Философия науки и техники
2022. Т. 27. № 1. С. 34–48
УДК 004.89

Philosophy of Science and Technology
2022, vol. 27, no. 1, pp. 34–48
DOI: 10.21146/2413-9084-2022-27-1-34-48

Т.Г. Лешкевич

Метафоры цифровой эры и Black Box Problem

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 рассмотрены эффекты цифровой эры, драйвером которой является искусственный интеллект. Основная цель состоит в том, чтобы сфокусировать внимание на проблеме «выхода из-под контроля» (Black Box Problem), непрозрачности искусственного интеллекта и возможности его злонамеренного использования. Сопряжены три взаимосвязанных направления. Во-первых, для анализа новообразований, генерируемых цифровизацией, используется потенциал метафор, позволяющих образно обрисовать имеющие место трансформации. Во-вторых, на фоне взрывного роста потребности в высоких технологиях обозначены негативные последствия функционирования искусственного интеллекта, в связи с чем формулируется ряд парадоксов научно-технического прогресса. В-третьих, рассматриваются перспективы технологического симбиоза и фиксируется процесс «конвергенции субъектности», понимаемый как взаимопроникновение естественных возможностей человека и ресурсов интеллектуальных систем. Анализ проведен с учетом отечественной и англоязычной литературы. Автор рассматривает аргументы цифровых алармистов и скептиков. Анализирует метафоры, указывающие на 1) тип современного существования – «лицом в экран» или «лицом в устройство», 2) особенности субъекта цифровой эры – «просмотрщик контента», 3) специфический тип цифровой рациональности – «аренда знания», цифровой мультитаскинг. Делаются выводы о необходимости усиления ответственности (так называемой алгоритмической ответственности), связанной с негативными последствиями использования искусственного интеллекта, и о необходимости расширения рефлексивного анализа, нацеленного на осмысление направленности развития искусственного интеллекта.

***Ключевые слова*:** цифровые метафоры, искусственный интеллект, неконтролируемые последствия, «конвергенция субъектности», «алгоритмическая ответственность»

Цифровые трансформации, выступая мегатрендом современности, направлены на создание когерентной (согласованной) среды будущего. Изменяющийся социотехнологический ландшафт, отражаясь на нормативном уровне включенности в цифровые взаимодействия, свидетельствует не только о расширяющемся проникновении искусственного интеллекта (ИИ) во все сегменты человеческой жизни, но и о тенденции замещения человеческого потенциала ресурсами интеллектуальных систем. Вместе с тем признание развития ИИ в качестве важнейшего приоритета науки не сопровождается столь же активным концептуально-рефлексивным анализом порождаемых им проблем и пределов допустимости его диктатуры. Несмотря на то, что проблематика, в фокусе которой оказывается ИИ, претерпевает уже третью волну бурного к ней интереса, ощутимо запаздывание в осмыслении рисков и возможных неконтролируемых процессов, запущенных его интенсивным внедрением. Одна из центральных проблем современного использования ИИ – *Black Box Problem*, или *проблема выхода из-под контроля*. Учение считают, что многие запрограммированные действия ИИ «непрозрачны», имея строго нормативный характер, они тем не менее представляют своеобразный «черный ящик». 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].

Отметим и то, что происходящие цифровые сдвиги сопровождаются амбивалентными оценками. Громко звучат голоса цифровых алармистов, оглашающих угрозы существованию человека в связи с распространением и совершенствованием технологий ИИ. Их тревожит перспектива вынужденного подчинения цифровому алгоритму, надличностный масштаб контроля и принуждения, зависимость от цифры и ситуация, когда текущая нейронная активность человека становится управляемой, доступной для проверок и вторжений. Негативная реакция столь велика, что близка к призыву: «задраить люки и не допустить вторжение цифрового врага!». Алармистам вторят цифровые скептики, обеспокоенные тем, что контуры общества будущего обусловлены эффектами цифровой детерминации, связаны с технологиями компьютерных симуляций и генерированием «множественного рождения» квазиреальных событий. Опасения вызывают манипулирование сознанием в киберпространстве, цифровое слабоумие, цифровое право и новое цифровое неравенство, маркируемое наличием или отсутствием доступа к сети, уровнем цифровых навыков и компетенций. Как было показано в докладе Нью-йоркского института AI Now (декабрь 2018 г.), одной из острых проблем является «углубление неравенства между теми, кто владеет, и теми, кто не владеет технологиями ИИ, а другой – критичность ошибок для тех, кто становится их жертвой» [Шнуренко, 2018–2019]. Валеологические угрозы связаны с наличием «электромагнитного смога», производимого современными электронными средствами коммуникации [Рахманин, Михайлова, 2014]. Имеет место и негативное воздействие так называемых «аудионаркотиков» – определенных звуковых ритмов, создаваемых с расчетом определенного влияния на потребителя [Андреев, Назарова, 2014]. К широко обсуждаемым темам относятся: цифровая

идентификация, прозрачность приватной жизни, цифровой контроль за поведением человека, где даже мобильные телефоны могут быть использованы для отслеживания траекторий и местонахождения индивида, обнаружения сведений о его склонностях и интересах. Ученые фиксируют, что у современного человека на фоне пандемии развивается ярко выраженная потребность не только в получении информации, но и в субъективном переживании ее избыточности или недостоверности [Солдатова, Рассказова и др., 2021]. Озабоченность цифровых скептиков поддерживается вопросами: зависит ли проект счастливого будущего исключительно от технологий? Как связано социальное и цифровое? Насколько поворот к управлению Большими данными способен обеспечить когерентную среду обитания, если на сегодняшний день правомерен вывод исследователей, допускающий понимание цифровой среды как множества «сетевых семиотических швов»? [Аршинов, Буданов, 2020, с. 110].

