УДК 001

Философия искусственного интеллекта: рамки этичности создания и эксплуатации цифровых двойников людей

Кидяев Вячеслав Вячеславович

Аспирант,
Российский экономический университет им. Г.В. Плеханова,
главный эксперт Управления по коммуникационной работе
и международной деятельности
Национального центра развития искусственного интеллекта
при Правительстве Российской Федерации,
109028, Российская Федерация, Москва, Покровский бульвар, 11;
e-mail: Slavakid@yandex.ru

Аннотация

Технологии искусственного интеллекта открыли для человечества спектр уникальных возможностей в ключевых областях: автоматизации, безопасности, здравоохранения, анализа информации и т.д. Однако за динамичным развитием технологий не успевает развитие философских идей и концепций. Некоторые ИИ-разработки, такие как создание цифровых копий умерших людей, создают этические дилеммы быстрее, чем человечество находит на них ответы, провоцируя риски и снижая вероятность безопасного и комфортного сосуществования естественного интеллекта с искусственным. Создание доверенного искусственного интеллекта, полностью подчиненного интересам человека и контролируемого им, является сегодня недостижимой целью, которая откладывается на все больший срок из-за отсутствия межотраслевого подхода компаний – разработчиков ИИ-моделей. Сегодня этика разработчика и этика разработки – тождественные понятия, вытекающие одно из другого, но можно ли распространить это убеждение на цифровых двойников, главная цель которых – быть копией конкретного человека? Определение границ прав людей и цифровых копий позволит избежать существенных рисков, связанных с защитой памяти умерших, авторскими правами, безопасностью персональных данных и другими аспектами. Остановить создание новых, более совершенных цифровых копий людей невозможно, но необходимо уже сейчас создать фундамент для безболезненной интеграции этих ИИ-решений в реальность. Сделать это можно с помощью законодательного регулирования, четкого разграничения прав на создание и использование цифровых копий, обучения ботов на наиболее достоверных базах данных, а также с помощью привлечения к разработке специалистов по антропологии, социологии, этике и, конечно, философии.

Для цитирования в научных исследованиях

Кидяев В.В. Философия искусственного интеллекта: рамки этичности создания и эксплуатации цифровых двойников людей // Контекст и рефлексия: философия о мире и человеке. 2024. Том 13. № 7А. С. 41-48.

Ключевые слова

Искусственный интеллект, большие языковые модели, Death Tech, цифровой двойник, человек-прообраз, галлюцинация ИИ, этика искусственного интеллекта, трансгуманизм.

Введение

Бурное развитие технологий искусственного интеллекта (ИИ) за последние 5-7 лет поставило перед философами, правоведами и разработчиками технологий множество новых вопросов, многие из которых сегодня считаются фундаментальными и не могли быть спрогнозированы и осмыслены ранее, до повсеместного распространения искусственного интеллекта. В основе таких вопросов находится этика искусственного интеллекта, границы которой не определены и постоянно расширяются, за счет все более глобального распространения инструментов ИИ и популяризации технологии в мире. Философские принципы этики искусственного интеллекта основываются, в первую очередь, на трех ключевых понятиях – безопасности, справедливости и человекоориентированности как при разработке, так и при применении и внедрении ИИ-решений. Хотя данная совокупность принципов является, по мнению ряда философов – Ника Бострома, Рея Курцвейла, Губерта Дрейфуса и других, достаточной для осмысления технологии и ее работы на данном этапе развития, она не способна стать универсальной этической доминантой в философии искусственного интеллекта. Главная причина этого – разнообразие технологий искусственного интеллекта. Технология машинного зрения, технология машинного обучения, технология обработки естественного языка, автоматизация роботизированных процессов, экспертные системы, а главное – большие языковые модели (БЯМ) требуют индивидуалистичного подхода при формировании границ этики ИИ. Развитие БЯМ позволило создавать цифровые копии людей – особых чат-ботов, обученных на информации о жизни и деятельности реальных людей-прообразов. В первую очередь были созданы копии известных исторических личностей, однако сегодня цифровое копирование становится коммерческим, выполняющимся в интересах отдельных людей без реально научной или образовательной необходимости. Нет сомнения, что в дальнейшем такие процессы не только продолжатся, но и станут массовым явлением. Это требует работы не только с технологиями, но и с этикой и философией их разработки и применения [Кодекс этики в сфере искусственного интеллекта, www].

