мере того, как субъект расширяет зону прокси, позволяя системам ИИ действовать от его имени, дает о себе знать прагматическая функция. Она связана с оптимизацией личностных физических усилий и установкой на получение немедленной реакции на возникший запрос. Мир офлайн потребления, в котором акт покупки и владение материальным товаром являлся основной ценностью, сменяется типом взаимодействий с окружающей средой посредством информационных интеракций на основе доверия к интеллектуальным системам. В этом отношении феномен доверия к ИИ тесно связан с успешными пробами и эквивалентным обменом, при котором намерения субъекта сопровождаются ожидаемыми результатами интеракций. Поэтому приложения с функциями «умных помощников» приоб-

ретают огромную популярность. Примеров услуг, предоставляемых интеллектуальными системами, множество: это и электронная покупка билетов, и сведения о пробках на дороге, наличии или отсутствии товаров, финансовый банкинг и пр. В силу того, что к 2020 г. человечество накопило предположительно 35 секстибайт данных, в инфосфере становится просто невозможно ориентироваться, не полагаясь на прокси-серверы. Зависимость от компьютеров и систем ИИ возросла многократно. Существуют данные, что в 2019 г. в мире цифровыми помощниками пользовались около 3,25 млрд человек, а к 2023 г. их число достигнет 8 млрд [Purwanto et al., 2020]. Согласно прогнозу К. Шваба, до 2025 г. должны произойти изменения, включающие в себя «1 трлн датчиков, подключенных к сети интернет; первый имеющийся в продаже имплантируемый мобильный телефон; 10% людей будут носить одежду, подключенную к сети интернет; 90% населения будут обладать регулярным доступом к сети» [Шваб, 2017, с. 39]. Всё это свидетельствует о трансформациях, учреждающих новый цифровой тип существования.

Доверие к ИИ и проблема алгоритмической ответственности

Выход из реального «аналогового» мира делает цифровую среду местом обитания современного человека, однако «цифровая онтология», по определению Ю. Хуэй, – это «единство, состоящее из множества формальных свойств» [Hui, 2016]. Оно представляет собой корпус организованных данных, которые существуют в виде кода и могут принимать визуальную, звуковую, текстовую форму. В обществе взаимосвязанных данных важность алгоритмов становится очевидной. Прокси-серверы, обеспечивающие доступ к услугам, замещают физический акт приобретения, а иногда и владения товарами. При этом функция замещения отношений с объектами реального мира цифровыми транзакциями позволяет включить в сферу человеческого опыта ранее физически затруднительные или даже невозможные взаимодействия. Прокси-серверы могут служить мостиками к труднодоступным и недоступным сферам нашего опыта. Так, можно путешествовать по Марсу, оказаться героем событий прошлого, строить различные варианты жизненного пути либо погружаться в среды, малодоступные в реальности, и осваивать различные типы нештатных ситуаций [Лешкевич, 2022, с. 55]. Однако, открывая большие возможности, прокси-культура как культура доверия сопряжена с многочисленными рисками суррогатных референтов. Исследователи выявляют новые формы уязвимости, осмысливая проблемы «фейков» и «пузырей фильтров» в Сети [Гуров, 2019, с. 11]. Это указывает на правовой аспект проблемы доверия ИИ и на острую потребность научиться контролировать тех, кто будет создавать серверы и программное обеспечение. В связи с чем громко заявляет о себе проблема алгоритмической ответственности, обусловленная двумя характерными чертами: автоматизированностью алгоритмических процессов и их непрозрачностью.

Острота проблемы алгоритмической ответственности связана с тем, что лица, принимающие решения, полагаются на выводы автоматизированной системы при минимальном участии человека или вовсе без него. Решение основывается на принципе компьютерной алгоритмизации: алгоритмизированные

системы принимают определенные входные данные и генерируют определенные выходные данные с помощью вычислительных средств. И, как заключают исследователи, в основе всего, что мы делаем на компьютере, лежат математические действия, которые имеют свои фундаментальные ограничения, определяющие границы технологий [Бруссард, 2020, с. 15]. В общем виде алгоритм – это организованный специальным образом набор шагов для обработки данных в направлении достижения поставленной цели. Алгоритмические системы как комплексные образования включают в себя не только сами алгоритмы, но и вычислительные сети, в которых они функционируют, и людей, которые их проектируют и используют, данные (и пользователей), на которых они воздействуют, а также субъектов (индивидуальных и корпоративных), которые предоставляют услуги с помощью алгоритмов [Мартыненко, Добринская, 2021, с. 176].

