Этическая мысль
2024. Т. 24. № 1. С. 129–143
УДК 17:004.8

Ethical Thought
2024, Vol. 24, No. 1, pp. 129–143
DOI: 10.21146/2074-4870-2024-24-1-129-143

А.С. Глуховский, А.Д. Дурнев, Д.В. Чирва

Распределенная моральная ответственность в сфере искусственного интеллекта

Глуховский Андрей Сергеевич – преподаватель. Университет ИТМО. Российская Федерация, 197101, г. Санкт-Петербург, Кронверкский пр., д. 49, лит. А.

ORCID 0000-0003-0052-0180
e-mail: aglukhovsky@itmo.ru

Дурнев Алексей Дмитриевич – преподаватель. Университет ИТМО. Российская Федерация, 197101, г. Санкт-Петербург, Кронверкский пр., д. 49, лит. А.

ORCID 0000-0002-1248-4685
e-mail: a.durnev.ph@gmail.com

Чирва Дарья Викторовна – преподаватель. Университет ИТМО. Российская Федерация, 197101, г. Санкт-Петербург, Кронверкский пр., д. 49, лит. А.

ORCID 0000-0001-7804-8803
e-mail: dvchirva@itmo.ru

Основной целью статьи является обоснование оптимальности концепции распределенной морали Лучано Флориди для описания модели отношения ответственности в сложных системах «человек – искусственный интеллект» (ИИ). Достаточно высокий уровень автономии в функционировании ИИ ставит под вопрос классический инструментализм в отношении этой технологии, так как она оказывается способной на принципиально неконтролируемые со стороны человека действия в процессе выполнения поставленной перед ней задачи. По этой же причине классическое определение моральной ответственности постепенно перестает соответствовать действительному положению дел в области разработки ИИ, что зачастую приводит к проблемам, тормозящим ее развитие. Поэтому в статье осуществляется концептуализация нового, соответствующего типу технологии способа описания отношений ответственности в сфере этики ИИ. Для достижения этой цели ИИ рассматривается как обладающий агентностью участник различных взаимодействий (на основе критериев

агентности Флориди), в том числе и социальных, но при этом ему не приписывается подлинная интенциональность, мотивация, осознанность и другие свойства, являющиеся необходимыми для атрибутирования агентности в классической модели отношения ответственности. Кроме обращения к описанию критериев агентности Флориди, обоснование такого рассмотрения ИИ в статье осуществляется с опорой на концепт текучести Аннамари Мол и Марианы де Лаэт, который разрабатывается ими в том числе и для обозначения способности технологий обладать агентностью без наличия у них свойств, считающихся необходимыми для атрибутирования агентности в классической модели описания роли технологий в социальных процессах. Технология ИИ представляется как сложная гетерогенная социотехническая система, конструируемая как человеческими, так и нечеловеческими агентами. Риск нивелирования значимости ответственности в такой системе преодолевается на концептуальном уровне посредством адаптации подхода объект-ориентированной морали Флориди. Смена акцента с субъекта действия на объект воздействия наделяет моральной значимостью всю ту сумму морально нейтральных действий, которые совершаются агентами в рамках социотехнической системы. Поскольку она производит морально значимые эффекты и воздействия, то выступает и носителем отношения ответственности, которое, как показано в статье, распространяется на каждого агента системы в равной мере, интенсифицируя тем самым моральную ответственность в сфере ИИ.

Ключевые слова: этика искусственного интеллекта, распределенное познание, распределенная ответственность, социотехническая система, агентность, объект-ориентированная мораль

Введение

Стремительное развитие современных технологий искусственного интеллекта (ИИ) бросает вызовы не только инженерной и научной мысли, но и философскому их осмыслению. Увеличение степени автономности технологий – проблема, подрывающая основания устоявшихся представлений о самой сущности таких явлений, как агентность и ответственность [Dignum, Dignum, 2020, 1701–1702]. Классическое понимание этих терминов базируется на тезисе, согласно которому и ответственность, и агентность могут быть атрибутированы исключительно человеку, поскольку только он может быть подлинным источником действия в силу наличия у него осознанности, целеполагания, интенциональности и мотивации [Лисанюк и др., 2014]. Из этого также следует, что только за человеком признается способность определять, при каких условиях наступает ответственность, как те или иные внешние факторы влияют на характер и степень ответственности. При таком подходе теряется специфика ИИ, поскольку он оказывается лишь простым инструментом, человеческую субъектность или подчиняющимся ей.

Инструментальное отношение к технологии предполагает утверждение за человеком способности полного контроля над функционированием технических объектов. Такой подход отрицает наличие реального взаимодействия между человеком и технологией и не допускает возможности обнаружения различных возможных форматов их совместной деятельности. Для классиче-

ского подхода к ответственности отсутствие у систем ИИ сознания и мотивационно-волевого компонента является достаточным основанием признания невозможности участия искусственных агентов в отношениях ответственности [Müller, 2021, 580]. Однако в сложноорганизованных системах взаимодействия человека и технологии, в которых в случае использования ИИ, обладающего достаточно высокой степенью автономности, представляется неадекватным и несправедливым приписывание полной ответственности человеку, поскольку он не всегда может выполнять экспертную функцию.