Большие языковые модели, ставшие основой генеративного искусственного интеллекта, сегодня являются наиболее распространенной в мире ИИ-технологией за счет создания тысяч генеративных ИИ-моделей разного размера и специфики, а также отдельных чат-ботов и нейросетей. Сегодня БЯМ имеют широкий вектор применений, от использования их в прогнозировании и аналитике до генерации фото, видео и текста. Основа для работы подобных моделей – база знаний, «скормленная» алгоритму для формирования «точки зрения» и фундамента знаний модели. Потенциально основой могут выступать любые знания, формирующие впоследствии специализацию модели. Это свойство позволяет использовать знания и особенности мышления, точку зрения и накопленный опыт конкретного человека, переведенные в машинопечатный текст, для создания модели-двойника человека. Создание при помощи БЯМ чат-ботов и цифровых копий, основанных на личностях людей, в том числе погибших, провоцирует целый спектр этических проблем, среди которых проблемы приватности, идентичности, уважения к памяти умерших, их авторские права, проблемы регулирования и контроля, проблемы создания множественных копий и т.д.

Проблема этики идентичности разработки и этики разработчика

Одним из ключевых вопросов, поднимающихся в философских дискуссиях, выступает проблема идентичности и аутентичности. Ряд западных философов, специализирующиеся на цифровой этике, предупреждают об опасности создания цифровых двойников. Обученные на истории жизни, взглядах и опыте человека боты способны не только заменить симулякром реальный опыт общения, но и представить новые данные, сведения и информацию, которые человек-прообраз мог не разделять, не желать разглашать или не иметь мнения на их счет. Еще одной областью риска становится недостаточная осведомленность людей, включая философов и даже разработчиков ИИ-технологий, о принципах работы нейронных сетей. Специфика генеративных моделей оставляет часть решений в «ведении самого ИИ», что позволяет создавать уникальный контент на основе имеющихся данных, но в то же время оставляет темные пятна в принципах и основах, на которых строятся выводы и «умозаключения» генеративных моделей [Белая книга искусственного интеллекта: европейский подход к совершенству и доверию, www].

Действительно, работа современных БЯМ может быть не до конца осознана даже самими ее создателями. Самопроизвольное, неконтролируемое, необъяснимое и нелогичное смешение первоначальных данных нейросетями или ботами называется «галлюцинацией» и на данный момент является одной из ключевых проблем, сдерживающих переход генеративного ИИ на следующий этап развития. Об этой проблеме ранее говорил создатель одной из крупнейших компаний, разрабатывающих генеративный ИИ «OpenAI», Сэм Альтман. Он заявлял, что разработчики не на 100% понимают, на каких нейронных принципах работает их самая мощная модель ChatGPT4-omni и не могут отследить причины галлюцинации моделей. Занимая отстраненную позицию, но продолжая развивать собственный продукт, разработчики создают угрозу этике ИИ, в первую очередь, связанную с использованием данных реальных людей – живых или мертвых. Это осложняет создание четко структурированной системы этики генеративного ИИ и в особенности чат-ботов, созданных на основе человеческой личности.

В этом разрезе актуальным представляется мнение итальянского философа Лючано Флориди, одного из современных философов в области цифровой этики. В своих трудах он не перекладывает ответственность за информацию, представляемую цифровыми копиями людей на сами копии, а ставит во главу угла человека, создателя или разработчика, задающего рамки этичности новой генеративной модели. Развивая идею Флориди, можно отметить, что риски, связанные с созданием цифровых двойников умерших людей, также должны лежать не на конечном пользователе и не на самой модели, а на разработчике, использующем определенный набор данных для формирования мировоззренческого фундамента нейросети и применяющем конкретные БЯМ для создания на их основе бота-двойника.

Однако сама по себе природа бота-двойника умершего человека создает ряд парадоксов и этических проблем. Одной из центральных проблем выступает нарушение этических норм по отношению к памяти умершего. Недобровольная, а при отсутствии мнения человека-прообраза она не может быть иной, оцифровка личности может быть законодательно допустима, но должно ли, при наличии подобной свободы, быть создано определенное мерило или система оценки, определяющая степень сходства прообраза и его цифровой копии? Чаще всего чат-боты, созданные как копии погибших известных личностей, не являются их 100% отражением, а представляют собой симбиоз нейросетевого алгоритма с базой знаний, на которой алгоритм был обучен [Флориди и др., 2023, www]. На текущем этапе развития технологий база знаний,

которую можно создать, используя труды автора и данные о нем, не может быть не только полной, но и не может претендовать даже на относительную объективность ввиду невозможности оцифровки субъективного опыта человека-прообраза. Невозможной представляется и попытка использовать материалы, созданные личностью, так как большая часть умозаключений, нравственных и идеологических основ человека, вне зависимости от сферы его деятельности, даже в XXI веке остается незасвидетельствованной, скрытой в подсознании специально или неосознанно. В этом случае возможность воспроизвести картину мира человека-прообраза в его цифровой копии отсутствует из-за релятивизма и ограниченного спектра данных о первоисточнике, а также ввиду отсутствия технических способов достоверного и полного цифрового копирования личности.