На наш взгляд, вклинивающаяся в ткань концептуального анализа метафоризация может послужить весомым дополнением, предлагающим ответы на эпистемологические вопросы о трансформациях цифровой эры, ведь сама цифровая реальность, по мнению исследователей, предстает во многом как математическая метафора [Замков, 2016]. Подчеркнем и то, что в эпоху сложностности (complexity) метафоризация выполняет роль адаптивного научного приема, во многом упрощающего понимание происходящего. Посредством метафоры можно образно представить эффекты цифровых трансформаций, постичь их во множестве самореференций. Оставаясь в поле смыслового измерения, метафора претендует на роль значимого компонента научного анализа, фокусируясь на ощутимых изменениях, вносимых масштабной цифровизацией.

Метафоры цифровой эры как научный прием рационального осмысления ее специфики

Среди порожденных цифровой эрой метафор первенство принадлежит метафоре, указывающей на тип современного существования «лицом в экран» или «лицом в устройство». Эпоха пандемии сделала этот способ существования общепринятым. Между тем как жизнь «лицом в экран» представляет собой совершенно иную практику для центральной нервной системы, порождающую многочисленные когнитивные деформации. Среди них: снижение концентрации внимания, ухудшение кратковременной и долговременной памяти, возрастание психической напряженности и тревожности, деформации мотивационной направленности и неизбежная потеря части богатства и сложности сенсорного опыта. В этом отношении весьма иллюстративен вывод о том, что физическое тело – это запрещенный эффект в Интернете. Погружаясь в онлайн-взаимодействия, человек не только во многом сокращает реальные физические контакты, но и вовлекает себя в новую зону риска, связанную с ситуацией телесного обездвиживания, низкой физической активности, столь присущей образу жизни современно инфомана. Серфинг по Сети и компьютерные интеракции становятся базовыми характеристиками его бытия. Интернет выступает конструктором жизнедеятельности и претендует на то, чтобы составить ядро культивируемых современностью практик. Статистика

ошеломляет: уже в 2019 г. эффект проникновения Интернета в современную жизнь у подростков (16–19 лет) достиг почти 100%, в возрастной категории 20–39 лет этот показатель варьируется от 94% до 97%, в категории 40–49-летних составляет 89%, для людей в возрасте 50–64 лет – 66%, у тех, кто старше 65 лет, достигает 36% [URL: https://www.rbc.ru/ technology and_media/13/01/2020/5e1876549a7947210b5ef636]. Свежие результаты статистических замеров дают значительное увеличение. Согласно данным на начало 2021 г. количество интернет-пользователей в мире выросло до 4,54 млрд человек, что на 7% больше прошлогоднего значения (+298 млн новых пользователей в сравнении с данными на январь 2019 г.). Почти 60% мирового населения уже в онлайн. Причем в России количество интернет-пользователей, по данным Digital 2020, составило 118 млн. Это значит, что Интернетом пользуются 81% россиян. Причем «сидят» в Интернете россияне по 7 часов 17 минут каждый день [https://www.web-canape.ru/business/internet-2020-globalnaya-statistika-i-trendy]. При этом возрастные и гендерные различия фиксируют, что молодежь проводит больше времени в Интернете, чем старшее поколение, а женщины проводят времени меньше, нежели мужчины. На начало 2020 г. в России рекламная аудитория Instagram насчитывала 44 млн человек, что составляет 36% от всего населения страны. Как сообщает App Annie, за прошлые 12 месяцев пользователи смартфонов загрузили более 200 млрд мобильных приложений, потратив совокупно 120 млрд долларов на приложения и покупки в них.

Приведенные данные позволяют говорить о появлении в XXI в. нового типа субъекта [Лешкевич, 2019], сопровождаемом формированием нонхьюман-проблематики. Поэтому вторая метафора указывает на человека, который выступает в статусе «просмотрщика контента» и реже в роли просьюмера – создателя собственного контента, что никак не уменьшает, а, напротив, увеличивает необходимость просмотра обновлений, сделанных другими. Серфинг по просторам Сети, которым так увлечен наш современник, хотя и мотивирован установкой на получение немедленной реакции на возникший запрос, на самом деле протекает, опираясь на гиперссылки, и представляет собой перескок с одной темы на другую, рождая постоянное утомление и усталость. Когнитивные акты, приобретая клиповый характер, генерируют эффект «коллажа». Отсутствует целостное восприятие содержания, понимание причинно-следственных связей и зависимостей. Феномен чтения становится техникой пропуска лишнего. «Просмотрщики контента» производят лишь поверхностное сканирование информации, не переводя ее в личностное знание. Индивид, будучи «просмотрщиком контента», пребывает по большей части в дорефлексивном, безрефлексивном или не сопряженном с высокой степенью рефлексии состоянии. Интернет-язык с ограниченным набором типичных социальных сигналов, но тем не менее ставший распространенным средством общения, относится специалистами к предрациональному – нижнему, пограничному уровню сознательности. Он включает в себя мемы, смайлики, флейминг, ожесточенные войны интернет-сообществ и их представителей.

Подчеркнем, однако, что функция просмотра контента и функция осмысления существенно различны, и в настоящее время самосознание массового

представителя цифровой эры не является социально значимым и востребованным запросом. Используя вывод Н. Лумана, можно сказать, что мир стал ареной коммуникативных процессов, из которых исчезают рефлексия и понимание [Луман, 2005]. Массмедиа, навязывая фреймы понимания ситуации, возлагают на себя квазирефлексивную миссию. Вследствие чего можно констатировать, что уровень рефлексии замещается интенсивностью инфомесседжей, которые сопровождаются эмоциональной реакцией, созданием квазисобытий, подчиняющим факты манипулятивным стратегиям либо случайными направлениями инет-активности [Лешкевич, Мотожанец, Катаева, 2020].