Таким образом, основой для построения концепции, замещающей классическую парадигму ответственности, выступает отказ от антропоцентризма и инструментализма в отношении анализа отношения «человек – ИИ» по причине несоответствия их предпосылок и понятийного инструментария действительному способу функционирования ИИ. Взамен в статье предлагается разработка анализа идеи целостной социотехнической системы, формируемой человеком и ИИ. Такая концепция позволяет ставить вопрос о распределении ответственности за последствия решений между всеми участниками системы «человек – ИИ».

Одна из ключевых особенностей функционирования такой социотехнической системы заключается в том, что оценка выработанных в ней решений может превосходить возможности когнитивных способностей отдельного человека (основного источника ответственности в классическом подходе). Данная особенность имеет место вследствие того, что в сфере разработок ИИ возникают ситуации, в которых отсутствует часть информации, необходимой для полного понимания всех обстоятельств функционирования технологии. Технически это происходит из-за того, что процесс обучения и вывода конкретного результата, основанного на большом количестве данных, не способен проследить ни один человек. Для самих разработчиков программа в некоторой степени является черным ящиком, т.е. слишком сложной [Латур, 2013, 25]. Стоит уточнить, что такая сложность в некоторых системах ИИ является принципиально неразрешимой. Вызвано это самой спецификой ИИ, связанной с тем, что технология нацелена на имитацию высших когнитивных функций человека, задействованных в том числе в творческих процессах. Реализуемая при этом алгоритмическая автономность является причиной ограничений в объяснимости и предвидении всех следствий функционирования ИИ. Тем самым технология превосходит ограничения инструментального подхода.

Разработчики ИИ осуществляют свою деятельность в условиях недостатка информации с двух сторон: как с этической, так и с научной. Разработчик не знает и не может знать как все возможные последствия функционирования ИИ, так и все возможные способы использования технологии, поскольку ее функционал может изменяться на стадии пользования спонтанно, не всегда прогнозируемым образом. Это может вести к возникновению дополнительных этически значимых последствий. Способность ИИ к непрогнозируемым изменениям в совокупности с тем, что он по определению не обладает строго заданной структурой и формой, так как каждый акт его функционирования привносит изменения в параметры, из которых состоит эта технология, позво-

ляет утверждать, что он обладает свойством текучести. Значимость применения этого концепта в рамках описания модели распределенной ответственности и его связи с атрибутированием агентности технологии ИИ будет описана в следующем разделе.

Новые технологии предполагают не просто автоматизацию некоторых производственных процессов, но их существенную трансформацию в логике взаимодействия человека и ИИ, позволяющую говорить о появлении нового типа социотехнической системы «человек – ИИ». Это такой тип системы, в которой определенной степенью автономии обладают и искусственный агент, и человек: как разработчик, так и пользователь. Все элементы системы (включая искусственных агентов) могут взаимодействовать друг с другом, корректируя действия друг друга и совместно влияя на итоговый результат взаимодействия [Müller, 2021, 585]. Исключение одного из участников взаимодействия существенным образом влияет на итоговый результат.

Как в такой системе может быть реализована ответственность за действия и какой становится ответственность человека в социотехнической системе, включающей ИИ?

Социотехническая система ИИ

Исходным тезисом нашей статьи выступает утверждение о том, что технологии функционируют не как автономный артефакт, а как сложная гетерогенная социотехническая система, конструируемая человеческими и нечеловеческими агентами. Этот тезис не является новым и, более того, после своего возникновения он уже претерпел некоторые трансформации. Впервые он появляется в контексте вопросов о производстве и управлении системами знаний (от теоретического производства до технического применения) с акцентом на гуманистические ценности и междисциплинарность подходов в исследовании и организации и был выдвинут во многом для того, чтобы продемонстрировать важность учета социальных аспектов в функционировании технологий [Cherns, 1976].

Однако в настоящее время можно отметить иную тенденцию использования данного тезиса, связанную, наоборот, с акцентом на активной роли технологий в работе системы. Эту тенденцию можно объяснить следующим образом. Усложнение технологий и увеличение их влияния на человеческое познание заставляют акцентировать внимание на агентности нечеловеческих акторов, которая выходит за рамки только лишь инструментального отношения к технологиям. Такого рода акценты обнаруживаются в понятиях распределенного познания (distributed cognition) [Hutchins, 1995] и расширенного сознания (extended mind) [Clark, Chalmers, 1998], а также попадают в поле внимания исследователей в области STS (акторно-сетевые теоретики и исследования инфраструктур [Star, 1999]). Внимание к агентности технологий позволяет увидеть и учесть вариативность их функций и трансформацию их возможностей в процессе использования. При этом современное состояние понятия социотехнической системы не предполагает главенствующей роли

агентов определенного типа. Так, Бруно Латур критикует как позицию социологизма, отводящую определяющую роль человеческим агентам, так и симметричную позицию технологизма, согласно которой определяющая роль в социотехнических системах отводится технологиям. Он предлагает заменить обе эти позиции утверждением гибридности социотехнической системы, человеческие и нечеловеческие отношения в которой переплетены до такой степени, что различение одних от других требует колоссальных трудов [Latour, 1992] и вряд ли будет являться целесообразным.