Сегодня создание цифровых копий становится массовым и доступным явлением. Уже созданы и функционируют чат-боты, основанные на личностях таких известных и ныне покойных личностей, как Кант, Пушкин, Менделеев, Сартр, Кафка и многих других. Контроль над созданием, формированием мировоззренческих основ при составлении базы данных, распространением и монетизацией сейчас всецело принадлежит одному человеку – разработчику. В отдельных случаях право собственности может переходить заказчику, государству или социальному институту. Случаев получения контроля над цифровым двойником, например родственниками умершего человека-прообраза, в сегодняшней судебной практике или редки, или исключены полностью.

Права цифровых копий и права на цифровые копии

Философские последствия создания цифровой копии умершего человека также обширны, как и этические. Создание бота, противоречащего идеалам своего первоисточника, ставит под сомнение уважение к личности и ее свободе выбора. Автономность личности и свобода выбора на протяжении веков остаются фундаментальным философскими ценностями, признанными классиками, такими как Иммануил Кант, Рудольф Штайнер, Альбер Камю и другими. Нарушение этой воли человека быть автономным, особенно после его смерти, можно рассматривать как форму посмертного насилия над личностью, искажающую и извращающую смысл ее образа в социокультурном пространстве. Будучи несовершенной копией, чат-бот становится не только и не столько цифровым артефактом, сколько актом этического нарушения, разрушающим идеалы и принципы прообраза, а также подменяющим представление о человеке, цифровой версией которого он является. Труды Джона Стюарта Милля о свободе личности по объективным причинам не могли поднимать вопросы цифрового копирования усопших, но выводы и положения, описанные в его работах, подводят к мысли о недопустимости принудительного копирования кого бы то ни было. Среди существующих юридических норм, косвенно касающихся вопросов распоряжения имуществом, в данном случае интеллектуальным, наиболее близкими примерами будут нарушения завещания умершего или использование тела без наличия прижизненного согласия личности на конкретную процедуру. Это поднимает вопросы о юридической защите прижизненных убеждений и возможности установить посмертные запреты на использование личности в цифровом формате. Вместе с развитием ИИ-технологий подобные меры должны стать социальной нормой, в противном случае преуменьшение их значимости приведет к бесконтрольному использованию данных погибших людей для создания на их основе цифровых двойников.

Система технологических методов «воскрешения» людей и животных сегодня является

настолько развитой, что была вычленена из подмножества ИИ-технологий в отдельную технологическую сферу – Death Tech – совокупность технологических решений, связанных со всеми аспектами человеческой смерти. Сама по себе концепция технологического воскрешения не является инновационной и появилась намного раньше, чем началось бурное развитие ИИ, включая и множество других технологических, биологических, физических и иных методов воскрешения умерших. Однако развитие ИИ-технологий придало области Death Tech новый импульс, предоставив инструмент, свойства которого позволяют добиться реальных результатов не только в воспроизведении отдельных свойств личности умершего, но в создании на основе прообраза новой информации. Развитие данной технологической сферы частично реализует идеи трансгуманизма, предлагая альтернативу физическому воскрешению – воскрешение цифровое. В этом контексте возникает философский вопрос: имеет ли право созданный бот представлять человека, который больше не может контролировать свою репрезентацию? Может ли он сохранить юридический статус прообраза и наследовать его интеллектуальную собственность? Ответы на эти вопросы в данный момент составляют семантическое ядро философии искусственного интеллекта, заходя в том числе на правовое и этическое поля, игнорирование которых недопустимо при выработке системы представлений о создании цифровых копий [Годовой отчет по индексу искусственного интеллекта: измерение тенденций области искусственного интеллекта, www].

Концепция коммерциализации памяти умерших также приобретает новые черты вместе с развитием технологий ИИ. Эксперты в области ИИ-технологий Кейт Кроуфорд и Шошана Зубофф уделяют особое внимание рискам, связанным с монетизацией цифровых двойников. Несанкционированные создание, использование и коммерциализация личности известного человека после его смерти, по сути, формируют новую форму цифровой эксплуатации личности, меняют представление о смерти в обществе и превращают нематериальные качества в товар. Развитие товарно-денежных отношений, основанных на создании и использовании цифрового аватара умерших людей, требует законодательных ограничений, как некогда торговля органами или рабская эксплуатация людей. Создание цифровых двойников с коммерческой целью без получения разрешения правообладателя также открывает новую область для совершения онлайн-преступлений и махинаций, а неэтичное и бесконтрольное использование данных о личности имеет потенциал стать инструментом для манипуляции родственниками и близкими погибшего, чья личности была оцифрована подобным образом.

