Человеческое в человеко-машинном гибриде искусственного интеллекта

Дарья Чирва
Национальный исследовательский университет ИТМО;
Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), Санкт-Петербург, Россия, dvchirva@itmo.ru.

Ключевые слова: антропоцентризм; инструментализм; агентность; моральная агентность; социотехническая система; искусственный интеллект; проблема ответственности.

Статья посвящена вопросу о том, что происходит со статусом человека в современный период распространения технологий искусственного интеллекта (далее — ИИ). Демонстрируется недостаточность ресурсов антропоцентризма для определения реального положения человека на том основании, что в качестве следствия он ведет к фиксации простой инструментальности за системами ИИ. Такое распределение ролей между человеком и технологией не соответствуют реальной сложности функционирования и взаимодействия человека и ИИ на современном этапе. ИИ в силу своего устройства обладает определенной степенью автономии, непрозрачности и непредсказуемости, что позволяет в целом говорить о нем как об агенте. Однако если допустить расширение поля морали в результате ее переориентации с субъекта действия на объект и его морально значимые изменения, то искусственные агенты могут расцениваться как моральные агенты (проект информационной этики Лучано Флориди).

Однако эти агенты проектируются и обучаются людьми на основе данных, которые производятся, собираются, размечаются людьми. Реальность производства технологии в ее социальном аспекте настолько сложна, что в итоге не представляется возможным учесть отдельный вклад всех агентов (человеческих и нечеловеческих) в ее функционирование. Поэтому для фиксации этой ситуации используется понятие социотехнической мультиагентной системы. Она характеризуется различными уровнями агентности: каузальной, моральной и моральной с возможностью несения ответственности. Человеческое в социотехнической системе является источником моральной агентности и ответственности в силу того, что они являются его сущностными свойствами. Дальнейшее направление исследования социотехнической системы определяется необходимостью проработки проблемы пробела в ответственности.

Введение

ОДНИМ из ключевых элементов понятийной системы координат в любом поле философского исследования является понятие человеческого. Его связь с другими центральными понятиями определяет его место и содержание. В зависимости от того, что стоит рядом с человеком: Бог или, к примеру, животное, — мы по-разному видим и самого человека. Сегодня статистически «человек» наиболее часто встречается в контексте разговора о технологиях в целом, об искусственном интеллекте (далее — ИИ) в частности. Что дает нам это соседство? Перечень новых понятийных затруднений, которые возникают при решении в том числе практических проблем. Новые системы ИИ не только успешно обрабатывают большие объемы данных, но и способны производить на их основе новые данные в виде текста, изображения, музыки, видео. Тем самым возникает вопрос об авторстве, о сути подлинного творчества.

Это значимые вопросы, но мы касаемся их здесь только с тем, чтобы обозначить, что стремительное развитие ИИ сначала производит понятие «естественный интеллект», а затем начинает конкурировать с ним. Конечно, *GPT* (*Generative Pretrained Transformer* — генеративный предобученный трансформер) не производит смыслы, а стремится построить статистически пригодную последовательность токенов. Тем не менее данная имитация производства смысла оказывает существенное влияние на человека. Сейчас *GPT* посредством волевого решения человека закрывается возможность давать ответы на критически значимые для человека вопросы, связанные с его здоровьем, например. Дискурс об ИИ все чаще принимает форму рассуждений об экзистенциальной угрозе для человечества. Здесь уместно вспо-

В работе представлены результаты проекта «Границы современной культуры: природа, технологии и социальные интерфейсы», выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2023 году.

мнить, о чем предупреждал нас еще Мартин Хайдеггер[1]. Если человек мыслит технологию как инструмент наращивания власти, то ее развитие в какой-то момент неизбежно начнет расцениваться им как угроза его существованию оттого, что орудие наращивает мощь и автономию. Сила и эффективность инструмента становится предметом не только гордости создателя, но и тревоги масс: вдруг орудие обернет свою мощь против создателя? Данная тенденция риск-ориентированного мышления находит свое выражение в различных нормативных документах по прикладной этике: этических рекомендациях о развитии ИИ[2], кодексах по этике в сфере ИИ[3]. Публичный дискурс, представленный в таких документах, нерефлексивным образом реконституирует антропоцентризм. Человек позиционируется как единственный носитель подлинного смысла и контроля над технологией, что должно выступать основой для решения всех затруднений и создавать ощущение контролируемости всего происходящего.