Теоретико-познавательный срез цифровых трансформаций обозначен метафорой «аренда знания», указывающей на специфический тип рациональности, связанной с использованием готовых информационных ресурсов и компиляцией контента Сети. Имеется в виду ситуация, когда индивид, обладая цифровыми навыками, находит содержащуюся в Сети нужную информацию, присваивает ее себе и выдает за эквивалент собственных умственных способностей. Эффект «аренды знания» или «заимствованных знаний» отражает несоответствие между высокоразвитыми цифровыми навыками, которыми характеризуется молодое поколение, и их неспособностью освоить концептуальную основу культурного наследия, обусловленную тем, что находящаяся в Сети информация автоматически личностным знанием стать не может. Здесь возникает сложная междисциплинарная проблема: как индикацию информации обратить в качество внутреннего опыта индивида? Причем заимствование контента Сети, т.е. «аренда знаний», предельно снижает осмысленность и рефлексивность и не способствует усилению мотивации к познанию. В ситуации, когда индивид передает огромную часть познавательной нагрузки поисковым системам, рискогенным оказывается, во-первых, то, что представители «поколения Гугл», как отмечают исследователи, запоминают не саму суть изучаемого вопроса, а путь к информации [Sparrow, Liu, Wenger, 2011]. Во-вторых, сам контент Сети может включать в себя ложные домыслы, псевдознания и сведения.

Тем не менее на сегодняшний день следует зафиксировать активно протекающий процесс «конвергенции субъектности», объединяющий естественные способности человека с возможностями и ресурсами Сети, фиксирующий их «гипервзаимосвязанность». Современник воспринимает «умные» устройства как интеллектуальных партнеров, в том числе с функцией трансактивной памяти, способствующей хранению информации на внешнем носителе. Это указывает на так называемый распределенный интеллект, поддерживаемый размещением новейших технологических устройств как на теле, так и внутри его.

Радикальная степень конвергенции современного человека и цифровых технологий отражена метафорой “Homosolus” (человек одинокий) или указанием на «хикикомори». Интерпретируя суть данных метафор, заметим, что масштабная цифровизация рождает образ одинокого, сосредоточенно нажимающего на клавиатуру человека, предпочитающего добровольную изоляцию. Поглощенный «черной дырой» монитора Homosolus в реальности оказывается «совершенно замкнутой в себе субъективностью». Цифровая жизнь, предоставляя симулятивные аналоги, ведет к отчужденности от реальных взаимодействий. Японский ученый Т. Саито, вводя термин «хикикомори», подчеркивал,

что образ жизни, отличающий хики, и сфера его самореализации ограничены исключительно киберпространством. И хотя в момент выхода книги популяция хикки составляла около 2 млн человек, ученый прогнозировал, что в будущем ее количество превысит 10 млн, а «синдром хикикомори» захватит весь мир. В настоящее время эта проблема получила название «проблемы 8050», так как родители, содержащие на иждивении 50-летних детей-хики, вступают в 80-летний возраст [Saito, 2013].

Заслуживает внимания метафора мультитаскинга, характеризующая современного человека в его стремлении использовать цифровые технологии для сопряжения различных режимов деятельности (онлайнового и офлайнового, досугового и профессионального). По мнению ученых, цифровая многозадачность – это не только способность совмещать работу с потреблением развлекательного контента, но и сама психологическая готовность переключаться с одного вида деятельности на другой, «пропускать через себя» разнонаправленные информационные потоки. По большей части цифровая многозадачность свойственна представителям молодого поколения [Roubal, 2015].

Обозначенный ряд метафор может быть дополнен метафорой «гибридного мира», объединяющего телесно-материальную и цифровую реальность как пространства одновременной принадлежности современного человека. Парадокс состоит в том, что потребности, ценности и смыслы «здесь и теперь» существующего индивида должны быть привязаны, с одной стороны, к непосредственному «наличному бытию» (запретить или отменить которое невозможно), а с другой – к возможностям и перспективам интеграции человека и технологий. Гибридный мир генерирует прессинг двойных стандартов, идущих как со стороны цифровых технологий, так и со стороны не потерявших свою силу традиционных регламентаций доцифровой эры; выдвигает удвоенные требования к формированию навыкообразующих форм человеческого опыта. Но по мере того, как возможности ИИ становятся все более мощными, его функционирование охватывается беспокойной подозрительностью; все чаще поднимается проблема, связанная с его уязвимостью и опасностью вредоносного использования.

Искусственный интеллект и Black Box Problem (проблема «выхода из-под контроля»)

Гонка в развитии ИИ делает очевидным тот факт, что ИИ выполняет своеобразную функцию демаркации, разделяя высокотехнологичные страны и те, которые не обладают технологиями ИИ. Мониторинг ресурсов позволяет обнаружить отличие направлений развития и использования ИИ в различных государствах. Так, значимым направлением в Китае является использование ИИ для масштабной системы фильтрации сетевых сообщений и передаваемых данных, мониторинга сетевых взаимодействий, прогнозирования митингов и манифестаций. В отношении Франции существует информация о стремлении нарастить кибервойска. Команда швейцарских ученых ведет масштабный международный проект по моделированию человеческого мозга с целью синтезирования всех знаний в единую полноценную карту активности мозга.

Возможности США в развитии информационных технологий, в том числе и информационной разведки, превосходят все остальные страны. Олигархи кремниевой долины фактически «курируют» Интернет. В России согласно Указу Президента РФ «О развитии искусственного интеллекта в РФ» 2019 г. заявленные цели, включающие в себя обеспечение роста благосостояния и качества жизни населения, обеспечение национальной безопасности и правопорядка, достижение устойчивой конкурентоспособности российской экономики, в том числе лидирующих позиций в мире в области ИИ, носят исключительно гуманный характер [Указ Президента РФ, 2019].

