В. В. ЛЕУШИНА, В. Э. КАРПОВ

ЭТИКА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В СТАНДАРТАХ И РЕКОМЕНДАЦИЯХ

Активное обсуждение вопросов этики ИИ связанно, с одной стороны, с интенсивным развитием и распространением технологий ИИ, а с другой – с повышенным интересом к этому вопросу со стороны широкой общественности (не считая политической, коммерческой и тому подобной заинтересованности). Очевидно, что для дальнейшего развития сфер применения ИИ стали необходимы стандарты и рекомендации, определяющие принципы этичного ИИ. В статье рассматриваются разрабатываемые и опубликованные нормативные документы международных и региональных организаций, в которых приводится схожее абстрактное описание этических принципов, которым должен соответствовать ИИ. Приводится краткое изложение сути стандартов и рекомендаций. Особое внимание уделено стандарту P7000 от IEEE, который не дает строгого определения этичных систем, но, наоборот, перекладывает эту обязанность на самих потребителей и разработчиков, учитывая, таким образом, их ценности и идеологию. Подробно рассматриваются и рекомендация ЮНЕСКО, в которых приводится описание универсальной модели этичного ИИ. Акцентируется внимание на присутствие в некоторых пунктах этого документа конъюнктурных и сомнительных пассажей. Также упоминается отечественный кодекс этичного ИИ, в большинстве своем состоящий из принципов, аналогичных пунктам рекомендации ЮНЕСКО. В работе делается вывод о причинах сходства большинства разработанных нормативных документов. В заключении поднимаются вопросы, касающиеся возможности реализации и соблюдения предлагаемых норм.

Ключевые слова: *этика, искусственный интеллект, рекомендации этики ИИ, Кодекс этики ИИ, этически обусловленное проектирование, стандарты IEEE, ISO, ГОСТ.*

The intensive AI development and, on the other hand, increasing interest among public opinion (excluding political, commercial, etc.) to ethics-related AI issues had led to the active discussions of this topic. Obviously, standards and guidelines that determine ethical AI principles had become necessary for

further development of AI applications. The article discusses regulations, which were published or are being drafted by international or regional organizations. These regulations also provided an abstract description of ethics principles that AI must correspond. The text below is a summary of the essence and structure. The article especially regards the IEEE P7000 standard, which is not provided with a strict definition of ethical systems, so consumers and developers has to define it themselves. Thus, their values and ideology are also considered. Moreover, the UNESCO recommendation describing a universal model for ethical AI is also discussed in detail. The article focuses on the presence of opportunistic and questionable passages at some points in this document. In conclusion, the paper mentions reasons for similarity of most of the developed regulatory documents, as well as raises issues of opportunities for feasibility of the proposed rules and compliance with them.

Keywords: *ethic, artificial intelligence, recommendation on the ethics of AI, The AI Ethics Code; Ethically Aligned Design, IEEE, ISO, GOST standards.*

Введение

Развитие технологий искусственного интеллекта (ИИ), их активное внедрение не только в производство, но и в повседневную жизнь, повлекли за собой необходимость разработки нормативных документов, регулирующих возникающие или потенциально возможные этические проблемы. Диапазон этих проблем достаточно обширен: от нарушения конфиденциальности и безопасности до явных угроз человеку, поэтому сегодня предпринимаются многочисленные попытки их стандартизации. Необходимо понимать, что этические вопросы, касающиеся ИИ, стоят перед всем мировым сообществом, а это означает, что необходимо разработать некую документальную, нормативную основу, которой смогут следовать все страны, чтобы на ее основе стало возможным сформулировать уточняющие стандарты или рекомендации, учитывающие собственные ценности, культурные традиции, моральные нормы различных стран. Сейчас, несмотря на то, что создание объемлющего этического кодекса для ИИ находится в начале пути, многие организации уже начали не только предлагать свое видение развития данного направления, но и разрабатывать стандарты и рекомендации, регулирующие этические вопросы ИИ в разных областях.

