И. Ю. АЛЕКСЕЕВА

ЭТИКА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА КАК ПРИКЛАДНАЯ ЭТИКА*

Этика искусственного интеллекта рассматривается главным образом как направление научных исследований, оформляющееся в XXI в. и объединяющее как ученых, работающих в области искусственного интеллекта (ИИ), так и философов. Это направление охватывает не только вопросы, касающиеся поведения и нравственного сознания людей, создающих и использующих системы искусственного интеллекта, но также вопросы «поведения» таких систем. Будучи одним из видов прикладной этики, этика ИИ не избегает осмысления глубинных философских проблем сущности человека, его свободы, статуса и перспектив в технологизирующемся мире. В статье уделено внимание взаимосвязям и пересечениям этики ИИ с родственными прикладными этиками – компьютерной, инженерной, «технической». Отмечается обусловленное прогрессом в области ИИ расширение сферы применимости термина «агент» в русском философском языке. Присоединяясь к позиции, согласно которой субъектом деятельности и ответственности за работу систем ИИ может быть только человек, автор обсуждает условия действенности способов этической регуляции, в том числе этических кодексов.

Ключевые слова: *искусственный интеллект, ИИ, этика искусственного интеллекта, философия техники, прикладная этика, искусственный агент, система ИИ, ответственность, этическая регуляция искусственного интеллекта, этический кодекс.*

The ethics of Artificial Intelligence is considered primarily as a field of scientific research that has been taking shape in the 21st century, uniting both scientists working in the field of artificial intelligence (AI) and philosophers. This direction covers not only issues related to the behavior and moral con-

* ***Для цитирования:*** Алексеева И. Ю. Этика искусственного интеллекта как прикладная этика // Философия и общество. 2024. № 3. С. 69–85. DOI: 10.30884/jfio/2024.03.06.

For citation: Alekseyeva I. Yu. Ethics of Artificial Intelligence as Applied Ethics // Filosofiya i obshchestvo = Philosophy and Society. 2024. No. 3. Pp. 69–85. DOI: 10.30884/jfio/2024.03.06 (in Russian).

sciousness of people who create and use artificial intelligence systems, but also issues of the "behavior" of such systems. As one of the types of applied ethics, AI ethics does not ignore understanding the deep philosophical problems of the essence of human, freedom, status and prospects of humans in a technologizing world. The article pays attention to the relations and intersections of AI ethics with related applied ethics – computer ethics, engineering ethics, "technical" ethics. Due to the progress in the field of AI, the expansion of the scope of application of the term "agent" in the Russian philosophical language is noted. Adhering to the position that only a person can be the subject of activity and responsibility for the operation of AI systems, the author discusses the conditions for the effectiveness of methods of ethical regulation, including ethical codes.

Keywords: *Artificial Intelligence, AI, ethics of Artificial Intelligence, philosophy of technology, applied ethics, artificial agent, AI system, responsibility, ethical regulation of Artificial Intelligence, ethical code.*

Философские дискуссии, касающиеся искусственного интеллекта (ИИ), его возможностей и перспектив, имеют достаточно долгую историю. В середине XX в., в период, который мы называем «осевым временем ИИ» [Алексеев, Алексеева 2021], было выдвинуто немало философских идей, до сих пор сохраняющих актуальность. Эти идеи относились главным образом к теории познания (гносеологии) и дали серьезный импульс ее развитию. Сегодня на первый план в дискуссиях вокруг ИИ выдвигается другой раздел философского знания – этика. Этика ИИ как новое направление исследований и новый член в растущем семействе так называемых прикладных этик достаточно внятно заявила о себе в конце 10-х – начале 20-х гг. нынешнего столетия.

1. Этика ИИ: новый взгляд на искусственную интеллектуальную систему

В 2018–2019 гг. журнал «Философия и общество» опубликовал серию статей по этой тематике. Первой в данной серии вышла статья В. Э. Карпова, П. М. Готовцева и Г. В. Ройзензона «К вопросу об этике и системах искусственного интеллекта» [Карпов и др. 2018]. Авторы статьи, работающие в области технических наук и занятые созданием систем искусственного интеллекта, предупредили, что «менее всего хотели бы вторгаться в область философии, профессионально занимающейся вопросами этики» [Там же: 86] и выдвинули на первый план проблему поиска и выработки таких этических императивов и норм, которые могут трактоваться как эври-

стики, используемые искусственной интеллектуальной системой (ИИС) в ходе планирования, целеполагания, «при совершении выбора того или иного действия, формирования системы оценок, целевых функций и прочего» [Карпов и др. 2018: 86]. Таким образом, был поставлен принципиально новый для академической этики вопрос об участии этого раздела философского знания в создании искусственных агентов, поведение которых определяется в числе прочего техническими «проекциями» нравственных установок и императивов. Принципиальное отличие этики ИИ от других видов прикладных этик проявляется в том, что если последние занимаются вопросами норм и установок, регулирующих принятие решений людьми, то этика ИИ включает в сферу рассмотрения не только основания решений и поведения человека, но также основания выбора, решений и поведения технической системы – прежде всего, автономной ИИС. В таком контексте выглядит закономерным вопрос, вынесенный в заглавие одной из работ В. Э. Карпова – «Может ли робот быть моральным агентом?» [Karpov 2020].