Принятие указанного тезиса в его современном значении влечет за собой необходимость признания за технологиями возможности преподносить сюрпризы даже своим разработчикам и расширять свой изначальный функционал [Кузнецов, 2020, 169]. Указание на такое свойство технологии можно обнаружить уже в текстах Норберта Винера: «…не всегда будет верным предположение, будто все новые функции машины были заблаговременно и явно предусмотрены ее конструктором» [Винер, 2018, 34]. Эта способность у современных технологий только возросла, так как новейшие системы ИИ демонстрируют высокий уровень автономности функционирования, и зачастую даже сами разработчики не способны исчерпывающим образом объяснить, как именно был получен тот или иной результат. Некоторые исследователи выделяют такое свойство технологий, как «эпистемическая замутненность» (epistemic opacity) [Durán, Jongsma, 2021]: хотя в целом разработчик понимает принципы работы системы, объяснимость конкретных выходных данных зачастую оказывается под вопросом, особенно в рекуррентном, генеративном ИИ. Усложнение функционирования технологий при рассмотрении во взаимодействии с человеком приводит к еще большей неразличимости вклада человеческих и нечеловеческих агентов в конечный результат работы ИИ. Поэтому анализ современных систем ИИ влечет переход от классической концепции технологий к рассмотрению их в качестве социотехнических систем.

Внимание к агентности сложных технологий, подобных ИИ, позволяет допустить, что они составляют с людьми распределенную социотехническую систему, центр которой заранее не определен, и ответственность за действия которой проблематична. Сама возможность социотехнической системы быть распределенной задается способностью как человеческих, так и нечеловеческих агентов быть активными участниками общей познавательной деятельности. Речь идет о концепте распределенного познания, в котором схватывается постоянная смена центра связей гетерогенных участников социотехнических систем, осуществляемая в процессе совместной деятельности человеческих и нечеловеческих агентов[1]. Именно проблематичность выделения центра связей и, следовательно, субъекта ответственности, заставляет нас предпочесть понятие «распределенности», а не понятие «расширенности» для характеристики социотехнической системы. О проблематичности включения технологий в социальные взаимодействия во втором случае свидетельствует то, что авторы, использующие концепт «расширенного сознания», предлагают гово-

[1] Подробнее об этом см.: [Hutchins, 1995; Шиповалова, 2019].

рить либо о социальном, либо о техническом расширении сознания [Pritchard, 2016]. Причард, трактуя расширение когнитивной системы посредством технологий, рассматривает последние лишь как нейтральные средства познающего человека, которые не производят непредсказуемых эффектов. От такого инструментального подхода к технологиям мы полагаем необходимым отказаться при рассмотрении их роли в функционировании распределенных социотехнических систем с использованием ИИ.

Отметим, что понятие распределенности в том виде, в котором оно вводится и разрабатывается Э. Хатчинсом, предполагает возможность использования его при характеристике системы. Так, Хатчинс в этом случае говорит о «гибридной системе» [Hutchins, 2014, 35]. Он понимает и использует концепт гибридности в том же значении, что и Латур, и другие исследователи акторно-сетевой теории (АСТ), а именно как способ описания действительного функционирования сложных социотехнических систем, в которых до неразличимости смешиваются технические и социальные отношения. Мы также сохраняем само понятие системы, предполагая ее гибкость и распределенный характер, включающий агентность технологий, в том числе ИИ (без наделения его подлинной интенциональностью). В таких системах оказывается невозможным выстраивание четкой иерархии по какому-либо критерию, а потому относительно них можно говорить о горизонтальном характере связей.

Помимо гибридности в анализе технологий, в исследованиях АСТ используются концепты текучести и агентности. В определении этих концептов мы следуем А. Мол и М. Лаэт. В контексте достижения цели нашей статьи в первом понятии выделяется отсылка к распределенному характеру агента, производящего и использующего технологии, а также способность самой технологии, включенной в такое распределение, к функциональным и сущностным, но при этом незапланированным и спонтанным трансформациям. Во втором для нас важна возможность отличить действенность (agency) актора от активности субъекта, где активность предполагает постановку цели, наличие намерения, тогда как агентность – только то, что действие производит определенные эффекты [Де Лаэт, Мол, 2017, 176]. Агентность может характеризовать как сложную технологию ИИ, так и всю социотехническую систему, функционирующую с ее использованием. Об активности далее мы будем говорить только в случае человеческой активности, предполагающей интенциональность[2].