Этические вопросы уже давно являются предметом активных массовых обсуждений, при этом из-за некорректных определений

и зачастую недопонимания сути проблем, возникает определенная путаница. Основная проблема заключается в том, что обсуждения такого рода часто происходят среди общественности, которая не имеет достаточной компетенции в данном вопросе. ИИ-системы нередко понимаются очень абстрактно, метафорично, без учета или осознания их специфики. Вследствие этого (и не только) вместо решения конкретного вопроса – этики ИИ – рассуждения чаще всего уходят совсем в другие области. Для обсуждения данной проблематики необходимо точно понимать отличительную особенность этики интеллектуальных/автономных систем (И/АС) от всех остальных этических областей: И/АС – это система, которая самостоятельно принимает критически важные для человека решения. Следовательно, в первую очередь вопросы этики должны касаться вопросов поведения и принятия решений [Карпов и др. 2018]. Иными словами, если И/АС принимает такие решения, то, естественно, эти решения должны оцениваться с точки зрения их моральности.

При этом этические соображения при принятии решений рассматриваются как некий дополнительный фильтр, применяющийся в ситуации с неоднозначными множественным выбором альтернатив. В работе А. В. Разина такая ситуация рассматривается с точки зрения свободы решения («принимаются в условиях неполной информации и являются вероятностными» [Разин 2019]), но в любом случае предполагается, что в И/АС должны быть заложены критерии выбора вместе со встроенными этическими ограничениями.

Непонимание этой специфики И/АС приводит к тому, что мы имеем разнообразные не очень внятные документы, обзор основных из которых приведен ниже.

1. Рекомендации

Основными интересующими рекомендациями, определяющими этичность ИИ, являются документы, представленные ЮНЕСКО и отечественным Альянсом в сфере ИИ. Оба документа описывают этические принципы, которым должен соответствовать ИИ, и призывают придерживаться их всем заинтересованным сторонам.

ЮНЕСКО. В конце 2021 г. генеральная конференция ЮНЕСКО [2021] утвердила рекомендацию об этических аспектах ИИ. В этом документе описываются ключевые характеристики, присущие этич-

ному ИИ. Таким образом ЮНЕСКО определяет универсальную модель ИИ, ценностные установки и принципы деятельности которой должны соблюдаться всеми интересантами. Более того, в документе прописана рекомендация государствам – членам ООН принять надлежащие меры, в том числе и на законодательном уровне, по соблюдению утвержденных этических принципов на их территориях. По мнению ЮНЕСКО, крайне важно выстроить доверие между человеком и системами ИИ, так как тогда можно будет применять данные системы на благо человечества. А добиться вызывающего доверия ИИ возможно лишь тотальным контролем на протяжении всего его жизненного цикла. Документ утверждает следующие ценностные установки и принципы деятельности:

1. Уважение, защита и поощрение прав человека, основных свобод и человеческого достоинства. ИИ должен защищать и повышать качество жизни человека вне зависимости от расы, цвета кожи и других дискриминирующих признаков, и он не должен наносить ущерб или подвергать опасности (социальной, экономической и т. п.) человеческое сообщество.

2. Благополучие окружающей среды и экосистем. ИИ должен соответствовать всем международным и национальным нормам по защите, восстановлению и развитию окружающей среды, так как здоровая экосистема необходима для выживания человечества.

3. Обеспечение разнообразия и инклюзивности. Система не должна ограничивать возможность человеческого выбора.

4. Жизнь в мирных, справедливых и взаимосвязанных обществах. Считается, что каждый человек является частью большего целого, и его процветание зависит от благополучия других людей. ИИ необходимо поощрять мир, инклюзивность, справедливость, равноправие и взаимосвязанность между людьми, не порождать разногласия и не ставить под угрозу сосуществование человека со всем, что его окружает.

5. Соразмерность и непричинение вреда. Деятельность ИИ-систем в процессе достижения законной цели не должна выходить за рамки необходимого. При наличии угрозы нанесения вреда следует оценить все соответствующие риски и принять меры по устранению вероятности появления угрозы. В случаях, когда рассматриваются необратимые или труднообратимые ситуации, окончательное решение должно оставаться за человеком.

6. Безопасность и защищенность. На всех этапах жизненного цикла ИИ необходимо учитывать риски безопасности (причинение вреда) и защищенности (уязвимость к кибератакам).