Потенциал и жизненный цикл цифрового двойника

Может ли умереть сама цифровая копия? Смерть копии представляет собой полное или частичное прекращение ее существования. Сегодня решение об умерщвлении цифрового двойника принимает разработчик или непосредственный владелец ИИ-модели, однако в будущем с развитием института собственности для аналогичных форм цифрового наследия право распоряжаться ими будет четко определено еще до того, как будет создана модель. Цифровой двойник также в обозримом будущем может быть подвергнут насильственным ограничениям или даже определенному аналогу казни, если будет достоверно известно, что его деятельность является вредительской или противоречит в той или иной форме человеку, с которого копия была сделана. Лишение копии прав может быть сопряжено с разрушением семантического ядра модели и ее «обезличиванием» – это одна из перспективных мер вывода цифрового двойника из строя без полного нарушения работоспособности всей модели. В

отдельных случаях деятельность копии может быть признана нелегальной и быть ограничена законодательно [Указ Президента Российской Федерации от 10.10.2019, www]. Аналогичным ограничениям сегодня подвергаются осужденные за определенные преступления, которым запрещено публиковать что-либо в сети Интернет от своего лица или они связаны необходимостью использовать дисклеймер-предупреждение перед каждым своим сообщением.

Цифровая копия умершего человека может сама послужить прообразом для копирования. Став первоисточником, копия создаст прецедент, при котором вторичные цифровые модели, использующие ее как первоисточник, по сути, станут цифровыми копиями цифровой копии. В этом случае юридическая сила на владение копией копии не будет равносильной, так как сама цифровая модель уже успеет сгенерировать информацию и данные, которые ранее не существовали и не создавались умершим человеком-первоисточником. Вопрос копирования цифровых копий также создает ряд этических проблем и юридических коллизий, при которых многократное копирование цифрового двойника покойного человека становится универсальным способом обхода ограничений, связанных с авторским правом, правом собственности и моральным правом. Маркировка и создание универсальных методов анализа сходств базы данных копии с данными умершего человека-прообраза могут выступить инструментами, сдерживающими бесконтрольное копирование цифровых двойников в каких бы то ни было целях.

Кроме рисков, технология имеет огромный потенциал использования в медицине, бизнесе, образовании, управлении, технологиях удаленного доступа и ИИ-технологиях поддержки принятия решений. Самой перспективной сферой для использования цифровых двойников умерших людей кажется медицина, в частности психиатрия, где деликатно воссозданные копии умерших могут служить уникальным и эффективным инструментом для проработки ментальных травм, эмоциональной поддержки и работы с утратой. Не вызывает сомнений, что будут созданы универсальные типы моделей, использование которых не будет наносить ущерб родственникам и близким погибшего. Напротив, подобные модели могут стать одним из важных достижений современной медицины благодаря беспрецедентным возможностям для работы с людьми, переживающими потерю близкого. Не менее персептивным кажется использование цифровых двойников в роли наставников и учителей. Так, лауреаты Нобелевских премий, признанные ученые и философы смогут заниматься образованием тысяч людей одновременно, без привязки ко времени, пространству и даже собственной жизни. В отличие от медицинского применения, цифровые двойники в сфере образования могут быть ограничены строго «профессиональными» аспектами личности умершего человека-прообраза, что упрощает их создание и снимает ряд этических дилемм и юридических ограничений. Боты-преподаватели не должны заменять реальных людей, но вполне могут стать помощниками и проводниками для обучения истории, литературе, живописи, физике, химии и многим другим наукам.

Заключение

Чат-боты, созданные на основе личности умершего человека, на данный момент являются полностью неэтичными, не регулируются законодательно, не регламентируются и являются до конца «цифровыми копиями» в полном смысле. Однако бурное развитие технологий ИИ в краткосрочной перспективе приведет к тому, что создание цифровых копий станет массовым и доступным явлением, рамки этики которого только предстоит определить. Широкая популярность цифровых двойников неизбежна ввиду функционала, которым они обладают.