В поголовном большинстве случаев мы сталкиваемся с трудностями приписывания и истребования ответственности за действия алгоритмических систем. Так, в ситуации, когда автоматизированная система кредитного рейтинга какого-либо банка отклоняет заявку клиента на получение кредита, ответственен ли за это банк, использующий алгоритм? На основе алгоритмизированных моделей могут быть составлены классификации с последующим принятием соответствующих решений в отношении живых людей. Работодатели используют алгоритмические инструменты при выборе сотрудников. Кредитодатели отслеживают онлайн поведение клиента, которое, как предполагается, коррелирует с кредитоспособностью, например, скорость, с которой потенциальные заемщики просматривают веб-сайты, определяют наиболее релевантные конкретному поисковому запросу веб-страницы. При этом «одним из типов входных данных является информация о том, на какие из ссылок, выданных ранее по тому же запросу, кликали пользователи» [Манович, 2015, с. 204]. Компании в целях оптимизации продуктивности и прогнозирования неудач мониторят онлайн активность своих работников, отслеживая их «цифровую тень», свидетельствующую о всех произведенных индивидом компьютерных интеракциях. Алгоритмы ранжируют потенциальных нелегальных иммигрантов, чтобы лица из групп повышенного риска подвергались более тщательному анализу. С использованием алгоритмических технологий принимаются решения об одобрении тех или иных кандидатур в качестве приемных родителей. Есть алгоритмы, которые подвергают цензуре любой нежелательный контент, исключая не только поисковые запросы, связанные с порнографией, насилием, разжиганием ненависти, но и блокируя высказывания, представляющие не совпадающие с официальной политикой независимые взгляды. При этом возникает этический вопрос: насколько справедливо формировать определенное мнение об индивиде на основе совокупных алгоритмизированных данных с встроенной программой индикаторов, использованием операций приоритизации и фильтрации?

И поскольку алгоритмы – это программные продукты, в конечном счете созданные человеком, следует принимать во внимание намерения разработчиков, а также желания группы людей и институциональных процессов, оказывающих влияние на их конструкцию. На сегодняшний день фиксируется как «сложность» алгоритмов, так и неясность правовых и институциональных

структур, в которых они функционируют. «Сложность» алгоритмов является одной из причин их непрозрачности. Действительно, в силу того, что алгоритмические коды скрыты за слоями технической сложности, они трудно отслеживаются и интерпретируются. И хотя автономное принятие решений является сутью алгоритмической силы, сами правила могут быть сформулированы непосредственно программистами или быть динамичными и гибкими на основе данных. Так, алгоритмы машинного обучения, основанные на усвоенных закономерностях в данных, позволяют другим алгоритмам делать решения умнее. Однако в тех случаях, когда результаты настолько важны или, напротив, беспорядочны и неопределенны, окончательное решение принимает оператор-человек. Вследствие этого исследователи с особой настоятельностью подчеркивают, что на первый план выходят вопросы ответственности за разработку и внедрение алгоритмов, а также проблема контроля над результатами их работы, зачастую заранее непрогнозируемыми [Willson, 2017].