Одной из новых опор антропоцентризма в цифровую эпоху является наличие проблемы ответственности в сфере ИИ, не имеющей простого решения. В текущей ситуации декларируется принцип несения человеком полной ответственности за действия систем ИИ. Тем не менее декларирование не может заменить аргументацию. Аналитического решения данная проблема еще не получила: посредством стратегического усиления антропоцентризма значение ИИ редуцируется до простого инструмента. Тем самым подлинное место человека в эпоху ИИ оказывается непроясненным в новых реалиях партнерства с ИИ.

Агентность искусственного интеллекта

Многие авторы еще в период, предшествующий распространению технологии генеративного ИИ, говорили о том, что данная технология в силу особенностей своего устройства не является простым инструментом[4]. (Стоит отдельно отметить, что из того, что

1. *Хайдеггер М.* Вопрос о технике // Время и бытие: ст. и выступ. М.: Республика, 1993.
2. Рекомендация об этических аспектах искусственного интеллекта ЮНЕСКО 2021 года // UNESDOC. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus.
3. Кодекс этики в сфере искусственного интеллекта. М.: Альянс в сфере искусственного интеллекта, 2023. URL: https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/Кодекс_этики_20_10_1.pdf.
4. *Rahwan I. et al.* Machine Behaviour // Nature. 2019. Vol. 568. № 7753.

ИИ не является простым инструментом, еще не следует, что он является полноценной искусственной личностью, обладающей интенциональностью, волей, истинным целеполаганием и т. п.[5] У нас не возникает каких-либо оснований для развертывания дискурса о том, что у самого ИИ есть некий план в отношении человека.) Принципиальная черта инструментальности — отсутствие каких-либо форм агентности. В классической антропоцентрической парадигме инструментальным оказывается все, за исключением человека. Если человек распоряжается технологией, осуществляя полный контроль за ее функционированием, то речь идет об инструменте, и его суть таким образом однозначно предопределена. Следовательно, «человеческое» антропоцентризма всякий раз оказывается под ударом там и тогда, когда на сцене появляются иные значимые агенты (животные или технические сущности), которые не поддаются полному контролю человека. Однако в некотором смысле системы ИИ уже вышли из-под контроля человека: для функционирования ИИ не требуется постоянный надзор и вмешательство человека на всех этапах жизненного цикла технологии. Об агентности ИИ свидетельствуют следующие его черты:

1. Наличие автономии как способности без непосредственного постоянного контроля человека производить новые данные, функционировать в неопределенных условиях.
2. Непрозрачность ИИ как отсутствие возможности для человека в режиме реального времени получить исчерпывающие объяснения процессов между так называемыми входом и выходом. Данный аспект иначе называется проблемой черного ящика.
3. Непредсказуемость ИИ как его способность производить операции, которые не заложены в него разработчиками и не ожидаются пользователями.

Наделение ИИ статусом минимальной агентности выводит нас за рамки инструментализма в отношении технологии. Однако имеется и тенденция по наделению ИИ статусом морального агента, заложенная в проекте информационной этики Лучано Флориди[6]. Он имеет своей целью поколебать незыблемость основ антропоцентризма в сфере этики. Так, в статье 2004 года Флори-

5. Однако дискурс о нравственности самого ИИ развивается в текущий момент. См., напр.: *Железнов А.* Мораль для искусственного интеллекта: перспективы философского переосмысления // Логос. 2021. Т. 31. № 6.
6. *Floridi L.* Information Ethics: On the Theoretical Foundations of Computer Ethics // Ethics and Information Technology. 1999. Vol. 1. № 1.