В этом контексте представляется весьма корректным определить ИИ как способность системы приобретать, обрабатывать и применять знания, где в объем понятия «знание» входят факты, информация и навыки, приобретенные в результате опыта или обучения. Система ИИ – это техническая система, которая используется для решения проблем [Боргест, 2019]. Согласно позиции В.К. Финна, строение интеллектуальных систем исчерпывается следующей структурой: (база знания + база фактов) + Решатель задач + Интерфейс интеллектуальной системы [Финн, 2009]. Однако отметим, что в противовес так понимаемому ИИ еще Ж. Пиаже определял интеллект человека как «прогрессирующую обратимость мобильных психических структур», как состояние равновесия, к которому тяготеют все последовательно расположенные адаптации сенсорно-моторного и когнитивного порядка, так же как и все ассимилятивные и аккомодирующие взаимодействия организма со средой [Пиаже, 1994]. Исходя из подобных заключений, становится очевидным, что перевести биологическую и психофизиологическую «элементную базу» когнитивных функций в «цифру» невозможно!

На фоне тенденции к неограниченному расширению сфер применения ИИ и обеспечения процесса принятия жизненно важных решений посредством обработки большого объема данных ИИ подпадает под алармистские настроения. Все настойчивее обозначается Black Box Problem – проблема «выхода из-под контроля». По мнению Дж. Баррата, при существующей демонстрации мощи и сложности ИИ весьма реальны опасения, связанные с тем, что ИИ будет вести себя непредсказуемо и непостижимо. Непредсказуемость будет сочетаться со случайностями, которые проистекают из-за сложности устройства. Вероятно и то, что поведение ИИ окажется несовместимым с нашим выживанием [Баррат, 2015]. Автор предлагает читателям поставить себя на место ИСИ – Искусственного Суперинтеллекта, который в тысячу раз умнее человека, решает задачи в миллиарды и триллионы раз быстрее человека, – с тем, чтобы понять, что ИСИ «захочет» получить доступ к энергии в той форме, которую ему удобнее всего использовать, «захочет» улучшить себя и «не захочет», чтобы его выключали или портили. Вполне допустимо, что ИСИ будет искать способы выйти за пределы охраняемого помещения, чтобы получить лучший доступ к ресурсам, при помощи которых он сможет защитить и усовершенствовать себя [Там же, с. 2]. Это доказывает, что развитие ИИ нуждается в мерах, четко определяющих технологии управления искусственным интеллектом.

Действительно, в современных условиях своевременная оценка рисков, обусловленных стремительным научно-технологическим прогрессом в области

ИИ, выдвигается в число приоритетных. Д. Сайклбек посвятил раздел своей книги описанию того, как воспринимаются возможные опасности ИИ [Cycleback, 2018]. Реальными угрозами, как отмечает автор, могут быть ошибки в программах, о которых программисты могут и не знать, а также то, что ИИ может начать действовать вопреки желаниям программистов, принимая неверные решения в наиболее важных ситуациях. Последствия, с которым может столкнуться человек при использовании ИИ в низменных целях, выражаются в кибератаках, авариях автономных транспортных средств, распространении Интернет-вирусов, «ботах» для социальных сетей, несанкционированном использовании персональных данных, угрозах личной безопасности и информационно-психологических угрозах, влияющих на сознание и поведение людей, и пр. Примечательно, что в докладе Римского клуба в отношении проблем цифровизации современного мира отмечается разрушительный характер данного процесса: «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].

Возможность того, что ИИ, используя собственные преимущества, станет корректировать, изменять себя и начнет действовать злонамеренно, вызывает особые опасения. Следует подчеркнуть, что по сравнению с программной уязвимостью использование искусственного интеллекта с преднамеренным вредом и иллюстрирующие эти инциденты ситуации выделяются в отдельное проблемное поле. Исследователи обращают внимание на то, что злонамеренное использование ИИ может иметь краткосрочные, среднесрочные и дальносрочные последствия. Предлагаются следующие варианты классификации злонамеренного использования искусственного интеллекта: по территориальному охвату (местный, региональный, глобальный), по степени наносимого ущерба (незначительный, значительный, крупный, катастрофический), по скорости распространения (медленный, быстрый, стремительный), по форме распространения (открытый, скрытый) [Пашенцев, 2019, с. 284].

Вследствие того, что нюансы процесса принятия решений понять все сложнее, остро встает вопрос: как сформировать и внедрить в искусственный интеллект алгоритм дружественного отношения к человеку? И этот вопрос звучит тем острее, чем понятней противоречие, состоящее в том, что основания алгоритмов принципиально формализуемы, а качество дружественности кодом формального алгоритма вряд ли может быть ухвачено. Проблемой остается и то, сможет ли область машинного зрения ИИ ориентироваться в условиях беспорядка в стихии мира людей и современной жизни. Иными словами, возможны ли корреляции между идеализированным пространством вычислений и беспорядочной реальностью с ее онтологической неопределенностью? А то, что ИИ, по заявлениям ученых, обладает «существенной эпистемологической непрозрачностью» [Humphreys, 2009, p. 618], является еще одним камнем преткновения. Совершенно очевидно, что для всей проблематики ИИ важно наличие ясности и информационной прозрачности. Право иметь «значимую информацию о логике» решения проблемы, «право на объяснение» воспринимается как основополагающее.

Вместе с тем непрозрачность ИИ на сегодняшний день оценивается как свойство, преодолеть которое невозможно.