И хотя в условиях масштабной цифровизации алгоритмическая ответственность возводится в ранг как общественно значимого, так и индивидуально значимого фактора, анализ этой проблемы выявляет амбивалентное к ней отношение. С одной стороны, замена человеческого ресурса автоматизированными системами может нейтрализовать человеческий фактор и связанную с ним предвзятость в процессе принятия решений [Zarsky, 2015]. Но, с другой – формы алгоритмизированного принятия решений могут обладать потенциалом воплощения определенной системы ценностей и воспроизводить предубеждения своих авторов [Nissenbaum, 2001]. В программное обеспечение могут быть внедрены алгоритмы, выработанные с учетом ценностей программиста-разработчика и анонимно содержащие в себе идею неравенства по признаку пола, расы или других этически значимых факторов. И тогда, опираясь на эти характеристики, алгоритмические решения могут усилить социальные предубеждения, увековечить вредные стереотипы и в конечном счете подорвать доверие к системе ИИ. Указывая на ситуацию, когда разработка и настройка алгоритмов тем или иным способом может дать преимущество различным заинтересованным сторонам в принятии определенного решения, исследователи ставят вопрос о том, что, если ценностные предпочтения и использование определенного критерия являются политически или каким-либо иным образом предвзятыми? [Diakopoulos, web]. Ведь очевидно, если алгоритм настроен на путь к ложным срабатываниям, то алгоритм пометит многое, как нарушающее авторские права (например, видео добросовестного использования). А если алгоритм настроен по-другому, то он пропустит многое из того, что, быть может, нарушает авторские права [Ibid., p. 7]. Вместе с тем, согласно имеющимся данным, около 46% европейцев положительно оценивают работу алгоритмов и только 20% озабочены негативными последствиями алгоритмического управления и принятия решений на основе алгоритмов [Grzymek, Puntschuh, 2019]. По результатам опроса ВЦИОМ, проводившегося в России в 2019 г., большая доля россиян положительно (48%) или, по крайней мере, нейтрально (31%) относятся к распространению технологий искусственного интеллекта: 68% не опасаются замещения технологиями ИИ

специалистов своей профессии, а 87% считают, что государство должно способствовать развитию этих технологий [ВЦИОМ, 2020].

На наш взгляд, для решения проблемы доверия к ИИ целесообразно различать специфику онтологии его развития, онтологии внедрения и поддержки, сопровождения и использования. Важна и градация взаимодействий человека и систем ИИ, которая, как правило, представлена четырьмя достаточно выраженными типами. Это ориентации на расширенное использование интеллектуальных систем, обеспечивающих решение так называемых «антропоморфных» задач интеллектуальной обработки информации, т.е. «интеллектуальная поддержка человека». Заметно и ограниченное обращение к ИИ с учетом «прерогатив человека», когда людям не нравится полагаться на ИИ и они предпочитают доверять экспертам. Исследователи также выделяют обращение к ИИ сугубо в прикладном аспекте с целью облегчения «рутинных операций», либо операций, направленных на решение «прикладных задач» [Пинчук, Тихомиров, 2019].

Еще один значимый аспект выявлен в исследованиях токсичного поведения ИИ компанией DeepMind. Отметим, что «под зонтиком» токсичности группируются способы генерирования оскорбительных выражений, включая язык ненависти, ненормативную лексику и угрозы [Gehman et al., 2020]. В отчете компании DeepMind был отмечен неожиданный эффект: «Несмотря на эффективность принципа блокировки триггерных фраз и оптимизации системы ответов без оскорблений, искусственный интеллект лишился почти всех слов, относящихся к меньшинствам, а также диалектизмов и в целом упоминаний маргинализированных групп [DeepMind…, 2021]. Но поскольку формы коммуникативных практик являются ценностно-чувствительными и могут значительно различаться в разных странах, культурах и социальных группах, а также принимая во внимание, что ИИ-системы не смогут, образно выражаясь, «вовремя прикусить себе язык», возникает правомерный вопрос. Может ли компьютерная алгоритмизация иметь доказательную силу, равную результатам рефлексивных обобщений, сделанных учеными-специалистами? Научная литература пестрит замечаниями относительно того, что многие запрограммированные действия ИИ «непрозрачны» и не совсем ясно, какой тип регулярности или корреляции между входами и выходами действительно имеет место. В то время как в некоторых случаях может присутствовать простая статистическая корреляция, в других она может относиться к добросовестной причинной закономерности [Zednik, 2021]. Насколько критично, что алгоритмы «не умеют» отличать причинно-следственную связь от корреляции, в то время как для науки причинно-следственная зависимость является альфой и омегой научного обоснования.