Подобные вопросы имели все основания быть понятыми как вызов ученым-этикам со стороны разработчиков ИИС. Ответом стала статья А. В. Разина «Этика искусственного интеллекта». Автор связывает статус морального агента с принципиальной возможностью ошибки. «Этика, – пишет А. В. Разин, – непосредственно начинается тогда, когда появляется способность реагировать на собственные ошибки, осуществлять рефлексию поведения, учитывая при этом мнения других людей. Такая же принципиальная возможность ошибки должна быть заложена и в работу искусственного интеллекта, чтобы можно было говорить о его этике в собственном смысле слова. Должны быть также выполнены условия коммуникации машин, их взаимных оценок и наличия у них феноменального опыта» [Разин 2019: 57]. А. В. Разин относит к области этики ИИ как прагматические вопросы, связанные в конечном счете с разработкой стандартов и сертификацией ИИС, так и общие проблемы свободы воли, ответственности, риска, раскрывающиеся новыми гранями в новых технологических контекстах. Признавая необходимость определенных этических ограничений, А. В. Разин обращает внимание и на опасность неоправданных ограничений, способных затормозить прогресс в области создания искусственных интеллектуальных систем, полезных для человека.

В статье Ф. Г. Майленовой «Люди и роботы: сбывающиеся прогнозы. Шаг длиной в столетие» [Майленова 2019] современные

этико-психологические проблемы, связанные с развитием ИИ, рассматриваются в историческом и футурологическом контекстах – от написанной в 2020 г. пьесы Карела Чапека «РУР» (название пьесы – аббревиатура названия воображаемой фирмы «Россумские универсальные роботы») до перспективы появления «киборгов» в результате развития практики имплантации электронных устройств в тело человека. Автор справедливо подчеркивает значимость использования технологий для развития интеллекта человека при сохранении творческого начала и человеческой сущности людей.

Следует отметить, что этика искусственного интеллекта имеет общие части не только с другими видами прикладных этик, но и с другими разделами философского знания, – например, с философией науки и техники. Этика ИИ включает и вечные вопросы «философии вообще» – о сущности человека, его месте в мире, о свободе и необходимости, разуме и вере; эти вопросы заново переосмысливаются в новых условиях на новом этапе научно-технологического и социального развития. Как своего рода философское завещание читаются сегодня слова А. И. Ракитова, завершающие опубликованную в год смерти ученого статью «Философия, автоматы, роботы и зримое будущее». Настаивая на необходимости осознания того, что в будущем решения, принимаемые ИИС, смогут радикально изменить социальную жизнь и оценку человеком своего «человеческого начала», А. И. Ракитов утверждал, что «роботизация, автоматизация и развитие всех форм искусственного интеллекта должны стать центральной проблемой философского дискурса современности» [Ракитов 2019: 47].

Вопрос о субъектности (квазисубъектности) искусственной интеллектуальной системы, об ограничениях на решения, которые могут быть доверены такой системе, обсуждался как этический и эпистемологический еще в середине XX в. [Алексеева 2020]. Сегодня подобные вопросы ставятся в зависимость от понимания природы и сущности морального агента [Перов, Головков 2022: 97]. Следует подчеркнуть, что вопрос о соотношении субъекта и агента деятельности является отнюдь не праздным для русского философского языка, испытывающего на протяжении нескольких десятилетий все более заметное влияние английского. Традиционно советские, а затем российские ученые, работающие в области теории познания, ставили в соответствие английскому слову “agent” слово «субъект», когда включали в сферу внимания англоязычные работы, относящиеся к проблемам субъекта познания и деятельности. Что же

касается этики, то здесь сложилась иная ситуация. В русских этических текстах термин «агент» употребляется достаточно давно и широко. Впрочем, в последние годы и в теории познания обсуждается онтолого-гносеологический статус агента как дополняющего классическую субъект-объектную схему, причем это обсуждение связано с развитием цифровых технологий. Речь в данном случае не идет о необходимости следовать за английским словоупотреблением, тем более что и там возникают недоразумения. Ярким примером последнего может служить ситуация с книгой «Сообщество разума» (“Society of Mind”) [Минский 2018], написанной М. Минским – ученым, внесшим значительный вклад в развитие искусственного интеллекта. В указанной книге Минский представил сознание человека как состоящее из множества агентов, но это вызвало настолько сильное недоумение англоязычных читателей, что впоследствии автор переименовал «агентов» в «ресурсы» [Его же 2020].