Black Box Problem усиливает критическое отношение к ИИ, порождает стремление к ограничению диапазона его применения, особенно в ситуациях, когда на него возлагаются функции основного агента человеческой жизни. Все чаще акцентируется необходимость разработать технологии аварийного выключения ИИ. Тем не менее следует отметить, что на фоне Black Box Problem широко распространена альтернативная практика, связанная с доверием интеллектуальным системам, онлайн-платформам и веб-сайтам, которые, принимая запросы, бескорыстно и беспристрастно действуют от имени пользователя. Такой тип взаимодействий, инициированных от имени «кого-то», назван прокси-культурой [Floridi, 2015]. Термином «прокси» обозначают ситуацию, когда информационная система действует «от лица» реального субъекта, причем контакт с реальным субъектом необязателен, т.к. значимой является компьютерная алгоритмизация. В настоящее время на основе технологий ИИ создано множество «умных» помощников с расширенными возможностями интерактивного взаимодействия. 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 [Moar, 2019]. По прогнозам корпорации Huawei, к 2025 г. в мире будет насчитываться свыше 40 млрд личных «умных» устройств, а у 90% пользователей устройств будут «умные» цифровые помощники (https://integral-russia.ru/2019/09/20/gosudarstvennoe-upravlenie-i-iskusstvennyj-intellekt-istoriya-i-perspektivy/).

В связи с этим представляется правомерным введение «алгоритмической ответственности», которая должна быть сфокусирована на решениях и их последствиях, принятых на основе алгоритма. Имеются в виду ситуации, когда лица, принимающие решения, всецело полагаются на результаты автоматизированной системы. Ссылаясь на то, что решения формируются на основе технологической обработки огромных объемов данных, они снимают с себя какую-либо ответственность. Совершенно очевидно, что для обеспечения как индивидуального, так и общественного блага введение «алгоритмической ответственности» весьма значимо. Действенно и всяческое поощрение культуры ответственности. В контексте «алгоритмической ответственности» большое значение имеет ответственное раскрытие уязвимостей ИИ и инструментов безопасности.

Парадокс научно-технического прогресса и перспектива технологического симбиоза

Негативные последствия для рынка труда в отношении рабочих мест, которые активно замещаются интеллектуальными системами, вытесняющими людей, различимы уже сейчас. Однако, во-первых, считается, что интеллектуальная система может куда качественнее справиться с исполнением однотипных функций, в то время как индивиду не всегда хватает специальных навыков в человеко-машинном взаимодействии. Во-вторых, применение технологий, как полагают, будет способствовать тому, чтобы сделать нашу жизнь более

независимой от субъективных предпочтений и противоречивой предвзятости. Масштабы предполагаемых трансформаций велики. Согласно некоторым прогнозам, к 2030 г. от 75 до 375 млн людей (от 3 до 14% мировой рабочей силы) окажутся вынуждены сменить сферу деятельности из-за того, что занимаемые ими рабочие места будут автоматизированы [Кловайт, Ерофеева, 2019, с. 59]. Это наталкивает на очередной парадокс цифровой эры, когда стремительное развитие научно-технического прогресса ведет к «опустошению» рынка труда и выбросу на улицу «лишних» людей, тем самым девальвируя фундаментальное понимание предметно-деятельной сущности человека. Любопытно и то, что в отношении интеллектуальных систем введен термин «условный сотрудник». Предрекается скорая смена традиционного типа управления управлением при помощи нейросетей. Но какая же участь уготована современному человеку, когда в конкурентную борьбу вступят интеллектуальные системы? В этой ситуации на передний план выходят проблемы регулирования развития ИИ, равно как и контроль за его совершенствованием. Логично предположить, что современная наука и институциональная мысль должны объединить свои усилия для того, чтобы найти и установить баланс между развитием интеллектуальных систем и количеством рабочих мест.

Тем не менее развертка моделей будущего в связи с использованием ИИ имеет различные конфигурации. В этой связи обращает на себя внимание исследование Р. Ферс и Э. Робинсон, в рамках которого выделено шесть типов социального отношения к будущему человеко-машинного взаимодействия на основании предложенных авторами критериев оптимизма/пессиммизма/ассамбляжности, а также стратегичности и гуманистичности. Движение человеческой цивилизации к тому, что они называют «роботопиями», авторы сопровождают следующими уточнениями: «Отношения между человеком и роботом могут быть черной сердцевиной нашего времени, с чертами из разных моделей – доброжелательными инструментами гуманистов-оптимистов, опасными бесчеловечными системами гуманистов-пессимистов, застывшим трудом гуманистов-стратегов и крайними другими теориями сборки – объединенными пока непостижимыми способами» [Firth, Robinson, 2021, p. 309]. Таким образом, будущее сопряжено со всей сложностью, неоднозначностью и неопределенностью человеко-машинного взаимодействия. Весьма экстравагантной, на наш взгляд, является позиция, отстаивающая симбиоз человека и технологий как новую счастливую эру в истории человечества. Как отмечают комментаторы, британский философ Д. Пирс откровенно выражает свою глубокую веру в «Три С-Цивилизацию»: Суперинтеллекта, Супердолголетия и Суперсчастья [Чеклецов, 2021]. Д. Пирс уверен, что генная инженерия с нанотехнологией избавят от страданий всю разумную жизнь, освободят весь живой мир от неприятных переживаний. Однако вариант «голова с проводами» – вариант, который наиболее часто ассоциируется с внутричерепной самостимуляцией, будет лишь одним из пунктов большого «меню». Натуралистический рай может, по мнению автора, быть реализован биотехнологическими средствами [Пирс, 2020, с. 11].