В силу того, что техническая непрозрачность скрывает и затемняет внутреннюю работу, компьютерные программы образно именуют «черными ящиками». И полагая, что в них «что-то» происходит, исследователи отмечают, что «медиальное устройство оказывается настолько сложным в силу количественного разнообразия входящей информации и методов работы с ней, что при каждой итерации возникает новая конфигурация (повторить которую по силам только самой программе, если сойдутся входящие условия)» [Латыпова, 2020,

с. 168]. Вместе с тем общая цель состоит в политике прозрачности, в том, чтобы следовать четкому раскрытию информации, относящейся к последствиям и принятию более обоснованных алгоритмических решений. При этом важны не только результаты работы алгоритма, но и то, как этот вывод доступен пользователю. Значимым представляется, во-первых, то, чтобы алгоритмические решения и их последствия могли бы быть представлены понятным языком, возможно, с несколькими уровнями детализации, которые интегрируются в решения, с которыми сталкиваются конечные пользователи, а также разработчики алгоритма. Во-вторых, имеет значение общение с разработчиками системы, предоставляющими полезную информацию: проектные решения, описания цели, ограничения и бизнес-правила, а также встроенные в систему основные изменения и внедренные детали, которые произошли с течением времени. В-третьих, ставится вопрос об эффективном взаимодействии с алгоритмами через процесс обратной инженерии. Предполагается, что, поняв отношения ввода-вывода алгоритма, можно понять, как он работает.

Поскольку концептуализация тематики «доверия к ИИ» находится в исходном состоянии, то в исследованиях можно встретить понятие «алгоритмическое доверие», а содержание понятия ответственности заменяется «технической обязательностью», полагающей беспрепятственное функционирование запрограммированного кода. Вместе с тем очевидно, что алгоритмическая ответственность, направленная на сокращение ненадежности и рисков, вступает в противоречие с обязательностью, обеспеченной запрограммированным кодом. Есть и противоречие между идеалом прозрачности и реальным функционированием алгоритмов. В связи с этим становится очевидным, что мы не можем остановиться лишь на осуждении отсутствия ответственности за алгоритмы и их эффекты. Важно предложить стандарты алгоритмической прозрачности, которые, признавая проблемы управления и бизнеса, обеспечивали бы предоставление полезной информации общественности.

Подводя итоги, отметим следующее. Парадокс доверия к ИИ порождает аргументы pro & contra. Так, с одной стороны, в докладе Римского клуба было отмечено: «Нет сомнения, что все положительные вещи, связанные с ИКТ и цифровыми технологиями, при рассмотрении их прямых последствий с точки зрения устойчивости, вызывают отрицательные эффекты первого порядка» [Von Weizsäcker, Wijkman, 2018, p. 46]. Однако, с другой, новейшие технологии и способы взаимодействия с ними признаны очередной ступенью техноэволюции и оцениваются как магистраль расширения человеческих возможностей [Clark, 2004]. След парадокса доверия к ИИ ощутим в сопоставлении известных характеристик, указывающих на объем памяти, во многом превышающий человеческую, огромную скорость обработки информации и принятия решения, с одной стороны, а с другой – с негативным когнитивным воздействием практики «аренды знания». Ситуация, когда индивид направляет поисковый запрос и, пользуясь ресурсами Сети, выдает найденную там информацию за собственное знание, свидетельствует о том, что доверие к ИИ оборачивается своей превращенной формой, насаждающей тип компилятивной

и безрефлексивной рациональности. В силу того, что алгоритмы опираются на математический формализм и рассчитаны на стационарные ситуации, вынужденная ориентация на принятие решений посредством цифровых алгоритмов грозит примитивизацией смыслосферы. Человек становится послушным реципиентом, весьма ограниченным в выборах и предпочтениях.