Возможность применения концепта текучести в описании ИИ как социотехнической системы задается тем, что между человеком и ИИ в настоящее время стирается граница, поскольку последний является в некотором смысле продолжением человеческой рациональности. Более того, зачастую выполнение им когнитивных задач (например, по обработке информации и генерации некоторых решений) значительно превышает человеческие возможности, что во многом и обуславливает стремительный рост внимания к этой технологии. Мол и Лаэт вводят концепт текучести, когда описывают функционирование

[2] Об использовании данной терминологии в применении к социотехническим системам с использованием цифровых технологий см.: [Шиповалова и др., 2021, 76].

втулочного водяного насоса, распространенного в Зимбабве. После проведенного ими эмпирического исследования, они пришли к выводу о том, что у зимбабвийского насоса нет одного зафиксированного режима существования, вместо этого он обладает множеством степеней и оттенков функционирования, обусловленных, с одной стороны, самой технологией, с другой – природным и социальным контекстом. Помимо указанного ключевого аспекта текучести важным является также тот факт, что текучесть – не надстроенный конструкт, она встроена в саму технологию. Более того, исследовательницы демонстрируют неотъемлемость этой характеристики насоса, так как именно ее наличие обеспечивает работу насоса в сложных природных и социальных условиях Зимбабве. Текучесть позволила насосу распространиться по всей стране и обеспечить водой большую часть населения. Таким образом, текучесть – это не просто метафора, а именно социотехническое свойство, характерное только для некоторых устройств и технологий. Вместе с тем насосу также приписывается агентность [Де Лаэт, Мол, 2017, 171]. В связи с этим Мол и де Лаэт делают вывод о том, что агенты могут быть как не-рациональными и не-человеческими, так и текучими. ИИ также обладает агентностью, хотя он и не является твердой сущностью, и не имеет четких границ.

Современный ИИ – это текучая технология. Его границы неустойчивы подобно границам зимбабвийского втулочного насоса, и он также собирается во взаимодействии разнообразных социальных и политических акторов, а также трансформируется в процессе своего использования. Он масштабируем от алгоритмической до сложной социотехнической системы, состоящей из человеческих и нечеловеческих агентов. ИИ обладает даже более высокой степенью текучести, нежели насос, поскольку компьютерную программу сложнее локализовать в пространстве с привязкой к конкретным необходимым материальным артефактам. Более того, пользователем ИИ или, по крайней мере, заинтересованным лицом, испытывающим на себе эффект от результатов его деятельности, является трудно определимая группа лиц, потенциально – все общество. Подобный подход во многом оказывается причиной актуальности этического дискурса касательно ИИ, поскольку результат его внедрения и развития потенциально может затрагивать всех людей, в том числе напрямую не использующих его.

В настоящее время под понятие ИИ попадает широкий класс различных технологий, таких как средства поддержки принятия решений, голосовые помощники, языковые модели, инструменты по генерации текстов, изображений, аудио и т.д. Их активное использование в творческой деятельности человека вызывает кризис классической идеи авторства, поскольку произведение получается в результате согласованного действия распределенной социотехнической системы. Агентность ИИ заключается в его способности выступать значимым фактором при выполнении той или иной задачи, оказывающим существенное влияние на результат. Это позволяет исследователям, как, например, Гомарту и Генниону говорить скорее о событиях, нежели о действиях в рамках ACT [Gomart, Hennion, 225–226]. В то время как действие подразумевает активное действующее лицо, некоторую субъективность, событие просто случается как эффект от констел-

ляции действий различных человеческих и нечеловеческих агентов, среды в целом. Имеет место своеобразная «диффузия»: с одной стороны, ИИ в силу своей текучести оказывается распределен по «человеческому», с другой стороны, человеческое оказывается рассеяно по технологической среде, результатом чего и является функционирование социотехнической системы ИИ, которая, несмотря на свою текучесть и распределенность, обладает агентностью.

Таким образом, сама технология ИИ рассматривается нами как распределенная социотехническая система с присущими ей свойствами текучести и агентности. Текучесть позволяет определить динамику социотехнической системы, что, на наш взгляд, существенно дополняет распределенность как статическую характеристику. Такой подход к изучению ИИ избегает антропоморфизма, поскольку, во-первых, технические объекты рассматриваются без наделения их интенциональностью, однако со способностью производить эффекты своими действиями. Во-вторых, агентность технических объектов выражается в их возможности привносить в функционирование системы шумы и искажения, дополняя неопределенность, свойственную человеческим действиям. В-третьих, технические устройства рассматриваются не как замещающие человеческую ответственность, но скорее как ограничивающие ее в части ответственности за создание условий (например, сбор данных), необходимых для принятия решений и действий. Ведь в сложных ситуациях, описываемых Э. Хатчинсом, также как и в ситуациях действия системы, включающей ИИ, становятся необходимыми различные источники информации и различные – человеческие и нечеловеческие – агенты ее получения и синтеза. Все эти элементы вместе составляют содержание процесса решения сложной задачи, который невозможно осуществить силами индивидуального сознания или даже действиями отдельной группы людей без использования распределения действий в социотехнической системе, что соответствующим образом обусловливает функционирование ответственности в такой системе.