Возможность реинкарнации, продолжения жизни после смерти, о которой человечество мечтало с самого своего зарождения, будет выступать неконтролируемым искушением. Современные представления о развитии подобной цифровой формы перерождения не соответствуют тем рискам и дилеммам, которые она уже порождает и будет порождать еще больше вместе с развитием технологии. Обеспечение своевременного философского осмысления, юридического контроля и этического регулирования являются необходимыми условиями для равномерного, осознанного и безопасного развития этого вида технологий искусственного интеллекта. Текущая позиция общества является преимущественно реактивной и не позволяет обеспечить достаточного уровня этичности и контроля над рисками для продолжения развития технологии с теми же темпами, что ведутся сегодня.

Библиография

1. Альянс в сфере искусственного интеллекта. Кодекс этики в сфере искусственного интеллекта. URL: https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/Кодекс_этики_20_10_1.pdf (дата обращения: 08.09.2024).
2. Белая книга искусственного интеллекта: европейский подход к совершенству и доверию. URL: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en (дата обращения: 10.09.2024)
3. Флориди Л. и др. Этические рамки для хорошего искусственного интеллекта: возможности, риски, принципы, рекомендации. 2023. URL: https://link.springer.com/chapter/10.1007/978-3-030-81907-1_3 (дата обращения: 10.09.2024).
4. Стэнфордский университет. Годовой отчет по индексу искусственного интеллекта: измерение тенденций области искусственного интеллекта. URL: https://aiindex.stanford.edu/report/ (дата обращения: 15.09.2024).
5. О развитии искусственного интеллекта в Российской Федерации: указ Президента Российской Федерации от 10.10.2019 // Официальный сайт Президента Российской Федерации. URL: http://www.kremlin.ru/acts/bank/44731 (дата обращения: 18.09.2024).

The philosophy of artificial intelligence: the ethical framework for creating and exploiting of digital human counterparts

Vyacheslav V. Kidyaev

Postgraduate Student,
Plekhanov Russian University of Economics,
Chief Expert of the Department of Communications and International Affairs
of the National Center for Artificial Intelligence Development
under the Government of the Russian Federation,
109028, 11 Pokrovsky Boulevard, Moscow, Russian Federation;
e-mail: Slavakid@yandex.ru

Abstract

The development of artificial intelligence technologies, in particular generative artificial intelligence, has opened up many unique opportunities for humanity. One of these possibilities was resurrection, which is available today by creating digital copies of people based on algorithms. This development has already attracted a lot of attention from developers and experts around the world, today digital copies of such famous personalities of the past and present as Immanuel Kant, Elon Musk, Alexander Pushkin, Karl Marx, Friedrich Nietzsche and many others have been created.

However, like all other potential forms of "resurrection", the creation of digital copies has provoked many risks and dilemmas associated with the development, application and rights to use them. Today, most of the key ethical decisions that determine the moral contour of the future AI model are made by the developer, a technical specialist focused primarily on the effectiveness and reliability of development. This only further widens the gap between the ethics and philosophy of AI and the actual implementation of such developments in the economic sphere. The lack of a sufficient number of tools to regulate the use of AI technologies to create digital copies creates a lot of fears and distrust of people around these developments, which requires a philosophical analysis of this issue and its ethical understanding.

For citation

Kidyaev V.V. (2024) Filosofiya iskusstvennogo intellekta: ramki etichnosti sozdaniya i ekspluatatsii tsifrovykh dvoinikov lyudei [The philosophy of artificial intelligence: the ethical framework for creating and exploiting of digital human counterparts]. *Kontekst i refleksiya: filosofiya o mire i cheloveke* [Context and Reflection: Philosophy of the World and Human Being], 13 (7A), pp. 41-48.

Keywords

Artificial intelligence, large language models, Death Tech, digital double, human prototype, hallucination And, ethics of artificial intelligence, transhumanism.

References

1. Alliance on Artificial Intelligence. Code of Ethics for Artificial Intelligence. Available at: https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/Кодекс_етики_20_10_1.pdf (Accessed: 09/08/2024).
2. Floridi L. et al. An Ethical Framework for Good Artificial Intelligence: Opportunities, Risks, Principles, Recommendations. 2023. URL: https://link.springer.com/chapter/10.1007/978-3-030-81907-1_3 (date of access: 10.09.2024).
3. On the development of artificial intelligence in the Russian Federation: decree of the President of the Russian Federation of 10.10.2019 // Official website of the President of the Russian Federation. URL: http://www.kremlin.ru/acts/bank/44731 (date of access: 18.09.2024).
4. Stanford University. Artificial Intelligence Index Annual Report: Measuring Trends in Artificial Intelligence. URL: https://aiindex.stanford.edu/report/ (date of access: 15.09.2024).
5. White Paper on Artificial Intelligence: A European Approach to Excellence and Trust. Available at: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en (Accessed: 09/10/2024)