Анализ парадокса доверия к ИИ с точки зрения функционального аспекта указывает на угрозы того, что ИИ, используя собственные преимущества, станет корректировать самого себя, выйдет из-под контроля и начнет действовать злонамеренно. Рядовой пользователь беспомощен перед ошибками цифровых алгоритмов, разработчики также подчас «разводят руками», не в состоянии объяснить глитч ИИ. Феномен доверия обусловлен потребностью в прозрачности и объяснимости работы ИИ, вместе с тем обеспечение доверия к ИИ тесно сопряжено с безопасностью его функционирования для жизни и деятельности человека. В силу этого аксиологический срез проблемы акцентирует необходимость согласованности системы человеческих ценностей с программами развития и использования ИИ. И хотя оценки доверия к системам ИИ формулируются по-разному, в них должны учитываться не только соображения эффективности, но и справедливости. Важный этико-правовой фактор связан с потребностью перестройки отношений с информационными технологиями с целью подчинения их человеческой рефлексии и осознанности.

Список литературы

Бруссард, 2020 – *Бруссард М.* Искусственный интеллект. Пределы возможного / Пер. с англ. Е. Арье. М.: Альпина нон-фикшн, 2020. 362 с.

ВЦИОМ, 2020 – Искусственный интеллект: угроза или возможность? // ВЦИОМ. Аналитический обзор. 2020. 27 января. URL: https://wciom.ru/index.php?id=236&uid=10132 (дата обращения: 23.05.2022).

Гарбук, 2020 – *Гарбук С.В.* Особенности применения понятия «доверие» в области искусственного интеллекта // Искусственный интеллект и принятие решений. 2020. № 3. С. 15–21.

Гуров, 2019 – *Гуров П.Н.* Опыт социально философского осмысления проблемы «фейков» и «пузырей фильтров» в Сети // Проблемы современного образования. 2019. № 3. С. 9–20.

Латыпова, 2020 – *Латыпова А.Р.* Между мутацией и глитчем: цифровая эволюция медиа // Эпистемология и философия науки. 2020. Т. 57. № 2. С. 162–178.

Лешкевич, 2022 – *Лешкевич Т.Г.* Человек-виртуал и передача культурных ценностей поколению эпохи цифры // Вопросы философии. 2022. № 3. С. 53–63.

Манович, 2015 – *Манович Л.* Как следовать за пользователями программ? // ЛОГОС. 2015. Т. 25. № 2 (104). С. 189–218.

Мартыненко, Добринская, 2021 – *Мартыненко Т.С., Добринская Д.Е.* Социальное неравенство в эпоху искусственного интеллекта: от цифрового к алгоритмическому разрыву // Мониторинг общественного мнения: экономические и социальные перемены. 2021. № 1. С. 171–192.

Никифорова и др., 2020 – *Никифорова В.Д., Никифоров А.А., Викторова В.А.* Экономические и управленческие аспекты внедрения финансовых технологий // Научный журнал НИУ ИТМО. Серия Экономика и экологический менеджмент. 2020. № 4. С. 99–105.

Пашенцев, 2019 – *Пашенцев Е.Н.* Злонамеренное использование искусственного интеллекта: новые угрозы для международной информационно-психологической безопасности и пути их нейтрализации // Государственное управление. Электронный вестник. 2019. № 76. С. 279–300.

Пинчук, Тихомиров, 2019 – *Пинчук А.Н., Тихомиров Д.А.* О взаимодействии человека и искусственного интеллекта: новая социальная реальность в представлении московских студентов // Социология и жизнь. 2019. № 3. С. 85–97.

Шваб, 2017 – *Шваб К.* Четвертая промышленная революция. М.: Издательство «Э», 2017. 208 с.

Clark, 2004 – *Clark A.* Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence. Oxford, England: Oxford University Press, 2004. 240 p.

DeepMind…, 2021 – DeepMind заявила Google, что не знает, как сделать искусственный интеллект менее токсичным // Хабр. 21.09.2021. URL: https://habr.com/ru/news/t/579196/ (дата обращения: 23.05.2022).

Diakopoulos, web – *Diakopoulos N.* Algorithmic Accountability Reporting: on the Investigation of Black Boxes // Columbia Journalism School. Tow Center for Digital Journalism. 2014. 33 p. URL: https://doi.org/10.7916/D8ZK5TW2 (дата обращения: 18.05.2022).