Распределенная ответственность в социотехнической системе ИИ

Какую форму обретает отношение ответственности в социотехнической системе? Разработчики могут наделять агентностью ИИ и переносить или не переносить на него долю ответственности. При этом проблема возникает с отсутствием у ИИ способности к моральной ответственности. На текущем уровне разработок ИИ складывается следующая ситуация: системы ИИ могут автономно принимать решения и совершать различные действия, влияющие на разного рода процессы, но при этом не могут быть агентами моральной ответственности, так как у них отсутствуют необходимые для этого свойства (интенциональность, целеполагание, осознанность и т.д.). В результате такая ситуация приводит к проблеме ответственности за последствия функционирования ИИ: кто несет ответственность, если у ИИ отсутствует моральная агентность,

а разработчики и пользователи не могут нести полную ответственность, так как они не контролирует весь процесс работы ИИ и не могут предвидеть все последствия его функционирования?

Широкое распространение получило представление о пробеле в ответственности (responsibility gap) в сфере ИИ, поскольку само устройство технологии связано с невозможностью выделить конкретное ответственное лицо, обладающее и интенциональностью, и имеющее мотивы, реализация которых приводит к морально значимым последствиям [Coeckelbergh, 2020; Matthias, 2004]. Наличие данной проблемы вызывает опасения в силу того, что в ходе реализации технологии могут возникать морально значимые события (например, неосознанная и систематическая дискриминация какой-либо социальной группы в ходе эксплуатации системы ИИ), за которые не наступает ответственность. Так, в случае с проявлением дискриминации предполагается, что среди разработчиков, распространителей или пользователей системы отсутствует тот, кто осознанно стремился бы осуществить дискриминацию: сама система ИИ не продуцирует какие-либо намерения, а место ответственного субъекта оказывается пустующим там и тогда, когда остается запрос на установление социального справедливого порядка. Кажется, что комбинация подхода к ИИ как социотехнической системе с позицией наличия пробела в ответственности ведет к девальвации моральной ответственности, к возникновению понятийных ограничений для определения моральной ответственности в контексте ИИ. Однако здесь стоит учесть, что тогда, когда речь заходит о тупике пробела в ответственности, моральная ответственность анализируется в рамках диалоговой модели, предполагающей ответственность как отношение, возникающее между двумя агентами, обладающими осознанием существующих обязательств, способностью взаимно признавать значимость друг друга, осознанно реализовывать в действиях ценностные установки и т.п. [Heinrichs, 2022]. Так, Кокельберг прямо говорит о проблеме «многих рук», имеющейся в случае беспилотных автомобилей и препятствующей решению проблемы ответственности. Очевидно, что само устройство технологии ИИ всегда подразумевает участие множества лиц: это и разработчики, и тестировщики, и распространители технологии, и ее пользователи, и инженеры аппаратной части. Их совокупная деятельность, очевидно, может оказывать этически значимое воздействие и на отдельных людей, и на общество в целом. Таким образом, диалоговая модель ответственности неприменима к анализу ИИ.

Кажется, что социотехническая система в целом не обладает моральным значением, поскольку в нее входят неоднородные элементы: как интенциональные агенты, носители морального сознания, так и технологические сущности, не обладающие сознанием. Сложная комбинация и система взаимосвязей агентов в социотехнической системе препятствует реализации прямого определения ответственных лиц среди интенциональных агентов, которым ответственность может быть вменена согласно диалогической модели. Скорее создаются условия, при которых человеческие агенты социотехнической системы имеют возможность перекладывать бремя полной ответственности за функционирование целой системы друг на друга. Возможность этого обу-

словлена тем, что вероятностное устройство ИИ, наличие связанной с этим проблемы объяснимости [Buijsman, 2022, 563–564] делают невозможным закрепление ответственности только за разработчиками или только за пользователями системы.

Для разрешения данного затруднения может быть использован понятийный аппарат концепции распределенной морали, разработанный Лучано Флориди на основе его исследований мультиагентных систем, акторами в которых могут являться люди, искусственные и гибридные объекты [Floridi, 2012, 729–731]. В своих исследованиях он частично исходит из предпосылок, на которых основывается и данная статья: увеличивающегося превалирования и автономии искусственных агентов и гибридных систем. Следует отметить, что последний термин оказывается близок к тому, что в данной статье обозначено как социотехническая система. Согласно Флориди, моральность искусственных объектов проистекает из их агентности. Последняя же в свою очередь обуславливается их интерактивностью (способностью вписываться в сеть различных взаимодействий), автономностью (способностью быть независимыми от человека в процессе функционирования) и адаптивностью (способностью вписываться в различную среду) [Floridi, Sanders, 2004, 349]. Иными словами, то, на что он опирается в данном случае, соответствует концептам текучести и распределенного познания, рассмотренным выше в этой статье.