Отметим, что Д. Пирс не единственный приверженец проекта развития, стимулированного сверхвозможностями ИИ. Верой в успешное человеко-машинное слияние пронизана доктрина Э. Кларка. Логика его рассуждений такова.

Поскольку сложная культурная среда обитания человека предстает по большей части как технологическая, то разум в стремлении адаптироваться к ней расширяет свои возможности, прибегая к интеллектуальным вычислительным устройствам. Коалиция с артефактом есть знак человеческого интеллекта, демонстрирующего «преобразующий потенциал этой коалиции» [Clark, 2004, p. 22]. Представляя направление «нейронного конструктивизма», ученый приветствует человеко-ориентированные технологии, считая, что такие технологии будут более походить на часть психического аппарата человека, нежели на внешние инструменты. «Умный мир» будет функционировать в такой тесной гармонии с биологическим мозгом, что проведение границы не будет служить ни юридическим, ни моральным, ни социальным целям [Ibid., p. 30]. К подобному выводу приходит и специалист в области физики М. Каку. Обладая ярко выраженным гуманитарным взглядом на проблему соотнесения «человеческого» и «компьютерного», он заключает: «Если вы спросите доктора Брукса, как человек может сосуществовать с суперумными роботами, он откровенно ответит: мы с ними сольемся» [Каку, 2015]. Таким образом, энтузиасты искусственного интеллекта и искусственных интеллектуальных систем уверены, что при всем превосходстве над человеческим мозгом их интеллектуальная мощь не может быть аморальна.

Резюмирующие замечания

В современной ситуации, когда постоянные трансформации становятся нормой, выживают, говоря языком П. Друкера, только лидеры перемен – те, кто чутко улавливают тенденции изменений и мгновенно приспосабливаются к ним, используя себе во благо открывающиеся возможности [Друкер, 2012]. В этих условиях тематика, связанная с анализом негативных последствий развития ИИ, требует неотступного и опережающего сопровождения социогуманитарной рефлексией. Использование ИИ, свидетельствующего о новой ступени социотехнологической эволюции, порывающей со своим неотцифрованным прошлым, должно быть поставлено под контроль человеческого разума. Философско-концептуальный анализ современной ситуации выявляет ряд парадоксов. Первый, экзистенциальный, заключается в двойственной принадлежности человека, состоящей в том, что, с одной стороны, потребности, ценности и смыслы «здесь и теперь» существования обусловлены наличным физическим бытием, а с другой – они должны быть привязаны к возможностям и перспективам цифровой реальности. Второй парадокс, технологический, показывает, что, несмотря на фиксацию Black Box Problem, а также проблему злонамеренного использования искусственного интеллекта, степень доверия к ИИ и интеллектуальным системам растет, невзирая на имеющий место эффект быстрого программного устаревания. Третий парадокс – парадокс прогресса – фиксирует, что стремительное научно-техническое развитие ведет к «опустошению» рынка труда и замещению человека. В противостоянии «слепому» развитию технологий современная наука и институциональная мысль должны объединить свои усилия для установления баланса между человеческим потенциалом и развитием интеллектуальных систем и ИИ.

Список литературы

Андреев, Назарова, 2014 – *Андреев И.Л., Назарова Л.Н.* Эволюция психического ландшафта информационной эпохи // Психическое здоровье. 2014. № 7 (98). С. 74–80.

Аршинов, Буданов, 2020 – *Аршинов В.И., Буданов В.Г.* Социотехнические ландшафты в оптике семиотически-цифровой сложностности // Вопр. философии. 2020. № 8. С. 106–116.

Баррат, 2015 – *Баррат Дж.* Последнее изобретение человечества. Искусственный интеллект и конец эры Homo sapiens. М.: Альпина нон-фикшн, 2015. 330 с.

Боргест, 2019 – *Боргест Н.М.* Стратегии интеллекта и его онтологии: попытка разобраться // Онтология проектирования. 2019. Т. 9. № 4 (34). С. 407–428.

Друкер, 2012 – *Друкер П.Ф.* Менеджмент. Вызовы XXI века. М.: Манн, Иванов и Фербер, 2012. 276 с.

Замков, 2016 – *Замков А.В.* Цифровая реальность как математическая метафора // Вестник Волжского университета имени В.Н. Татищева. 2016. Т. 2. № 4. С. 176–184.

Каку, 2015 – *Каку М.* Будущее разума. М.: Альпина-нон-фикшн, 2015. 500 с.

Кловайт, Ерофеева, 2019 – *Кловайт Н., Ерофеева М.* Работа в эпоху разумных машин: зарождение невидимой автоматизации // Логос. 2019. Т. 29. № 1. С. 53–80.

Лешкевич, 2019 – *Лешкевич Т.Г.* Цифровые трансформации эпохи в проекции их воздействия на современного человека // Вестник ТГУ. 2019. № 439. С. 103–109.

Лешкевич, Мотожанец, Катаева, 2020 – *Лешкевич Т.Г., Мотожанец А.А., Катаева О.В.* Цифровая детерминация и трансформации смысложизненной рефлексии. Ростов н/Д; Таганрог: Изд-во Южного федерального университета, 2020. 196 с.

Луман, 2005 – *Луман Н.* Реальность массмедиа. М.: Праксис, 2005. 256 с.

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. Вып. 76. Октябрь. С. 279–300.

Пиаже, 1994 – *Пиаже Ж.* Избранные психологические труды. М.: Международная педагогическая академия, 1994. 674 с.

Рахманин, Михайлова, 2014 – *Рахманин Ю.А., Михайлова Р.И.* Окружающая среда и здоровье: приоритеты профилактической медицины // Гигиена и санитария. 2014. № 5. С. 5–10.