ди и его соавтор, Джефф Сандерс, доказывают, что искусственные агенты могут быть также моральными агентами. Флориди и Сандерс вводят понятие морального агента не через операцию определения, а посредством построения его характеристики на определенном уровне абстракции[7]. Они исходят из того, что вводить сущности как таковые (сопоставляя этот уровень рассуждения с рассуждением о кантианской вещи в себе), то есть вне конкретного контекста, невозможно. Исследование ведется на определенном уровне абстракции, который не артикулируется четко либо в случае отсутствия рефлексии, либо в силу распространения этого уровня абстракции на уровень общих убеждений.

Характеристики ИИ-агента, позволяющие информативно описать его в контексте модели агента, следующие:

- интерактивность (способность отвечать на изменения окружающей среды, изменяя свои состояния);
- автономия (способность изменять состояния в соответствии со своими внутренними принципами и без дополнительного влияния внешней среды);
- адаптивность (изменение внутренних правил перехода от одного состояния к другому на основе обратной связи от окружающей среды).

Авторы также вводят дополнительный критерий моральности для искусственных агентов:

> Действие квалифицируется как моральное, если и только если оно приводит к морально значимому добру или злу. Агент является моральным агентом, если и только если он способен производить морально квалифицируемое действие[8].

Моральный агент является субъектом подотчетности (*accountability*), если он послужил источником для морально значимого действия, независимо от того, человек он или нет. Данную идею безвиновной (*faultless*), или объектно-ориентированной, морали (*object-oriented-morality*) Флориди развивал и в своих последующих работах[9]. В его концепции фокус внимания смещается с самого

7. *Floridi L., Sanders J. W.* On the Morality of Artificial Agents // Minds and Machines. 2004. Vol. 14. № 3. P. 351.
8. Ibid. P. 364.
9. *Floridi L.* Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions // Philosophical Transactions. 2016. Vol. 374. № 2083; *Idem.* Distributed Morality in an Information Society // Science and Engineering Ethics. 2012. Vol. 19.

агента и наличия у него определенных морально значимых намерений (по такой модели строится классический антропоцентристский дискурс морали) на конечное состояние объекта воздействия. В рамках объектно-ориентированного подхода существенным является не то, кто совершает то или иное действие, а то, приводит ли действие агента к морально значимым изменениям состояния объекта воздействия. При этом на определенном уровне рассмотрения (или абстракции) и люди, и организации, и искусственные агенты могут рассматриваться как моральные агенты, потому что они способны влиять на моральный статус других агентов[10].

Однако данное расширение морального пространства, на наш взгляд, происходит не за счет увеличения числа искусственных агентов, а за счет людей-агентов, способных формировать новые социальные, партнерские отношения с не-людьми. Сам Флориди отмечает, что в рамках объектно-ориентированной морали, которая включает в число моральных агентов искусственные сущности, нет оснований для приписывания ответственности искусственным агентам. Ее отсутствие оставляет их на уровне действенных причин определенных событий и не возводит на уровень подлинных моральных агентов классической этики. Искусственные агенты, в силу отсутствия у них подлинной интенциональности сознания, воли, не заслуживают похвалы или порицания в силу их неприменимости. Ответственность способны нести только люди. Однако каким образом она функционирует в новой среде, которую мы разделяем вместе с искусственными агентами?

Социотехническая система

Люди активно взаимодействуют с ИИ, получая в результате новые возможности. Алгоритмы помогают опытным пользователям решать рутинные рабочие, организационные и коммуникативные задачи. Сервисы *GPT* предоставляют возможность получить индивидуального помощника буквально каждому человеку. Такие взаимодействия с ИИ существенным образом трансформируют сложившиеся связи между людьми. Это позволяет нам говорить о том, что в контексте вопроса о человеческом в эпоху ИИ стоит говорить о гибридной сущности человека и машины, действу-

10. *List C.* Group Agency and Artificial Intelligence // Philosophy and Technology. 2021. Vol. 34. № 4.

ющих совместно[11]. Однако поскольку сам машинный интеллект представляет собой не только результат вычислительной деятельности, не исключительно техническую сущность, но и результат взаимодействия некоторого числа людей в процессе его разработки, то корректно говорить не о том или ином гибриде конкретного человека (как пользователя) с конкретной системой ИИ, а о социотехнической системе, в которую входит большое число людей и технология ИИ.