Идеи Флориди также важны для формирования концепции распределенной ответственности, поскольку он пытается преодолеть антропоцентризм, присутствующий в классической этической парадигме [Floridi, 2012, 728]. Рассматривая понятие агентности, он демонстрирует, что ни сознание, ни проистекающие из него интенциональность и свободная воля не являются необходимыми для признания агентности. Флориди и Сандерс называют это «неразумной моралью» (mindless morality) [Floridi, Sanders, 2004, 349]. Кроме того, феномен «неразумной морали» не является специфическим лишь для искусственных агентов, он предполагает распространение данного принципа и на человеческих агентов.

Подход Флориди допускает, что каждое отдельное действие каждого отдельного агента мультиагентной системы может быть морально нейтральным или иметь минимальное моральное значение [Floridi, 2012]. Однако также, как в системе распределенного познания, знание как таковое не производится отдельным субъектом, но принадлежит всей системе взаимосвязанных участников процесса познания, которую осуществляют в равной мере сознательные и технические агенты, в системе распределенной морали морально нагруженное действие не имеет определенного источника или автора. Вместо этого оно является результатом организации взаимодействий в сложной мультиагентной или социотехнической системе. Распределенная мораль, таким образом, представляет собой свойство сложных систем [Ibid., 729]. Она предполагает, что результат работы всей системы может носить как нейтральный, так и существенный (положительный или отрицательный) эффект, в то время как каждый агент системы совершает морально нейтральные либо минимально значимые действия. Это преломление осуществляется именно в ходе взаимодействия

агентов в условиях недостатка информации и когнитивной ограниченности каждого из них. Таким образом, эффект взаимодействия агентов не равен сумме их отдельных действий.

Должным образом отрегулированная координация, кооперация отдельных агентов системы может приводить к достижению желаемых морально значимых результатов. Их оценка осуществляется не на уровне декларируемых мотивов сознательных агентов в социотехнической системе, а на уровне качества влияния системы на объекты ее воздействия.

Дальнейшее построение модели распределенной ответственности для социотехнической системы ИИ осуществляется посредством адаптации объект-ориентированной морали Флориди [Floridi, 2016]. Моральное измерение системы вводится не через субъекта, автора действия, обладающего качествами морального агента (интенциональность, воля, сознание), но через объект воздействия, к которому могут относиться и отдельные люди, и определенные социальные группы, и человечество в целом, и экосистема планеты. Технические и технологические сущности, функционирование которых может приводить к морально значимым изменениям состояний объекта воздействия, в конце концов получают моральное значение.

Распространение ответственности в рамках социотехнической системы осуществляется по пути обратного распределения: если совокупность действий, не содержащих никаких морально значимых коннотаций (например, написание кода, токенизация текстового массива и т.п.), может приводить к морально значимым последствиям, то и ответственность за состояние объекта воздействия может обратным образом распространяться на всю систему в целом. Отсутствие явных мотивов, намерений, наличие которых обычно служит основанием для вменения ответственности субъекту действия, не препятствует обратному распределению ответственности в социотехнической системе в силу того, что она изначально строится на не диалогических отношениях агентов, намеренным образом реализующих те или иные моральные ценности. Сложные взаимодействия в социотехнической системе служат причиной возникновения морально значимого состояния объекта воздействия данной системы и приводят к реализации имплицитные моральные ценности. Следовательно, в области этики в сфере ИИ акцент должен ставиться не на намерениях тех или иных заинтересованных лиц, а на значении возможного воздействия технологии на морально значимые объекты. Моральная ответственность в рамках распределенного подхода касается всей социотехнической системы в целом, что ведет не к девальвации ответственности отдельных агентов системы, но, напротив, к ее интенсификации, поскольку ответственным в равной мере оказывается каждый агент. Распределенная ответственность предполагает, что каждый участник взаимодействия несет ответственность за наступающее событие, поскольку без их взаимодействия события бы не было.

Итак, идея распределенной ответственности основывается на ключевых характеристиках социотехнической системы: распределенность, текучесть, агентность ее акторов. В ситуации распределенного познания имеет место ограничение когнитивных возможностей человеческих агентов. Возможно-

сти ИИ по сбору и обработке информации влияют на этот аспект. Текучесть означает в данном случае подвижность центра принятия и реализации решений в рамках социотехнической системы, невозможность локализовать когнитивный процесс решения задачи в индивидуальном сознании.

Заключение

Понятийный потенциал концепции распределенной морали Флориди был использован в статье для решения задачи непротиворечивого определения того, как осуществляются отношения ответственности в сложных социотехнических системах ИИ. Посредством анализа ИИ как текучей технологии, обладающей агентностью, в статье был обоснован тезис о том, что система «человек – ИИ» является распределенной социотехнической системой, в которой человеческие и нечеловеческие агенты выполняют совместные задачи через распределенное познание и, шире, – распределенное действие. В результате в статье была предложена и описана концепция распределенной моральной ответственности в социотехнической системе «человек – ИИ».