Этика искусственного интеллекта, охватывая проблемы, относящиеся ко всем видам ИИС, включает в себя и вопросы, касающиеся интеллектуальных роботов [Середкина 2020]. Эти вопросы иногда характеризуют как относящиеся к «робоэтике», или этике робототехники. Следует отметить, что робоэтика как направление (занимающееся прежде всего интеллектуальными роботами) была обозначена раньше, чем более широкая область этики ИИ. Первый международный симпозиум по робоэтике состоялся в 2004 г. в Италии [Veruggio 2005]. Робоэтика была заявлена здесь как направление, охватывающее этические, социальные, гуманитарные и экологические аспекты робототехники. Робоэтику, как и в целом этику ИИ, невозможно отделить от философии и социологии техники, – не только потому, что этическая теория составляет часть философии, но и вследствие комплексного характера социогуманитарных проблем, относящихся к разработке и применению роботов и других ИИС [Горохов, Декер 2013].

К настоящему времени количество работ по этике ИИ, опубликованных на русском языке, отнюдь не ограничивается указанными выше. Учитывая актуальность этики ИИ, уже проведенные исследования и текущие дискуссии, можно было бы ожидать появления монографий или сборников научных статей по этой тематике. Представляется, что отсутствие таковых обусловлено не в последнюю очередь утвердившейся в нашей стране системой оценки на-

учной деятельности, когда во главу угла ставятся публикации статей в определенного рода журналах.

Что касается англоязычных книг, то следует упомянуть по крайней мере несколько изданных за последние годы. В 2020 г. вышел сборник статей «Этика искусственного интеллекта» под редакцией С. М. Ляо [Ethics… 2020]. В числе авторов – как ученые, работающие в области ИИ, так и философы. Спектр обсуждаемых проблем очень широк. Здесь и «встраивание» этики в машину, и этика беспилотных автомобилей, автономные системы в военной сфере и сексуальные роботы, и, конечно же, футурологические дискуссии о перспективах сверхинтеллекта и правах искусственных интеллектуальных систем. В том же году опубликовано «Оксфордское руководство по этике ИИ» под редакцией М. Д. Дубера, Ф. Паскуале и С. Даса [Dubber *et al*. 2020]. Этот сборник содержит статьи как по общим проблемам этики ИИ, так и по вопросам применения ИИ в разных областях – в военной сфере, биомедицинских исследованиях и здравоохранении, в правовой сфере, в управлении миграцией. Среди статей общего характера – те, что посвящены отношениям человека с ИИ, роли профессиональных норм в сфере ИИ, кодексам и стандартам, вопросам прозрачности и ответственности. В 2022 г. вышла книга известного итальянского философа Л. Флориди «Этика искусственного интеллекта» [Floridi 2022]. Автор трактует искусственную интеллектуальную систему как новый вид агента, управление которым может соответствовать или не соответствовать этическим требованиям. При этом Флориди уделяет основное внимание возможностям применения ИИ для благих целей, включая такие, как устойчивое развитие и борьба с изменением климата.

Таким образом, этика ИИ охватывает широкий круг разнородных вопросов как собственно прикладного, так и теоретического характера. Следует подчеркнуть, что этот вид прикладной этики выходит на глубинные философские проблемы, связанные с осмыслением статуса человека в технологизирующемся мире, пониманием сущности человека и перспектив человечества. В этом отношении этика ИИ имеет сходство и пересечения с биоэтикой. Тело человека и его сознание стали полем «схождения», конвергенции биомедицинских и цифровых технологий, а перспективы такой конвергенции, понимаемые как перспективы «технобиоэволюции» человека, выводят философские дискуссии в область футурологии.

2. Поведение машины и ответственность человека

Выражение «искусственный интеллект» (ИИ) используется, во-первых, для обозначения определенной сферы деятельности людей, а во-вторых, для характеристики особого рода технических систем, создаваемых в этой сфере. Искусственный интеллект в первом смысле – направление исследований и разработок, занимающееся аппроксимацией способностей, образующих естественный интеллект [Финн 2009]. Напомним, что русское слово «аппроксимация» происходит от латинского “approximare”, означающего «приближаться». Аппроксимация в технике предполагает замещение, упрощение, огрубление. Результатом аппроксимации становятся в конечном счете технические системы, которые называют искусственными интеллектуальными системами (ИИС), системами искусственного интеллекта (СИИ) или кратко – искусственным интеллектом (ИИ). Именно в этом, втором, смысле чаще всего и употребляется выражение «искусственный интеллект».