Искусственный интеллект сегодня представляет собой переплетение и взаимосвязь ценностных установок людей с технологическими решениями. Фактически корректно говорить о человеко-машинном гибриде ИИ. Нейросетевая технология в силу своего технического устройства является социотехнической системой. Нейросети проектируются и обучаются людьми на основе данных, которые производятся, собираются, размечаются людьми. Социальный и институциональный контекст действий, совершаемых людьми, производящими и использующими технологию, настолько сложен, что не представляется возможным учесть отдельный вклад всех агентов (человеческих и нечеловеческих) в итоговое решение. На уровне разработки, обучения и тестирования ИИ во многом определяется человеческим воздействием. У ИИ, когда речь идет о социотехнической системе, нет собственной воли, мотива и целей, но в него вкладывается достаточно много человеческого: различного рода предубеждения разработчиков, их нерефлексируемые установки в отношении всего того, что мы называем значимым, правильным и т. п., которые задействуются на уровне проектирования, обучения и тестирования моделей. Решение задачи генерируется ИИ, и, помимо так называемого черного ящика, на решении не всегда прогнозируемым и контролируемым образом скажутся представления тех людей, что создают и используют технологию.

В этой связи возможно несколько подходов к описанию отношений между человеческими и нечеловеческими агентами. Классический подход подразумевает контроль человека и полную его ответственность за все результаты функционирования ИИ. Это так называемая модель «человек в цикле» (*human-in-the-loop*). Однако она не соответствует текущему уровню развития технологии в силу того, что технология не является простым инструментом, действия которого целиком и полностью контролируются че-

11. *Латур Б.* Пересборка социального: введение в акторно-сетевую теорию. М.: ИД ВШЭ, 2014.

ловеком, как это предполагается данной парадигмой. Технология обладает высоким «творческим»[12] потенциалом (способностью решать задачи, превосходящие когнитивные способности человека), возможностью автономного функционирования и использования в рамках социального взаимодействия человека и машины, в ходе которого деятельность человека и функционирование искусственного агента корректируются и видоизменяются.

Альтернативой модели контроля является модель «человек вне цикла» (*human-outside-the-loop*), допускающая бесконтрольное функционирование автономной технологии, в ходе которого с неизбежностью возникают пробелы в ответственности[13]. Они имеют место в таких ситуациях, в которых человек не может быть признан ответственным лицом в силу того, что не обладает технической возможностью осуществлять контроль за функционированием высокоразвитой технологии. Вменить ответственность технологии не представляется возможным в силу того, что она не является живой и не способна стремиться к удовольствию и избегать страдания. Такие ситуации, безусловно, связаны с высокими рисками и не должны возникать[14]. Однако не только на уровне оценки, но и на понятийном уровне данная модель несовершенна, поскольку в целом модель «человек вне цикла» исходит из базовой предпосылки инструментального понимания технологии и основывается на предположении дуализма человека и технологии (так же как и модель «человек в цикле»).

12. «Творческое» заключено в кавычки в силу того, что в вопросе о том, корректно ли называть результат работы генеративного ИИ продуктом творчества, нет однозначного решения. Или же, следуя Флориди, можно сказать, что на разных уровнях абстракции будут даны разные ответы.
13. Пробелы в ответственности (*responsibility gap*) — ситуации с функционированием ИИ, в которых невозможно определить ответственное лицо либо в силу большого числа агентов, повлиявших на результат, одним из которых является ИИ, либо в силу того, что мы не можем приписать ответственность ИИ там и тогда, когда знаем, что его функционирование послужило причиной события. Причем если бы на месте ИИ во втором случае был человек, то мы бы однозначно определили его в качестве ответственного. Впервые понятие «пробел в ответственности» использовал Андреас Маттиас. См.: *Matthias A.* The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata // Ethics and Information Technology. 2004. Vol. 6. № 3.
14. Именно на этом основании ряд исследователей приходит к выводу, что автономные системы вооружения не должны использоваться, ибо это нарушает принцип справедливой войны. См., напр.: *Sparrow R.* Killer Robots // Journal of Applied Philosophy. 2007. Vol. 24. № 1.