7. Справедливость и отказ от дискриминации. ИИ-система должна способствовать социальной справедливости, соблюдать принципы непредвзятости и следовать нормам международного права. Разработчики ИИ обязаны минимизировать проявления дискриминации и предвзятости. Кроме того, для обеспечения справедливости на всех этапах жизненного цикла ИИ пристальное внимание необходимо уделять цифровому неравенству между странами и внутри них.

8. Устойчивость. Технология на основе ИИ не должна мешать формированию устойчивых обществ.

9. Право на неприкосновенность частной жизни и защита данных. Для обеспечения защиты человеческого достоинства, личной независимости и способности человека выступать субъектом действия необходимо соблюдать неприкосновенность частной жизни. Вся деятельность ИИ-системы должна осуществляться в рамках международного права с учетом ценностных установок и принципов рекомендации ЮНЕСКО.

10. Подконтрольность и подчиненность человеку. ИИ-система для принятия решений никогда не сможет заменить человека в качестве конечного субъекта ответственности и подотчетности. Поэтому этическая и правовая ответственность всегда в той или иной мере возлагается на конкретные физические или действующие юридические лица. Вопросы жизни и смерти никогда не должны передаваться ИИ-системам. (Например, в Китае, в котором технологии ИИ уже давно применяются в судебной сфере, подобные системы выполняют роль помощника судьи, способного собирать, хранить, находить и выдавать искомую информацию, а также предоставлять заключение, сделанное на основе анализа предыдущих похожих дел. При этом окончательное решение все равно остается за судьей, то есть за человеком [СМИ… 2021].)

11. Прозрачность и объяснимость. Человека необходимо информировать о том, какие решения были приняты при помощи ИИ, а также, при необходимости, понятно их обосновать.

12. Ответственность и подотчетность. Люди, ответственные за ИИ-системы, обязаны уважать, защищать и поощрять права и сво-

боды человека, содействовать охране окружающей среды, а также предусматривать наличие механизмов контроля, позволяющих проводить проверку и аудит нарушений ИИ-систем.

13. Осведомленность и грамотность. Необходимо улучшать понимание гражданами ИИ-технологий. Таким образом, отношение к ИИ станет зависеть от того, как они влияют на сферу прав человека и на окружающую его среду.

14. Многостороннее и адаптивное управление и взаимодействие. Использование данных должно осуществляться на основе международного права и уважения национального суверенитета, а также для реализации инклюзивного управления в сфере ИИ привлечь внимание к участию во всех этапах жизненного цикла ИИ-систем широкого круга заинтересованных лиц.

Дополнительно в рекомендации прописаны стратегические меры, задачей которых является практичная реализация описанных ранее ценностных установок. Все стратегические меры делятся на приоритетные области, в которых их необходимо реализовать. Всего их двенадцать:

1. Оценка этического воздействия.
2. Этичное управление и руководство.
3. Политика в отношении данных.
4. Развитие и международное сотрудничество.
5. Окружающая среда и экосистемы.
6. Гендерное равенство.
7. Культура.
8. Образование и научные исследования.
9. Коммуникация и информация.
10. Экономика и рынок труда.
11. Здоровье и социальное благополучие.
12. Мониторинг и оценка.

В каждом разделе достаточно подробно описано, каким образом должны поступать государства-члены для обеспечения того или иного принципа, однако к некоторым пунктам рекомендации возникают вопросы.

Например, пункт 88 («целевые ассигнования на финансирование программ в поддержку гендерного равенства» и «меры по целевому финансированию программ и использованию гендерно неспецифического языка в целях расширения представленности де-

вушек и женщин в области естественных наук, техники, инженерии и математики») заставляет задуматься о его целесообразности. Во-первых, он больше относится к глобальной проблеме гендерного равенства, нежели к узкой проблеме в рамках ИИ, а во-вторых, продвижение идеи «неспецифического языка» и вовсе заставляет сомневаться в его осмысленности.