Соответственно, этика искусственного интеллекта охватывает две большие группы вопросов. Во-первых, это вопросы профессионального поведения и самосознания людей, участвующих в создании и эксплуатации систем и технологий ИИ, в принятии решений относительно разработки и использования таких систем и технологий. Речь идет об ученых, инженерах, менеджерах, других работниках, вовлеченных в соответствующие процессы. Во-вторых, это вопросы «поведения» ИИС, интеграции ИИС в человеческое общество, взаимодействия ИИС с человеком и между собой, этический статус ИИС.

Упомянутые группы вопросов взаимозависимы и переплетены друг с другом. Создание устройств и систем, которые при соблюдении надлежащих правил использования работают на благо человека и не причиняют ему вреда – общая норма профессиональной деятельности в любой сфере техники и технологий. Дискуссии по вопросам этики инженерной деятельности («инженерной этики», “engineering ethics”), как и по общим этическим вопросам, связанным с развитием техники, начались более ста лет назад. Опыт осмысления этой проблематики полезен для этики ИИ как направления, обязанного своим возникновением появлению новых видов технических устройств и технологий.

Первая из известных нам попыток теоретического осмысления этических проблем инженерной деятельности была предпринята

в начале прошлого столетия П. С. Осадчим, профессором Электротехнического института императора Александра III. Изданная в 1911 г. брошюра П. С. Осадчего «К вопросу о принципах профессиональной этики инженеров» [Осадчий 1911] стала своеобразным продолжением проходивших в VI (электротехническом) отделении Императорского русского технического общества дискуссий по поводу кодекса профессиональной этики электротехников. К настоящему времени опубликован значительный массив работ по проблематике, которую иногда обозначают как «этика технической деятельности», «этика техники» или «техническая этика». А. Грунвальд характеризует эту область как объединяющую «этическую рефлексию условий, целей и средств разработки, производства, использования и утилизации техники» [Грунвальд 2014: 26], в том числе этическую оценку разных вариантов решений, связанных с представлениями о будущем человека и общества.

Если в начале прошлого столетия самой передовой технологической областью была электротехника, то сегодня в числе таковых находятся технологии искусственного интеллекта. Конкретизация общего ориентира моральной ответственности разработчика и производителя техники применительно к созданию систем ИИ зависит не в последнюю очередь от предназначения системы, ее возможностей и функций.

Более ста лет назад в упомянутой выше фантастической пьесе чешского писателя Карела Чапека «РУР» была поднята проблема «воспитания» искусственной интеллектуальной системы, «встраивания» в нее определенных ценностей, представлений о правилах поведения и взаимодействия с человеком. Эта проблема мыслилась автором пьесы (кстати, придумавшим само слово «робот») как важная до такой степени, что одно из подразделений воображаемой корпорации, производящей «умные» универсальные машины, получило название «Институт психологии и воспитания роботов». Позже писателем-фантастом Айзеком Азимовым были сформулированы знаменитые «три закона робототехники». Первый из этих законов – «Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинен вред», второй – «Робот должен повиноваться всем приказам, которые дает человек, кроме тех случаев, когда эти приказы противоречат первому закону», третий – «Робот должен заботиться о своей безопасности в той мере, в которой это не противоречит первому или второму законам». По сути, «законы Азимова» являются этическими

установками, своего рода «нравственными императивами», которые должны быть реализованы в эвристиках, «закладываемых» разработчиком в поведение робота.

О невозможности точного выполнения «трех законов робототехники» написано достаточно много. Например, обращают внимание на «проблему рамки», связанную с трудностями отделения релевантной информации, необходимой для принятия решения, от нерелевантной в условиях, когда робот получает из окружающей среды огромные массивы данных [Середкина 2020]. Одна из важных проблем этики ИИС (не только робота) связана с неспособностью системы предвидеть последствия решений и действий – будь то действие в материальном мире, порождение текста на заданную тему или обнаружение патологии в организме больного. Однако критики «законов Азимова» предъявляют к ним требования, уместные в отношении законов физики, но не применимые к велениям нравственности. Последние, в отличие от законов природы, неизбежно нарушаются людьми (даже признающими правомерность этих велений) в силу тех или иных обстоятельств. Невозможность предвидеть все последствия собственных действий, ситуации, когда разные обязательства вступают в противоречие друг с другом и нужно определить приоритеты, наконец, необходимость выполнения воинского долга, предполагающая уничтожение противника и риски для мирного населения – все это порождает нравственные коллизии, когда выбор той или иной линии поведения во многом зависит от воли человека. В этом контексте вполне обоснованным выглядит приведенное выше утверждение А. В. Разина, что об этике ИИС можно говорить только в том случае, если ИИС обладает свободой воли. Данное утверждение можно уточнить, добавив, что речь идет не о свободе воли как таковой, а о некотором технотронном аналоге свободы воли, аппроксимации человеческой свободы воли применительно к технической системе. Однако свобода предполагает ответственность, а потому тема свободы закономерным образом приводит к теме ответственности. Последняя включает в себя множество вопросов, среди которых есть не только новые, но и такие, что имеют внушительную историю обсуждения.