Технология ИИ предполагает не просто автоматизацию некоторых рутинных процессов, но их существенную трансформацию в логике взаимодействия человека и ИИ, позволяющую говорить о появлении нового типа социотехнической системы. Анализ отношений ответственности в рамках такой системы возможен в терминах модели «человек над циклом» (*human-on-the-loop*)[15]. В данной модели человек и ИИ формируют систему, в которой определенной степенью автономии обладают и искусственный агент, и человеческий агент. Они могут взаимодействовать друг с другом, корректируя действия друг друга. Оба влияют на итоговый результат взаимодействия. Исключение одного из участников взаимодействия существенным образом влияет на итоговый результат. В условиях широкого использования систем ИИ в социально значимых контекстах источником изменения состояния системы является вся сложная сеть взаимодействующих агентов: это и разработчики, и сама система ИИ, и ее пользователи. Эта сеть совершает распределенные действия. Определение индивидуального вклада агента в случае распределенной социотехнической системы затруднено. Каждый из составляющих ее агентов действует так, а не иначе в силу того, что он находится в рамках данной системы и возможности его действий зависят от всей совокупности действий связанных с ним других агентов. Именно поэтому мы говорим о человеко-машинном гибриде ИИ, социотехнической системе, анализируемой по модели «человек над циклом».

Человеческое

Человеческое остается на своем месте в составе сложной сети отношений среди новых искусственных агентов, успешно имитирующих или воссоздающих когнитивный и коммуникативный форматы человеческого существования. Однако это человек не возвышающийся-над, а включенный-в сложную сеть. Остается ли что-то специфически человеческое, что позволяет нам не декларативным, а аналитическим образом выявлять человеческий компонент социотехнической системы? На наш взгляд остается, и это то, что наделяет мультиагентную социотехническую систему способностью нести ответственность, в дополнение к развитию моральной агентности.

15. *Strasser A.* Distributed Responsibility in Human-Machine Interactions // AI and Ethics. 2022. Vol. 2. № 3.

Существует несоответствие между имеющейся практикой определения ответственности и реальностью технологии, с которой мы имеем дело сегодня. Это связано со сложностью социального контекста действий, совершаемых с использованием ИИ, — в него включены те, на ком может сказаться результат действия: пользователи технологии, разработчики, менеджмент, принявший решение о разработке и ее выводе на рынок для определенной аудитории, авторы законодательного регулирования сферы использования технологии. Практически невозможно выявить одного деятеля, который бы осознанно контролировал все происходящее и мог бы нести ответственность. Более того, в такой ситуации представляется невозможным и несправедливым приписывать решение, действие и ответственность за него одному человеку. Однако именно так устроен классический дискурс об ответственности. В реальности же имеет место сложный комплекс взаимосвязей, который образуют люди, создающие и использующие технологию, с технологией. При этом сам контекст действия (особенно социального действия) в текущий момент в значительной мере определяется технологиями.

В первую очередь стоит принять тот факт, что в определенной мере моральная агентность свойственна и ИИ. Она приписывается человеческим и нечеловеческим акторам, когда мы допускаем, что они выступают в качестве причины некоторой новой возникающей последовательности событий, имеющей моральное значение. Такое значение возникает всякий раз, когда функционирование социотехнической системы связано с влиянием на систему ценностей той или иной социальной группы. Так происходит в силу того, что ее элементами оказываются люди, носители ценностной перспективы. К тому же системы ИИ как подсистемы социотехнической системы влияют на результат когнитивной и социальной деятельности человека. Действия социотехнической системы могут приводить к морально значимым последствиям, особенно в социально значимых контекстах. Однако ИИ-агентность обусловлена функционированием алгоритмов. Она не выступает достаточным основанием для того, чтобы приписать полную агентность, включающую в себя компонент ответственности. Моральная агентность систем ИИ, таким образом, является привходящим, а не сущностным свойством систем ИИ, включенных в социотехническое целое, которое они формируют вместе с людьми. Подлинная моральная агентность остается сущностным свойством человека.