Строго говоря, встраивание в социотехническую систему ИИ для человеческого агента автоматически означает его включение в потенциально морально ненейтральный контекст и должно вести к осознанию им возможности наступления моральной ответственности. В связи с этим особое значение приобретает способность человека прогнозировать наступление определенных последствий совершаемых им действий. Из-за наличия у ИИ свойств текучести и агентности, а также того, что он является социотехнической системой, состоящей из разнородных человеческих и нечеловеческих агентов, взаимодействующие с ним люди могут переносить с себя на технологию ответственность за потенциально большое количество процессов. Важным выводом статьи является то, что в силу распределенного характера ответственности в социотехнических системах ИИ, человеческим агентам следует повышать степень своей собственной моральной бдительности, чтобы воспрепятствовать возникновению ситуаций, в которых снимается их личная моральная ответственность.

Раскрытый в статье подход к моральному определению распределенной ответственности может быть дополнен разработкой правовых механизмов ответственности в сфере ИИ, выделяющих долевую ответственность тех или иных агентов социотехнической системы, определяемую на основе договорных отношений агентов или иных нормативно-правовых актов. Распределенная моральная ответственность здесь выступает основанием для возможного наделения агентов правовой ответственностью, но сама она при этом не интерпретируется в терминах долей в силу того, что моральное значение социотехническая система получает без опоры на осознанность, интенциональность отдельных входящих в нее агентов.

Distributed Moral Responsibility in the Field of Artificial Intelligence

Andrei S. Glukhovskii, Aleksei D. Durnev, Daria V. Chirva

Andrei S. Glukhovskii – ITMO University. 49, bld., A Kronverksky Pr., St. Petersburg, 197101, Russia.

ORCID 0000-0003-0052-0180
e-mail: aglukhovsky@itmo.ru

Aleksei D. Durnev – ITMO University. 49, bld. A, Kronverksky Pr., St. Petersburg, 197101, Russia.

ORCID 0000-0002-1248-4685
e-mail: a.durnev.ph@gmail.com

Daria V. Chirva – ITMO University. 49, bld. A, Kronverksky Pr., St. Petersburg, 197101, Russia.

ORCID 0000-0001-7804-8803
e-mail: dvchirva@itmo.ru

The main goal of the article is to justify the optimality of Luciano Floridi's concept of distributed morality for describing a model of the relationship of responsibility in complex human – AI systems. The high level of its autonomy in functioning does not correspond to the classical instrumentalist approach to technology, as it turns out to be capable of fundamentally uncontrollable actions in the process of performing the task assigned to it. For the same reason, the classical definition of moral responsibility does not adequately correspond to the actual situation in the field of artificial intelligence development, which often leads to problems that hinder the development of this technology. Therefore, the article conceptualizes a new technology-appropriate way of describing the relationship of responsibility in the field of artificial intelligence ethics. To achieve this goal artificial intelligence is considered as a participant in various interactions, including social ones, possessing agency, but at the same time as not attributed with true intentionality, motivation, awareness and other properties that are necessary for attributing agency in the classical model of the relationship of responsibility. In addition to referring to Floridi's description of agency criteria, the justification of such a consideration of AI is based on Annamarie Mol and Marianne de Laet's concept of fluidity which is developed by them, among others, to denote the ability of technologies to possess agency without having the properties considered necessary for attributing agency in the classical model of the role of technologies in social processes.

AI technology is analyzed as a complex heterogeneous sociotechnical system made up of both human and non-human agents. The risk of leveling the significance of responsibility in such a system is overcome at the conceptual level by adapting the approach of Floridi's patient-oriented morality. Changing the emphasis from the subject of action to the object of influence gives moral significance to the entire sum of morally neutral actions that are performed by agents within the framework of a sociotechnical system. Since it produces morally significant effects and impacts, it also acts as a bearer of the relation of responsibility which, as shown in the article, applies to each agent of the system equally thereby intensifying moral responsibility in the field of AI.

Keywords: ethics of artificial intelligence, distributed cognition, distributed responsibility, sociotechnical system, agency, patient-oriented morality

Литература / References

Винер Н. Корпорация «Бог и голем» / Пер. с англ. В. Желнинова. М.: Изд-во АСТ, 2018.

Weiner, N. *Korporatsiya "Bog i golem"* [God & Golem, Inc.], transl. by V. Zhelninov. Moscow: AST Publ., 2018. (In Russian)

Де Лает А., Мол А. Зимбабвийский втулочный насос: механика текучей технологии // Логос. 2017. Т. 27. № 2. С. 171–232.

De Laet, M., Mol, A. "Zimbabviiskii vtulochnyi nasos: mekhanika tekuchei tekhnologii" [The Zimbabwe Bush Pump: Mechanics of a Fluid Technology], *Logos*, 2017, Vol. 27, No. 2, pp. 171–232. (In Russian)

Кузнецов А.Г. Туманности нейросетей: «черные ящики» технологий и наглядные уроки непрозрачности алгоритмов // Социология власти. 2020. Т. 32. № 2. С. 157–182.

Kuznetsov, A.G. "Tumannosti neirosetei: «chernye yashchiki» tekhnologii i naglyadnye uroki neprozrachnosti algoritmov" [Neural Network Nebulae: 'Black Boxes' of Technologies and Object-Lessons from Opacities of Algorithms], *Sociology of Power*, 2020, Vol. 32, No. 2, pp. 157–182. (In Russian)

Латур Б. Наука в действии: следуя за учеными и инженерами внутри общества / Пер. с англ. К. Федорова, науч. ред. С. Миляева. СПб.: Изд-во Европейского ун-та, 2013.