Солдатова, Рассказова, Неяскина, Ширяева, 2021 – *Солдатова Г.У., Рассказова Е.И., Неяскина Ю.Ю., Ширяева О.С.* Потребность в информации и отношение к цифровым технологиям как факторы критичного и некритичного распространения новостей о пандемии // Вестник Московского Университета. Серия 14. Психология. 2021. № 1. С. 170–195.

Указ Президента РФ, 2019 – Указ Президента Российской Федерации от 10.10.2019 г. № 490. О развитии искусственного интеллекта в Российской Федерации. URL: http://www.kremlin.ru/acts/bank/44731 (дата обращения: 11.02.2022).

Финн, 2009 – *Финн В.К.* Искусственный интеллект // Энциклопедия эпистемологии и философии науки. М.: Канон +, 2009. С. 316–318.

Чеклецов, 2021 – *Чеклецов В.В.* Диалоги гибридного мира // Философские проблемы информационных технологий и киберпространства. 2021. № 3 (19). С. 99–116.

Шнуренко, 2018–2019 – *Шнуренко И.* Искусственный интеллект на грани нервного срыва // Эксперт. 2018–2019. № 1–3. С. 38–41.

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

Cycleback, 2009 – *Cycleback D.* Philosophy of Artificial Intelligence: A Critique of the Mechanistic Theory of Mind. Florida: Universal-Publishers BocaRaton, 2009. 190 p.

Firth, Robinson, 2021 – *Firth R., Robinson A.* Robotopias: mapping utopian perspectives on new industrial technology // International Journal of Sociology and Social Policy. 2021. No. 41 (3/4). P. 298–314.

Floridi, 2015 – *Floridi L.* A Proxy Culture // Philosophy and Technology. 2015. Vol. 28. No. 4. P. 487–490.

Humphreys, 2009 – *Humphreys P.* The philosophical novelty of computer simulation methods // Synthese. 2009. Vol. 169. P. 615–626.

Moar, 2019 – *Moar J.* The Digital Assistants of Tomorrow. White Paper. Basingstoke (UK): Juniper Research Ltd., 2019.

Pearce, web – *Pearce D.* The Hedonistic Imperative. URL: https://www.hedweb.com/hedethic/tabconhi.htm (дата обращения: 11.02.2022).

Roubal, 2015 – *Roubal O.* Fast-Time Digital Age and Lifestyle Changes // Marketing identity: Digital Life, Part II, Conference Proceedings from International Scientific Conference 10th–11th November 2015. Trnava: Publishing house of Michal Vaško, Prešov, Slovak Republic, 2015. P. 206–219.

Saito, 2013 – *Saito T.* Hikikomori: Adolescence Without End. Minneapolis, MN: University of Minnesota Press, 2013. 216 p.

Sparrow, Liu, Wenger, 2011 – *Sparrow B., Liu J., Wenger D.M.* Google effects on memory: Cognitive consequences of having information at our fingertips // Science. 2011. Vol. 333. No. 6043. P. 776–778.

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.

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

Metaphors of the digital age and the Black Box Problem

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 effects of the digital era, the driver of which is AI. The main goal is to focus on the Black Box Problem, “opacity of AI” and the possibility of Malicious Use of Artificial Intelligence. Three interconnected directions are interfaced. Firstly, in the context of the analysis of the digital age, the potential of metaphors is used, which makes it possible to describe digital transformations figuratively. Secondly, due to the growing demand for high technologies, the negative consequences of using AI are considered and a number of paradoxes of scientific and technological progress are formulated. Thirdly, the article examines the widespread practice of trust in intelligent systems, as well as the prospects for technological symbiosis. The analysis is based on the Russian and English-language literature. The author analyzes metaphors that indicate the type of modern existence – “face-to-screen” or “face-to-device” (1); features of the subject of the digital age – “content viewer” (2); the specifics of digital rationality – “knowledge rent”, digital multitasking (3). Attention is drawn to the process of “convergence of subjectivity”. The issue of malicious use of AI is discussed. The author draws conclusions about the need for “algorithmic responsibility” and expanding the field of reflective analysis aimed at studying the consequences of using AI.

Keywords: digital metaphors, artificial intelligence, uncontrollable consequences, “convergence of subjectivity”, “algorithmic responsibility”

References

Andreev, I.L, Nazarova, L.N. "Evolyuciya psihicheskogo landshafta informacionnoj epohi" [The Evolution of the Mental Landscape of the Information Age], *Mental health / Psihicheskoe zdorov'e*, 2014, no. 7 (98), pp. 74–80. (In Russian)

Arshinov, V.I., Budanov, V.G. "Sociotekhnicheskie landshafty v optike semioticheski-cifrovoj slozhnostnosti" [Sociotechnical landscapes in optics of semiotic-digital complexity], *Voprosy Filosofii,* 2020, vol. 8, pp. 16–116. (In Russian)

Barrat, J. *Poslednee izobretenie chelovechestva. Iskusstvennyj intellect i konec ery Homo sapiens* [Our Final Invention: Artificial Intelligence and the End of the Human Era]. Moscow: Izdatel'stvo "Al'pina non-fikshn" Publ., 2015. 330 pp. (In Russian)

Borgest, N.M. "*Strategii intellekta i ego ontologii: popytka razobrat'sya"* [Strategies of intelligence and its ontology: an attempt to understand], *Ontology of Designing / Ontologiya proektirovaniya*, 2019, vol. 9, no. 4 (34), pp. 407–428. (In Russian)

Cheklecov, V.V. "Dialogi gibridnogo mira" [Dialogs of a hybrid world], *Philosophical problems of information technology and cyberspace / Filosofskie problem informacionnyh tekhnologij i kiberprostranstva*, 2021, no. 3 (19), pp. 99–16. (In Russian)

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

Cycleback, D. *Philosophy of Artificial Intelligence: A Critique of the Mechanistic Theory of Mind.* Florida: Universal-Publishers Boca Raton, 2009. 190 pp.