Если предположить исключение человека из социотехнической системы, то вместе с ним пропадает и привходящее свойство моральной агентности систем ИИ. Ответственность рассматривается как отношение, которое возникает между моральным агентом и его объектом в результате некоего действия, источником которого выступает сам моральный агент. Приписывание такого отношения системам ИИ является невозможным в силу того, что они не обладают подлинной интенциональностью и не имеют мотивационного осознаваемого компонента в своем функционировании. Человеческое оказывается носителем уникальной агентности в рамках социотехнической системы ИИ, которое своим присутствием привносит в эту систему не только значение моральной агентности, но и возможность распределения ответственности. Решение этой проблемы предполагает дальнейшую аналитику типов ответственного отношения, продуцируемых социотехнической системой. Дальнейшие усилия по аналитике состава социотехнической системы, типов ответственного отношения, обозначаемых сейчас одним термином, должны составить одно из ключевых направлений социогуманитарной экспертизы в сфере искусственного интеллекта.

Библиография

Железнов А. Мораль для искусственного интеллекта: перспективы философского переосмысления // Логос. 2021. Т. 31. № 6. С. 95–122.

Кодекс этики в сфере искусственного интеллекта. М.: Альянс в сфере искусственного интеллекта, 2023. URL: https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/Кодекс_этики_20_10_1.pdf.

Латур Б. Пересборка социального: введение в акторно-сетевую теорию. М.: ИД ВШЭ, 2014.

Рекомендация об этических аспектах искусственного интеллекта ЮНЕСКО 2021 года // UNESDOC. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus.

Хайдеггер М. Вопрос о технике // Время и бытие: ст. и выступ. М.: Республика, 1993. С. 221–238.

Floridi L. Distributed Morality in an Information Society // Science and Engineering Ethics. 2012. Vol. 19. P. 727–743.

Floridi L. Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions // Philosophical Transactions. 2016. Vol. 374. № 2083. P. 1–13.

Floridi L. Information Ethics: On the Theoretical Foundations of Computer Ethics // Ethics and Information Technology. 1999. Vol. 1. № 1. P. 37–56.

Floridi L., Sanders J. W. On the Morality of Artificial Agents // Minds and Machines. 2004. Vol. 14. № 3. P. 349–379.

List C. Group Agency and Artificial Intelligence // Philosophy and Technology. 2021. Vol. 34. № 4. P. 1213–1242.

Matthias A. The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata // Ethics and Information Technology. 2004. Vol. 6. № 3. P. 175–183.
Rahwan I, Cebrian M., Obradovich N., Bongard J., Bonnefon J.-F., Breazeal C., Crandall J. W., Christakis N. A., Couzin I. D., Jackson M. O., Jennings N. R., Kamar E., Kloumann I. M., Larochelle H., Lazer D., McElreath R., Mislove A., Parkes D. C., Pentland A., Roberts M. E., Shariff A., Tenenbaum J. B., Wellman M. Machine Behaviour // Nature. 2019. Vol. 568. № 7753. P. 477–486.
Sparrow R. Killer Robots // Journal of Applied Philosophy. 2007. Vol. 24. № 1. P. 62–77.
Strasser A. Distributed Responsibility in Human-Machine Interactions // AI and Ethics. 2022. Vol. 2. № 3. P. 523–532.

A HUMAN ELEMENT IN A HUMAN-MACHINE HYBRID OF ARTIFICIAL INTELLIGENCE

Daria Chirva. St. Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO University); National Research University Higher School of Economics (HSE University), St. Petersburg, Russia, dvchirva@itmo.ru.

Keywords: anthropocentrism; instrumentalism; agency; moral agency; sociotechnical system; artificial intelligence; responsibility problem.