Floridi, 2015 – *Floridi L.* A Proxy Culture // Philosophy and Technology. 2015. Vol. 28. P. 487–490.

Gehman et al., web – *Gehman S., Gururangan S., Sap M., Choi Y., Smith N.A.* RealToxicityPrompts: Evaluating neural toxic degeneration in language models // arXiv:2009.11462. 2020. URL: https://arxiv.org/pdf/2009.11462 (дата обращения: 18.05.2022).

Grzymek, Puntschuh, 2019 – *Grzymek V., Puntschuh M.* What Europe Knows and Thinks About Algorithms Results of a Representative Survey. Gütersloh: Bertelsmann Stiftung, 2019. 38 p.

Hui, 2016 – *Hui Y.* On the Existence of Digital Objects. Minneapolis, London: University of Minnesota Press, 2016. 336 p.

Lehman et al., 2020 – *Lehman J., Clune J., Misevic D., Adami C., Altenberg L., Beaulieu J., Yosinski J. et al.* The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities // Artificial life. 2020. № 26 (2). P. 274–306.

Leshkevich, Motozhanets, 2022 – *Leshkevich T., Motozhanets A.* Social Perception of Artificial Intelligence and Digitization of Cultural Heritage: Russian Context // Applied Sciences. 2022. № 12 (5). P. 2712.

Mumford, 2013 – *Mumford A.* Proxy Warfare: War and Conflict in the Modern World. Cambridge: Polity Press, 2013. 141 p.

Nissenbaum, 2001 – *Nissenbaum H.* How computer systems embody values // Computer. 2001. Vol. 34. № 3. P. 118–120.

Purwanto et al., 2020 – *Purwanto P., Kuswandi K., Fatmah F.* Interactive Applications with Artificial Intelligence Applications: The Role of Trust among Users // Foresight and STI Governance. 2020. Vol. 14. № 2. P. 64–75.

Schmidt et al., 2020 – *Schmidt P., Biessmann F., Teubner T.* Transparency and trust in artificial intelligence systems // Journal of Decision Systems. 2020. Vol. 29. № 4. P. 260–278.

Von Weizsäcker, Wijkman, 2018 – *Von Weizsäcker E.U., Wijkman A.* Come On! Capitalism, Short-termism, Population, and the Destruction of the Planet. N.Y.: Springer, 2018. 220 p.

Welbl et al., 2021 – *Welbl J., Amelia Glaese A., Uesato J., Dathathri S., Mellor J., Hendricks L., Pushmeet K., Coppin K., Huang P.* Challenges in Detoxifying Language Models // Findings of the Association for Computational Linguistics: EMNLP. 2021. P. 2447–2469.

Willson, 2017 – *Willson M.* Algorithms (and the) everyday // Information, Communication & Society. 2017. Vol. 20. № 1. P. 137–150.

Zarsky, 2015 – *Zarsky T.* The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making // Science, Technology, & Human Values. Vol. 41. № 1. P. 118–132.

Zednik, 2021 – *Zednik C.* Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence // Philosophy & Technology. 2021. № 34. P. 265–288.

The paradox of trust in artificial intelligence and its rationale

Tatiana G. Leshkevich

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

The article is devoted to the analysis of the phenomenon of trust in digital intelligent systems that are actively involved in the transformation of social practices, decision-making processes and become "new" mediators of our existence. As AI technologies run into trouble, the main challenge is to analyze the paradox of trust in AI, which includes both trust in online systems, servers and software, as well as failures demonstrated by artificial intelligence. The purpose of the article is to study the paradox of trust in the AI system, taking into account the comparison of "pro & contra" arguments. This involves, firstly, identifying the essence and specifics of the proxy culture as a culture of trust. Secondly, consideration of AI opacity and algorithmic responsibility becomes important. The theoretical basis is the modern work of Russian and foreign researchers. The methodological strategy includes a comparative analysis and comparison of the socio-humanitarian understanding of trust and the specifics of trust in AI. The conclusions are based on the following statements. The AI trust paradox is accompanied by an obscure compromise, the essence of which is to ignore all the risks due to the usability of fast intelligent systems. Algorithmic responsibility, aimed at reducing the unreliability and threats of using AI, conflicts with the obligation provided by the programmed code. The priorities of technological evolution are to develop standards for algorithmic transparency that ensure the disclosure of information related to the consequences of algorithmic decisions.