Вопросы ответственности людей и организаций за работу технических систем обсуждаются достаточно давно, при этом этические составляющие переплетаются с составляющими юридическими, но не совпадают с последними. В большинстве случаев моральная ответственность не может (и не должна) совпадать с ответственно-

стью, установленной законом. К тому же есть ситуации, когда гражданско-правовая или уголовная ответственность определена, но вопрос моральной ответственности может оставаться неясным или должен быть решен отрицательно (примером служит «ответственность без вины»). За работу технических систем отвечают люди, но далеко не всегда возможно определить, кто именно и за что ответственен при неполадках в работе систем и причинении вреда человеку.

В середине 80-х гг. XX в. массовое распространение компьютеров привело к появлению такого интеллектуального направления как «компьютерная этика». В рамках этого направления философы вместе с инженерами и учеными, работавшими в сфере компьютерных технологий, рассматривали широкий круг проблем, включающий, наряду с прочими, вопросы ответственности за неполадки в работе программного обеспечения, условия доступа к частной информации, накапливаемой в базах данных, процессы централизации и децентрализации власти в компьютерную эпоху, этические аспекты интеллектуальной собственности и коммерческой тайны [Ethical… 1985; Алексеева, Шклярик 2007]. В русле компьютерной этики обсуждались и некоторые этические вопросы ИИ, – например, вопрос об ограничениях на принятие решений компьютером [Вейценбаум 1982; Moor 1979].

Быстрые изменения в сфере цифровых технологий, побуждающие исследователей включать в поле зрения все новые явления и процессы, способствовали возникновению новых названий для обозначения соответствующих интеллектуальных направлений, имеющих существенные пересечения с «компьютерной этикой» и наследующих значительную часть ее проблематики. Так, появились публикации по «информационной этике», «электронно-коммуникативной этике», «информационно-технологической этике», «этике в сфере информационных технологий» [Малюк и др. 2011]. Характеризуя этот процесс в 2016 г., мы писали: «Есть основания полагать, что в будущем, на фоне стремительного технологического развития, возникнут новые познавательно-ориентировочные комплексы, которые будут охватывать существенную часть обсуждаемых сегодня проблем, добавлять новые проблемы, использовать новые (или хорошо забытые старые) подходы. Главное состоит в том, что в рамках подобных познавательно-ориентировочных комплексов формулируются и изучаются значимые вопросы технологизированного бытия все более технологизирующихся людей и обществ» [Алексе-

ева, Аршинов 2016: 56]. Сегодня мы можем утверждать, что одним из видов таких познавательно-ориентировочных комплексов является этика искусственного интеллекта.

Обсуждая проблемы этики ИИ, мы принимаем во внимание правоведческие исследования, касающиеся видов и оснований юридической ответственности в сфере цифровых технологий, ИИ и робототехники, порядка принятия решений, а также правосубъектности искусственных интеллектуальных систем [Юридическая… 2023; Синицын 2023]. Следует подчеркнуть, что в юридической науке статус системы ИИ является предметом дискуссий. Так, Т. Я. Хабриева принимает во внимание три ипостаси ИИ с юридической точки зрения: во-первых, «в качестве объекта субъективного права или правового режима (например, объекта исключительных прав, определенного режима эксплуатации)», во-вторых, «как инструмент правового регулирования» и в-третьих, «в качестве субъекта права». При этом отмечается, что возможность наделения искусственного интеллекта статусом субъекта права (например, в рамках концепции электронного лица, аналогичного юридическому лицу) решительно отвергается правоведами, отстаивающими «постулаты классической юриспруденции и традиционные, проверенные столетиями юридические конструкции правосубъектности» [Хабриева 2022: 78].