Также вызывают недоумение пункты 92 («следует поощрять гендерное разнообразие в сфере связанных с ИИ научных исследований <…> посредством предоставления девушкам и женщинам льготного доступа к данной области деятельности») и 93 («содействовать созданию репозитория передового опыта в области стимулирования участия женщин, девушек и недостаточно представленных групп населения во всех этапах жизненного цикла ИИ-систем»). Пункт 92 не дает равные права женщинам и девушкам, а наоборот, наделяет их преимуществом перед другими в виде льготного доступа. К пункту 93 возникают вопросы по поводу словосочетания «недостаточно представленные группы населения»: этот термин слишком абстрактен, а в документе не приведены уточняющие разъяснения этого понятия.

Выводы. Рекомендации ЮНЕСКО – первый в истории глобальный этический нормативный документ ИИ, который определяет универсальную модель этичного ИИ. В них приводится описание как самих принципов, так и практических методов их реализации. Однако рекомендации полны конъюнктурных и сомнительных пассажей; очевидно, что некоторые пункты рекомендаций подвергнуты излишне сильному влиянию современных европейских веяний.

Национальный кодекс этики. В 2021 г. ряд ведущих российских компаний с большим ИТ-потенциалом утвердил первый национальный кодекс этики [Кодекс... 2021]. Последний устанавливает общие этические принципы и некие стандарты поведения, которым следует руководствоваться всем, кто занимается созданием, внедрением или использованием технологий ИИ. Согласно VIII бюллетеню «Цифровая экономика: международная повестка», Совет Федерации предложил закрепить кодекс в законодательном праве [Цифровая... 2021].

Этот документ по своей структуре схож с рекомендациями ЮНЕСКО: в нем приводятся абстрактные описания этических

норм, которым обязан подчиняться ИИ, а главная задача развития технологий ИИ заключается в защите интересов и прав как отдельного человека, так и людей в целом. Согласно кодексу, разработчики ИИ-систем должны полностью осознавать всевозможные риски и ответственно подходить к вопросам влияния ИИ на общество. Наивысшей ценностью при разработке должны считаться права и свобода человека. Естественно, технологии ИИ на любом этапе создания или эксплуатации должны соответствовать законодательству Российской Федерации (РФ), а их деятельность – не дискриминировать пользователей по расе, национальности, политическим взглядам, религии, возрасту, полу, социальному и экономическому статусу. Недопустимым является также вмешательство в личную жизнь человека. (Разумеется, речь здесь должна идти не о технологиях как таковых, а лишь о продуктах ИИ-технологий.)

В отличие от остальных международных норм, кодекс поощряет внедрение добровольной сертификации разработанных технологий ИИ согласно правовым нормам (кодекса и законодательства РФ). Другой отличительной особенностью рекомендации является учет идентификации ИИ в общении с человеком: обычный пользователь должен быть проинформирован о том, что он взаимодействует с ИИ, когда деятельность касается вопросов прав человека или критических сфер его жизни, а также иметь возможность по желанию прекратить такое взаимодействие.

Церемония подписания кодекса прошла в рамках I форума «Этика искусственного интеллекта: начало доверия» 26 октября 2021 г. Одним из результатов встречи стали рекомендации от участников форума [Рекомендации... 2021]. Это весьма примечательный ритуальный документ. Правда, рекомендация № 5 о создании совета по этике, состоящего из «авторитетных представителей общества – артистов, композиторов, художников, юристов, философов, филологов, психологов, педагогов, социологов, врачей, спортсменов», хорошо иллюстрирует причину сомнений в конструктивности и продуманности подобного рода документов.

Выводы. В целом большинство этических принципов, прописанных в кодексе (соответствие закону, отсутствие дискриминации, непричинение вреда и т. п.), схожи или повторяют принципы, указанные в рекомендации ЮНЕСКО и не противоречат другим опубликованным нормативным документам, затрагивающим сферу эти-

ки ИИ. Стоит отметить, что, по сравнению с ЮНЕСКО, этические принципы национального кодекса более абстрактны и в меньшей степени спекулятивно-конъюнктурны.

2. Стандарты

Основными разработчиками стандартов этики ИИ с полным основанием можно назвать Институт инженеров в области электротехники и электроники (IEEE) и Международную организацию по стандартизации (ISO). Обе организации придерживаются политики добровольного соблюдения их стандартов. Если же говорить об отечественной стандартизации (ГОСТ), то пока ее активность сконцентрирована на регулировании вопросов, связанных с ИИ-продукцией как таковой, хотя уже в краткосрочных планах рассматриваются и вопросы этики ИИ.