Latour, B. *Nauka v deistvii: sleduya za uchenymi i inzhenerami vnutri obshchestva* [Science in Action: How to follow Scientists and Engineers through Society], transl. by K. Fedorova, ed. by S. Milyaeva. St. Petersburg: European Univ. Publ., 2013. (In Russian)

Шиповалова Л.В. Распределенное познание – аналитика и проблематизация концепта // Цифровой ученый: лаборатория философа. 2019. Т. 2. № 4. С. 175–190.

Shipovalova, L.V. "Raspredelennoe poznanie – analitika i problematizatsiya kontsepta" [Distributed Cognition – Analytics and Problematization of the Concept], *Tsifrovoi uchenyi: laboratoria filosofa*, 2019, Vol. 2, No. 4, pp. 175–190. (In Russian)

Шиповалова Л.В., Чернышева Л.А., Гизатуллина Э.Г. Цифровые технологии управления в действии, или об активности граждан вокруг платформы «Активный гражданин»// Социология науки и технологий. 2021. Т. 12. № 1. С. 71–87.

Shipovalova, L.V., Chernysheva, L.A., Gizatullina, E.G. "Tsifrovye tekhnologii upravleniya v deistvii, ili ob aktivnosti grazhdan vokrug platformy 'Aktivnyi grazhdanin'" [Digital Governance Technologies in Action, or On the Activity of Citizens around the Platform "Active Citizen"], *Sotsiologia nauki i technologii,* 2021, Vol. 12, No. 1, pp. 71–87. (In Russian)

Философия ответственности / Под ред. Е.Н. Лисанюк, В.Ю. Перова. СПб.: Наука, 2014.

Filosofiya otvetstvennosti [Philosophy of Responsibility], eds. E.N. Lisanyuk, V.Yu. Perov. St. Petersburg: Nauka Publ., 2014. (In Russian)

Buijsman, S. "Defining Explanation and Explanatory Depth in XAI", *Minds and Machines*, 2022, Vol. 32, pp. 563–584.

Cherns, A. "The Principles of Sociotechnical Design", *Human Relations*, 1976, No. 29 (8), pp. 783–792.

Clark, A., Chalmers, D. "The Extended Mind", *Analysis,* 1998, Vol. 58, Issue 1, pp. 7–19.

Coeckelbergh, M. *AI Ethics*. Cambridge, Massachusetts: The MIT Press, 2020.

Dignum, V., Dignum, F. "Agents are Dead. Long Live Agents!", *Proc. of the 19th International Conference on Autonomous Agents and Multi Agent Systems*, eds. N. Yorke-Smith, B. An, A.E.F. Seghrouchni, G. Sukthankar. Richland, SC: IFAAMS, 2020, pp. 1701–1705.

Durán, J.M., Jongsma, K.R. "Who is Afraid of Black Box Algorithms? On the Epistemological and Ethical Basis of Trust in Medical AI", *Medical Ethics*, 2021, No. 47, pp. 329–335.

Gomart, E., Hennion, A. "A Sociology of Attachment: Music Amateurs, Drug Users", *Actor Network Theory and After*, eds. J. Law, J. Hassard. Oxford: Blackwell Publishers, 1999, pp. 220–247.

Heinrichs, J.-H. "Responsibility Assignment won't Solve the Moral Issues of Artificial Intelligence", *AI Ethics*, 2022, No. 2, pp. 727–736.

Hutchins, E. *Cognition in the Wild,* Cambridge, Massachusetts: MIT Press, 1995.

Hutchins, E. "The Cultural Ecosystem of Human Cognition", *Philosophical Psychology*, 2014, Vol. 27 (1), pp. 34–49.

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 of the Royal Society*, 2016, Series A, No. 374 (2083), pp. 1–13.

Flodiri, L., Sanders, J.W. "On the Morality of Artificial Agents", *Minds and Machines*, 2004, Vol. 14, pp. 349–379.

Latour, B. "Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts", *Shaping Technology / Building Society: Studies in Sociotechnical Change*, eds. W. Bijker, J. Law. London: MIT Press, 1992, pp. 225–259.

Matthias, A. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata", *Ethics and Information Technologies*, 2004, Vol. 6, pp. 175–183.

Müller, V.C. "Is it Time for Robot Rights? Moral Status in Artificial Entities", *Ethics and Information Technologies*, 2021, Vol. 23, pp. 579–587.

Pritchard, D. "Intellectual Virtue, Extended Cognition and the Epistemology of Education", *Intellectual Virtues and Education: Essays in Applied Virtue Epistemology*, ed. by J. Baehr. London: Routledge, 2016, pp. 113–127.

Star, S.L. "The Ethnography of Infrastructure", *American Behavioral Scientist*, 1999, Vol. 43, No. 3, pp. 377–391.