Этические рекомендации, принимаемые в настоящее время авторитетными международными и национальными организациями, не предполагают, что система ИИ может обладать субъектностью и нести ответственность за собственную работу. В таких рекомендациях речь идет об ответственности людей и организаций – физических и юридических лиц, так называемых «акторов ИИ», в качестве которых выступают ученые, программисты, инженеры, специалисты по работе с данными, коммерческие предприятия, университеты, государственные и частные структуры и т. д. [Recommendation…]. В конце 2021 г. рядом российских компаний (образовавших впоследствии Альянс в сфере искусственного интеллекта) был подписан «Кодекс этики в сфере искусственного интеллекта». В пункте 2.2. (с заголовком «Ответственное отношение») сказано: «Акторы ИИ должны ответственно относиться к вопросам влияния СИИ (систем искусственного интеллекта. – *И. А.*) на общество и граждан на каждом этапе жизненного цикла СИИ, включая неприкосновенность частной жизни, этичное, безопасное и ответственное использование персональных данных, к характеру, степени и раз-

меру ущерба, который может последовать в результате использования технологий и СИИ, а также при выборе и использовании аппаратных средств и программного обеспечения, задействованных на различных жизненных циклах СИИ» [Кодекс…].

В. В. Леушина и В. Э. Карпов считают наиболее удачным примером этического регулирования стандарты, созданные в рамках глобальной инициативы Института инженеров электротехники и электроники по этике автономных и интеллектуальных систем [Chatila, Havens 2019]. В рамках этой инициативы рассматриваются вопросы информационной открытости автономных систем, необъективности алгоритма и многие другие. Определение принципов этически выверенного проектирования (Ethically Aligned Design) ориентировано на обеспечение конфиденциальности, устойчивости, подотчетности, эффективности. Характеризуя базовый стандарт IEEP700, Леушина и Карпов отмечают: «Важной отличительной особенностью IEEE P7000 является то, что сам стандарт не определяет, что этично, а что неэтично. Этика в нем упоминается лишь как принцип поведения, который помогает людям судить о том, что правильно, а что нет. Напротив, определение этих норм документ перекладывает на саму организацию» [Леушина, Карпов 2022: 133]. Следует подчеркнуть, что здесь, как и в вышеупомянутых примерах, этические рекомендации призваны регулировать деятельность людей, что лишь опосредованно сказывается на качествах искусственной интеллектуальной системы: «Факт использования данного стандарта не может гарантировать, что спроектированную и построенную систему можно назвать этичной, поскольку «“этичность” этой системы зависит от приверженности этическим принципам разработчиков и пользователей» [Там же: 133]. Иными словами, соблюдение правил этики людьми повышает вероятность, однако еще не является гарантией того, что результатом станет ИИС, которую можно отнести к числу «заслуживающих доверия».

Выражение «искусственный интеллект, заслуживающий доверия» (более точное, хотя и более длинное, чем «доверительный искусственный интеллект») относится и к работе людей, и к работе систем ИИ. В «Этическом руководстве по искусственному интеллекту, заслуживающему доверия», подготовленном в 2018 г. Высокоуровневой экспертной группой Еврокомиссии, заслуживающий доверия искусственный интеллект характеризуется как обладающий следующими основными свойствами. Во-первых, он должен соответствовать правовым нормам, во-вторых, соответствовать этиче-

ским принципам и ценностям, в-третьих, должен быть надежным в техническом и социальном плане. Авторы «Руководства» справедливо замечают, что «системы ИИ, даже созданные с хорошими намерениями, могут причинять непреднамеренный вред» [Ethics… 2018: 2]. Здесь мы возвращаемся к вопросу о «поведении» ИИС, который не сводится к вопросам профессионального поведения создателей ИИС.

Создатели ИИС не всегда могут понять, как система достигает тех или иных результатов, почему действует таким, а не иным образом. Это дает основания говорить о феномене «фундаментальной замутненности» действий самообучающихся систем [Войскунский 2022]. В марте 2023 г. группа экспертов и промышленников (в числе которых был знаменитый Илон Маск) опубликовала открытое письмо с призывом остановить на полгода разработку систем искусственного интеллекта, превосходящих по мощности GPT-4. В качестве причины указаны потенциальные риски, которые несут такие системы обществу. Авторы выражали обеспокоенность тем, что в последнее время разработчики искусственного интеллекта стали участниками неконтролируемой гонки по изготовлению все более мощных «цифровых умов» (“digital minds”), действия которых непонятны и непредсказуемы даже для их создателей и не могут контролироваться надежным образом [Pause… 2023].