Druker, P.F. *Menedzhment. Vyzovy XXI veka* [Management. Challenges for the 21st Century]. Moscow: Izdatel'stvo Mann, Ivanov i Ferber Publ., 2012. 276 pp. (In Russian)

Finn, V.K. "Iskusstvennyj intellekt" [Artificial Intelligence], *Encyclopedia of Epistemology and Philosophy of Science / Enciklopediya epistemologii i filosofii nauki*. Moscow: Kanon +, 2009, pp. 316–318. (In Russian)

Firth, R., Robinson, A. "Robotopias: mapping utopian perspectives on new industrial technology", *International Journal of Sociology and Social Policy,* 2021, no. 41 (3/4), pp. 298–314.

Floridi, L. "A Proxy Culture"*, Philosophy and Technology*, 2015, vol. 28, no. 4, pp. 487–490.

Humphreys, P. "The philosophical novelty of computer simulation methods", *Synthese*, 2009, vol. 169, pp. 615–626.

Kaku, M. *Budushchee razuma* [The Future of the mind]. Moscow: Al'pina-non-fikshn, 2015. 500 pp. (In Russian)

Klovajt, N., Erofeeva, M. "*Rabota v epohu razumnyh mashin: zarozhdenie nevidimoj avtomatizacii*" [Work in the Age of Intelligent Machines: The Rise of Invisible Automation], *Logos*, 2019, vol. 29, no. 1, pp. 53–80. (In Russian)

Leshkevich, T.G. "Cifrovye transformacii epohi v proekcii ih vozdejstviya na sovremennogo cheloveka" [Digital Transformation of the Era in the Projection of Their Impact on the Modern Man], *Tomsk State University Journal / VestnikTomskogo gosudarstvennogo universiteta*, 2019, no. 439, pp. 103–109. (In Russian)

Leshkevich, T.G., Motozhanets, A.A. Kataeva, O.V. *Cifrovaya determinaciya i transformacii smyslozhiznennoj refleksii* [Digital determination and transformation of meaningful reflection]. Rostov-na-Donu; Taganrog: Izdatel'stvo Yuzhnogo federal'nogo universiteta, 2020. 196 pp. (In Russian)

Luman, N. *Real'nost' massmedia* [The Reality of Massmedia]. Moscow: Praksis, 2005. 256 pp. (In Russian)

Moar, J. *The Digital Assistants of Tomorrow. White Paper.* Basingstoke (UK): Juniper Research Ltd., 2019.

Pashencev, 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], *Pablic administration E-jornal / Gosudarstvennoe upravlenie. Elektronnyj vestnik*, 2019, no. 76, pp. 279–300. (In Russian)

Pearce, D. *The Hedonistic Imperative.* [https://www.hedweb.com/hedethic/tabconhi.htm, accessed on 11.02.2022].

Piaget, J. *Izbrannye psihologicheskie trudy* [Selected psychological works]. Moscow: Mezhdunarodnaya pedagogicheskaya akademiya, 1994. 674 pp. (In Russian)

Rahmanin, Yu.A., Mihajlova, R.I. "Okruzhayushchaya sreda I zdorov'e: prioritety profilakticheskoj mediciny" [Environment and Health: Priorities for Preventive Medicine], *Hygiene and sanitation / Gigiena i sanitariya*, 2014, no. 5, pp. 5–10. (In Russian)

Roubal, O. "Fast-Time Digital Age and Lifestyle Changes", *Marketing Identity: Digital Life, Part II, Conference Proceedings from International Scientific Conference* 10th–11th November 2015. Trnava: Publishing house of Michal Vaško, Prešov, Slovak Republic, 2015, pp. 206–219.

Saito, T. *Hikikomori: Adolescence Without End,* Minneapolis, MN: University of Minnesota Press, 2013. 216 pp.

Shnurenko, I. "Iskusstvennyj intellekt na grani nervnogo sryva" [Artificial intelligence on the verge of a nervous breakdown], *Expert / Ekspert*, 2018–2019, no. 1–3, pp. 38–41. (In Russian)

Soldatova, G.U., Rasskazova, E.I., Neyaskina, Yu.Yu., Shiryaeva, O.S. "Potrebnost' v informacii i otnoshenie k cifrovym tekhnologiyam kak factory kritichnogo i nekritichnogo rasprostraneniya novostej o pandemii" [The Need for Information and the Attitude towards Digital Technologies as Factors of Critical and Uncritical Dissemination of Pandemic News], *Moscow University Psychology Bulletin / Vestnik Moskovskogo Universiteta. Seriya 14. Psihologiya,* 2021, no. 1, pp. 170–195. (In Russian)

Sparrow, B., Liu, J., Wenger, D.M. "Google effects on memory: Cognitive consequences of having information at our fingertips", *Science*, 2011, vol. 333, no. 6043, pp. 776–778.

Ukaz Prezidenta Rossijskoj Federacii ot 10.10.2019 g. no. 490. "O razvitii iskusstvennogo intellekta v Rossijskoj Federacii" [On the development of artificial intelligence in the Russian Federation] [Digital source]. URL: http://www.kremlin.ru/ acts/bank/44731 (In Russian)

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

Zamkov, A.V. "Cifrovaya real'nost' kak matematicheskaya metafora" [Digital Reality as Mathematical Metaphor], *Vestnik of Volzhsky University after V.N. Tatishcheva / Vestnik Volzhskogo universiteta imeni V.N. Tatishcheva*, 2016, vol. 2, no. 4, pp. 176–184. (In Russian)

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