Keywords: artificial intelligence, proxy culture, trust, algorithmic responsibility, AI opacity

References

Broussard, M. *Iskusstvennyj intellekt. Predely vozmozhnogo* [Artifical Unintelligence. How Computers Misunderstand the Word], transl. by C. Arie. Moscow: Alpina non-fiction Publ., 2015. 362 pp. (In Russian)

Clark, A. *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence.* Oxford, England: Oxford University Press, 2004. 240 pp.

"DeepMind zayavila Google, chto ne znaet, kak sdelat' iskusstvennyi intellekt menee toksichnym" [DeepMind told Google that it does not know how to make artificial intelligence less toxic], *Habr*, September 21, 2021 [https://habr.com/ru/news/t/579196/, accessed on 23.05.2022]. (In Russian)

Diakopoulos, N. "Algorithmic Accountability Reporting: on the Investigation of Black Boxes", *Columbia Journalism School. Tow Center for Digital Journalism*, 2014. 33 pp. [https://doi.org/10.7916/D8ZK5TW2, accessed on: 18.05.2022].

Floridi, L. "A Proxy Culture", *Philosophy and Technology*, 2015, vol. 28, pp. 487–490.

Garbuk, S.V. “Osobennosti primeneniya ponyatiya ‘doverie’ v oblasti iskusstvennogo intellekta” [The Feature of Using the Concept of “Trust” in the Area of Artificial Intelligence], *Iskusstvennyj intellekt i prinyatie reshenij* [Artificial Intelligence and Decision Making], 2020, no. 3, pp. 15–21. (In Russian)

Gehman, S., Gururangan, S., Sap, M., Choi, Y. & Smith, N.A. “Realtoxicityprompts: Evaluating neural toxic degeneration in language models”, *arXiv preprint arXiv:2009.11462*, 2020 [https://arxiv.org/pdf/2009.11462, accessed on 18.05.2022].

Grzymek, V., Puntschuh, M. *What Europe Knows and Thinks About Algorithms Results of a Representative Survey*. Gütersloh: Bertelsmann Stiftung, 2019. 38 p.

Gurov, P.N. “Opyt social’no filosofskogo osmysleniya problemy ‘fejkov’ i ‘puzyrej fil’trov» v Seti’ [The experience of social and philosophical explanation of the problem of “fakes” and “filter bubbles”], *Problemy sovremennogo obrazovaniya* [Problems of modern education], 2019, no. 3, pp. 9–20. (In Russian)

Hui, Y. *On the Existence of Digital Objects*. Minneapolis, London: University of Minnesota Press, 2016. 336 pp.

“Iskusstvennyj intellekt: ugroza ili vozmozhnost'?” [Artificial intelligence: threat or opportunity?], *VCIOM. Analiticheskij obzor* [Analytical review of the All-Russian Center for the Study of Public Opinion], January 27, 2020 [https://wciom.ru/index.php?id=236&uid=10132, accessed on 23.05.2022]. (In Russian)

Latypova, A.R. “Mezhdu mutaciej i glitchem: cifrovaya evolyuciya media” [Between Mutation and Glitch: Digital Evolution of the Media], *Epistemologiya i filosofiya nauki* [Epistemology & Philosophy of Science], 2020, vol. 57, no. 2, pp. 162–178. (In Russian)

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Yosinski, J. et al. “The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities”, *Artificial life*, 2020, no. 26 (2), pp. 274–306.

Leshkevich, T.G. “Chelovek-virtual i peredacha kul'turnyh cennostej pokoleniyu epohi cifry” [The Virtual Person and Transmitting Cultural Values to the Digital Generation], *Voprosy Filosofii*, 2022, no. 3, pp. 53–63. (In Russian)

Leshkevich, T., Motozhanets, A. “Social Perception of Artificial Intelligence and Digitization of Cultural Heritage: Russian Context”, *Applied Sciences*, 2022, no. 12 (5), p. 2712.