Профессиональная подготовка людей, которым предстоит создавать технику, традиционно предполагала усвоение не только естественно-научных, научно-технических и инженерных знаний, но в неразрывной взаимосвязи с этим – ценностных ориентиров, образцов поведения, одобряемого профессиональным сообществом, а также знакомство с поучительными примерами того, «как не следует делать», с примерами поведения предосудительного. Образование и последующая профессиональная деятельность способствовали усвоению этических норм, существенную часть которых составляют так называемые неписаные правила, и лишь небольшую – явно сформулированные предписания, отчасти совпадающие с юридическими. Этика инженера в этом смысле существует так же долго, как долго существует профессия инженера. В. Г. Горохов справедливо отмечал, что действенность профессиональной этики определяется такими факторами, как наличие профессионального сообщества, развитое самосознание профессионалов, а также деятельность социальных структур, обеспечивающих моральное поведение [Горохов 2013: 59]. Однако в новых областях техники и науки по-

добные факторы определить затруднительно. В качестве показательного примера В. Г. Горохов рассматривал вопрос о «наноэтике» (возникший в связи с развитием нанотехнологий и перспективами нанотехнонауки) и приходил к выводу, что в этой междисциплинарной области нет сложившегося профессионального сообщества, а действие рыночных механизмов ведет скорее к демонтажу ранее созданных социальных институтов, чем к возникновению новых.

Подобные соображения актуальны и в контексте этики искусственного интеллекта. Своды этических правил и рекомендаций, принятые в последние годы международными и национальными организациями, рассчитаны на регулирование деятельности людей разных профессий и родов занятий, имеющих дело с системами ИИ на всех этапах жизненного цикла таких систем. Здесь возникают и сложности взаимодействия этосов разных профессий, и вопросы использования механизмов этической регуляции в целях недобросовестной конкуренции. Например, сообщения о том, что Open AI разрабатывает специальную систему ИИ, которая будет оценивать безопасность других систем, вызвали серьезные опасения, связанные с перспективой монополизации контролирующих функций большими компаниями и навязыванием неоправданных ограничений остальным участникам рынка высоких технологий.

«Если исходить из предпосылок технологического детерминизма, то техническая этика выступает лишь как своеобразное “музыкальное сопровождение” к техническому развитию. Но если техническое развитие детерминировано, тогда вообще не возникает вопросов управления этим развитием в том направлении, которое желательно для общества или является этически справедливым», – писал В. Г. Горохов в статье «Этика в технике» [Горохов 2014: 15]. Не соглашаясь с подобным ограничением задач этики, автор настаивал на необходимости регулирования технологического развития и подчеркивал возрастающую роль этических норм в условиях риска.

Сегодня слова о необходимости этической регуляции искусственного интеллекта можно найти в официальных документах, относящихся к научно-технической и промышленной политике в нашей стране. В тексте «Национальной стратегии развития искусственного интеллекта на период до 2030 г.» в качестве одного из направлений комплексной системы регулирования указана «разработка этических правил взаимодействия человека с искусственным интеллектом» [Указ… 2019: 49 ж]. Мы можем добавить к этому,

что разработка действенных и в должной мере обоснованных этических правил немыслима без развития научных исследований в области этики искусственного интеллекта и применения результатов таких исследований в подготовке решений, значимых для общества и человека.

Литература

Алексеев А. П., Алексеева И. Ю. Судьба интеллекта и миссия разума. М. : Проспект, 2021.

Алексеева И. Ю. Субъектность искусственного интеллекта: старые вопросы в новых контекстах // Информационное общество. 2020. № 6. С. 2–7.

Алексеева И. Ю., Аршинов В. И. Информационное общество и НБИКС-революция. М. : ИФ РАН, 2016.

Алексеева И. Ю., Шклярик Е. Н. Что такое компьютерная этика? // Вопросы философии. 2007. № 9. С. 60–72.

Вейценбаум Дж. Возможности вычислительных машин и человеческий разум. От суждений к вычислениям. М. : Радио и связь, 1982.

Войскунский А. Е. Эвристики человеческие и нечеловеческие // Вестник РУДН. Сер.: Психология и педагогика. 2022. № 2. С. 195–208.

Горохов В. Г. Философия техники и инженерная этика // Ведомости прикладной этики. Вып. 42. Этика инженера: через понимание к воспитанию / под ред. В. И. Бакштановского, В. В. Новоселова. Тюмень : ТюмГНГУ, 2013. С. 41–61.

Горохов В. Г. Этика в технике // Научно-техническое развитие и прикладная этика / отв. ред. В. Г. Горохов, В. М. Розин. М. : ИФ РАН, 2014. С. 11–22.

Горохов В. Г., Декер М. Технологические риски как социальная проблема при разработке и внедрении интеллектуальных автономных роботов // Глобальное будущее 2045. Конвергентные технологии (НБИКС) и трансгуманистическая эволюция / под ред. Д. И. Дубровского. М. : МБА, 2013. С. 82–93.

Грунвальд А. Техническая этика // Научно-техническое развитие и прикладная этика / отв. ред. В. Г. Горохов, В. М. Розин. М. : ИФ РАН, 2014. С. 23–32.