The article is devoted to the question of what happens to the status of a human being in the modern period of artificial intelligence technology (hereinafter — AI) dominance. Anthropocentrism is treated as an insufficient conceptual means for the identification of the real position of a human, because it leads to the fixation of simple instrumentality behind AI systems. Such a distribution of roles between humans and technology does not correspond to the real complexity of the functioning and interaction of humans and AI at the present stage. AI, by virtue of its construction, has a certain degree of autonomy, opacity and unpredictability, which allows us to generally talk about it as an agent. However, if we allow the expansion of the field of morality as a result of its reorientation from the subject of action to the object and its moral state's significant changes, then artificial agents can be regarded as moral agents (based on Luciano Floridi's Information ethics project).

However, these agents are designed and trained by humans based on data that is produced, collected, and marked up by humans again. The reality of technology production in its social aspect is so complex that, as a result, it is not possible to take into account the separate contribution of all agents (human and non-human) to its functioning. Therefore, the concept of a sociotechnical multi-agent system is used to designate this situation. It is characterized by different levels of agency: causal, moral and moral with the possibility of responsibility. The human in the sociotechnical system is a source of moral agency and responsibility due to the fact that they are its essential properties. The further direction of the research is connected with the question of how the problem of eliminating the responsibility gap in the realm of the sociotechnical system should be solved.

DOI: 10.17323/0869-5377-2024-6-203-214

References

Floridi L. Distributed Morality in an Information Society. *Science and Engineering Ethics*, 2012, vol. 19, pp. 727–743.

Floridi L. Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions. *Philosophical Transactions*, 2016, vol. 374, no. 2083, pp. 1–13.

Floridi L. Information Ethics: On the Theoretical Foundations of Computer Ethics. *Ethics and Information Technology*, 1999, vol. 1, no. 1, pp. 37–56.

Floridi L., Sanders J. W. On the Morality of Artificial Agents. *Minds and Machines*, 2004, vol. 14, no. 3, pp. 349–379.

Heidegger M. Vopros o tekhnike [Die Frage nach der Technik]. *Vremia i bytie: stat'i i vystupleniia* [Time and Being: Articles and Speeches], Moscow, Respublika, 1993, pp. 221–238.

Kodeks etiki v sfere iskusstvennogo intellekta [The Code of Ethics in the Field of Artificial Intelligence], Moscow, AI Alliance Russia, 2023. Available at: https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/Кодекс_этики_20_10_1.pdf.

Latour B. *Peresborka sotsial'nogo: vvedenie v aktorno-setevuiu teoriiu* [Reassembling the Social. An Introduction to Actor-Network-Theory], Moscow, HSE Publishing House, 2014.
List C. Group Agency and Artificial Intelligence. *Philosophy and Technology*, 2021, vol. 34, no. 4, pp. 1213–1242.
Matthias A. The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, 2004, vol. 6, no. 3, pp. 175–183.
Rahwan I, Cebrian M., Obradovich N., Bongard J., Bonnefon J.-F., Breazeal C., Crandall J. W., Christakis N. A., Couzin I. D., Jackson M. O., Jennings N. R., Kamar E., Kloumann I. M., Larochelle H., Lazer D., McElreath R., Mislove A., Parkes D. C., Pentland A., Roberts M. E., Shariff A., Tenenbaum J. B., Wellman M. Machine Behaviour. *Nature*, 2019, vol. 568, no. 7753, pp. 477–486.
Rekomendatsiia ob eticheskikh aspektakh iskusstvennogo intellekta IuNESKO 2021 goda [UNESCO Recommendation on the Ethics of Artificial Intelligence]. UNESDOC. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus.
Sparrow R. Killer Robots. *Journal of Applied Philosophy*, 2007, vol. 24, no. 1, pp. 62–77.
Strasser A. Distributed Responsibility in Human-Machine Interactions. *AI and Ethics*, 2022, vol. 2, no. 3, pp. 523–532.
Zheleznov A. Moral' dlia iskusstvennogo intellekta: perspektivy filosofskogo pereosmysleniia [The Moral of Artificial Intelligence: A Chance to Reconsider Philosophy]. *Logos* (Russia), 2021, vol. 31, no. 6, pp. 95–122.