Manovich, L. “Kak sledovat’ za pol’zovatelyami programm?” [How to follow Software Users], *Logos*, 2015, vol. 25, no. 2 (104), pp. 189–218. (In Russian)

Martynenko, T.S., Dobrinskaya D.E. “Social’noe neravenstvo v epohu iskusstvennogo intellekta: ot cifrovogo k algoritmicheskomu razryvu” [Social Inequality in the Age of Algorithms: From Digital to Algorithmic Divide. Monitoring of Public Opinion: Economic and Social Changes], *Monitoring obshchestvennogo mneniya: ekonomicheskie i social’nye peremeny* [Monitoring of Public Opinion: Economic and Social Changes], 2021, no. 1, pp. 171–192. (In Russian)

Mumford, A. *Proxy Warfare*: *War and Conflict in the Modern World*. Cambridge: Polity Press, 2013. 141 pp.

Nikiforova, V.D., Nikiforov, A.A., Viktorova, V.A. “Ekonomicheskie i upravlencheskie aspekty vnedreniya finansovyh tekhnologij” [Economic and Managerial Aspects of Financial Technology Implementation], *Nauchnyj zhurnal NIU ITMO. Seriya Ekonomika i ekologicheskij menedzhment* [Scientific Journal NRU ITMO. Series: Economy and Environmental Management], 2020, no. 4, pp. 99–105. (In Russian)

Nissenbaum, H. “How computer systems embody values”, *Computer*, 2001, vol. 34, no. 3, pp. 118–120.

Pashentsev, E.N. “Zlonamerennoe ispol’zovanie iskusstvennogo intellekta: novye ugrozy dlya mezhdunarodnoj informacionno-psihologicheskoj bezopasnosti i puti ih nejtralizacii” [Malicious Use of Artificial Intelligence: New Threats to International Psychological Security and Ways

to Neutralize Them], *Gosudarstvennoe upravlenie. Elektronnyj vestnik* [Public administration. E-journal], 2019, no. 76, pp. 279–300. (In Russian)

Pinchuk, A.N., Tihomirov, D.A. "O vzaimodejstvii cheloveka i iskusstvennogo intellekta: novaya social'naya real'nost' v predstavlenii moskovskih studentov" [On the Interaction of Human and Artificial Intelligence: a new Social Reality in the Minds of Moscow Students], *Sociologiya i zhizn'* [Sociology and life], 2019, no. 3, pp. 85–97. (In Russian)

Purwanto, P., Kuswandi, K., Fatmah, F. "Interactive Applications with Artificial Intelligence Applications: The Role of Trust among Users", *Foresight and STI Governance*, 2020, vol. 14, no. 2, pp. 64–75.

Schmidt, P., Biessmann, F., Teubner, T. "Transparency and trust in artificial intelligence systems", *Journal of Decision Systems*, 2020, vol. 29, no. 4, pp. 260–278.

Schwab, K. *Chetvertaya promyshlennaya revolyuciya* [The Fourth Industrial Revolution]. Moscow: «E» Publ., 2017. 208 pp. (In Russian)

Von Weizsäcker, E.U., Wijkman, A. *Come On! Capitalism, Short-termism, Population, and the Destruction of the Planet*. N.Y.: Springer, 2018. 220 pp.

Welbl, J, Amelia Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L., Pushmeet, K., Coppin, K., Huang, P. "Challenges in Detoxifying Language Models", *Findings of the Association for Computational Linguistics: EMNLP*, 2021, pp. 2447–2469.

Willson, M. "Algorithms (and the) everyday", *Information, Communication & Society*, 2017, vol. 20, no. 1, pp. 137–150.

Zarsky, T. "The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making", *Science, Technology, & Human Values*, 2015, vol. 41, no. 1, pp. 118–132.

Zednik, C. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence", *Philosophy & Technology*, 2021, no. 34, pp. 265–288.