Карпов В. Э., Готовцев П. В., Ройзензон Г. В. К вопросу об этике и системах искусственного интеллекта // Философия и общество. 2018. № 2. С. 84–105.

Кодекс этики в сфере ИИ [Электронный ресурс]. URL: https://ethics.a-ai.ru/ (дата обращения: 12.12.2023).

Леушина В. В., Карпов В. Э. Этика искусственного интеллекта в стандартах и рекомендациях // Философия и общество. 2022. №. 3. С. 124–140.

Майленова Ф. Г. Люди и роботы: сбывающиеся прогнозы. Шаг длиной в столетие // Философия и общество. 2019. № 3. С. 95–105.

Малюк А. А., Полянская О. Ю., Алексеева И. Ю. Этика в сфере информационных технологий. М., 2011.

Минский М. Сообщество разума. М. : АСТ, 2018.

Минский М. Машина эмоций. М. : АСТ, 2020.

Осадчий П. С. К вопросу о принципах профессиональной этики инженеров. СПб. : Тип. А. Бенке, 1911.

Перов В. Ю., Головков В. В. Прикладная этика: «цифровое» переосмысление // Ведомости прикладной этики. 2022. Вып. 60. С. 91–102.

Разин А. В. Этика искусственного интеллекта // Философия и общество. 2019. № 1. С. 57–73.

Ракитов А. И. Философия, роботы, автоматы и зримое будущее // Философия и общество. 2019. № 3. С. 35–48.

Середкина Е. В. Этические аспекты социальной робототехники // Человек. 2020. Т. 31. № 4. С. 109–127.

Синицын С. А. Личные неимущественные права и безопасность человека в виртуальном пространстве // Журнал зарубежного законодательства и сравнительного правоведения. 2023. Т. 19. № 1. С. 13–24.

Указ Президента Российской Федерации от 10.10.2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» [Электронный ресурс]. URL: http://www.kremlin.ru/acts/bank/44731 (дата обращения: 12.12.2023) .

Финн В. К. К структурной когнитологии: феноменология сознания с точки зрения искусственного интеллекта // Вопросы философии. 2009. № 1. С. 88–103.

Хабриева Т. Я. Право, искусственный интеллект, цифровизация // Человек и системы искусственного интеллекта / отв. ред. В. А. Лекторский. СПб. : Юридический центр, 2022. С. 71–97.

Юридическая концепция роботизации / под ред. Ю. А. Тихомирова, С. Б. Нанбы. М. : Проспект, 2023.

Chatila R., Havens J. C. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems // Intelligent Systems, Control and Automation: Science and Engineering. 2019. Vol. 95. Pp. 11–16.

Dubber M. D., Pasquale F., Das S. The Oxford Handbook of Ethics of AI. Oxford : Oxford University Press, 2020.

Ethical Issuesin the Use of Computers / ed. by D. Johnson, J. Snapper. Belmont, CA : Wadsworth Pub, 1985.

Ethics Guidelines for Trustworthy AI. 2018 [Электронный ресурс]. URL: https://www.euractiv.com/wp-content/uploads/sites/2/2018/12/AIHLEG (дата обращения: 12.12.2023).

Ethics of Artificial Intelligence / ed. by S. M. Liao. Oxford : Oxford University Press, 2020.

Floridi F. The Ethics of Artificial Intelligence. Oxford : Oxford University Press, 2022.

Karpov V. E. Can a Robot Be a Moral Agent? [Электронный ресурс] : Artificial Intelligence. RCAI 2020. Lecture Notes in Computer Science. 2020. Vol. 12412. Cham : Springer. DOI: 10.1007/978-3-030-59535-7_5.

Moor J. Are There Decisions Computer Should Never Make? // Nature and System. 1979. No 1. Pp. 217–229.

Pause Giant AI Experiments: An Open Letter. 2023 [Электронный ресурс]. URL: https://www.reuters.com/technology/musk-experts-urge-pause-training-ai-systems-that-can-outperform-gpt-4-2023-03-29/ (дата обращения: 12.12. 2023).

Recommendation on the Ethics of Artificial Intelligence [Электронный ресурс]. URL: https://www.unesco.org/ru/artificial-intelligence/recommendation-ethics (дата обращения: 12.12.2023).

Veruggio G. The Birth of Roboethics [Электронный ресурс] : IEEE International Conference on Robotics and Automation Workshop on Robo-Ethics Barcelona. 2005. April 18. URL: https://www.semanticscholar.org/paper/The-birth-of-roboethics-Veruggio/8fe33312dd2fed75c3d5d4075b70f70c88f4e83c (дата обращения: 12.12.2023).