IEEE. В 2016 г. в рамках IEEE была начата глобальная инициатива по определению принципов создания этически обусловленных интеллектуальных и автономных систем (И/АС). Основные цели проекта заключались в том, чтобы регламентировать деятельность по исследованию, разработке и применению И/АС с точки зрения соблюдения различного рода этических норм в самых разных аспектах. Это, по мнению участников инициативы, позволит развивать технологии ИИ во благо человечества [Chatila, Havens 2019]. По их мнению, разрабатываемые системы должны не только разрешать технические проблемы, но и учитывать выгоду для человечества. Это позволит обеспечить доверие между человечеством и разрабатываемыми технологиями, что, в свою очередь, приведет к возрастанию благополучия людей. Одним из результатов этой инициативы является документ «Этически обоснованное проектирование» (“Ethically Aligned Design”), коллективная разработка которого продолжается и сегодня [IEEE… 2019]. Практическим воплощением инициативы стала серия стандартов IEEE P7000, в которых рассматриваются конкретные вопросы, возникающие на стыке технологических и этических областей. Серия включает в себя 14 активных проектов, четыре из которых – P7000, P7005, P7007 и P7010 – завершили свою работу публикацией соответствующих стандартов.

IEEEP 7000. Модельный процесс для решения этических проблем в ходе проектирования системы. Предусматривает ряд проце-

дур, позволяющих организациям учитывать этические аспекты на всех этапах изучения и разработки концепции [IEEE… 2021]. Цель стандарта – дать организациям возможность разрабатывать системы с учетом индивидуальных и общественных этических ценностей (таких как конфиденциальность, устойчивость, подотчетность и др.), а также критериев, которые обычно учитываются при разработке (например, эффективность). Основное внимание уделяется тому, как реализовать на практике этические ценности, от которых обычно зависит разработка и внедрение технологий.

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. 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.

IEEE P7010. Измеряет воздействие И/АС на людей. Цель стандарта – руководство по разработке продукта, выявление областей, в которых требуются улучшения, управление рисками и выявление преднамеренных и непреднамеренных пользователей, воздействующих при помощи И/АС на благополучие человека [IEEE… 2020]. На данный момент в рамках этой серии стандартов поднимаются и другие интересные вопросы: информационная открытость автономных систем (IEEE P7001), процесс сохранения интеллектуальных данных (IEEE P7002), необъективность алгоритма (IEEE P7003) и т. п.

Выводы. Разработанный стандарт определяет варианты разрешения некоторых этических вопросов, но не дает строгоо понимания этичности систем, а наоборот, перекладывает эту обязанность на самих эксплуатантов и разработчиков, учитывая, таким образом, их ценности, идеологию и т. п. Пожалуй, IEEE – первая организация, подошедшая к стандартизации вопросов, связанных с этикой ИИ, с таким видением.

ISO. В ISO решением этических вопросов в сфере ИИ занимается технический подкомитет ISO/IEC JTC 1/SC 42 «Искусственный интеллект». Позиция организации заключается в том, что

международные стандарты помогут создать этическую основу для разработки и эксплуатации систем в будущем. На данный момент ведется разработка стандарта ISO/IEC AWI TR 24 368 «Искусственный интеллект. Этические и социальные проблемы. Общие положения» [ISO 2022]. В этом документе приводится информация о принципах, процессах и методах в области этических и общественных проблем, применяемых в деятельности подкомитета.

В нем описаны следующие принципы, которые ISO считает критически важными при оценке этичности ИИ:

- подотчетность: владельцы, разработчики, пользователи и лица, представляющие систему ИИ, ответственны за полное соответствие правовым нормам процесса проектирования, разработки, обучения алгоритмов и эксплуатации системы ИИ;
- ответственность: в системе ИИ полностью реализован заявленный функционал, а сама система используется строго по назначению;
- объяснимость: либо решения ИИ понятны конечному пользователю (желательно) или хотя бы третьей независимой стороне, либо ИИ может самостоятельно доступно объяснить их;
- достоверность: результат выполнения ИИ задач постоянен и не зависит от времени;
- безопасность: необходимо разработать систему, которая будет предотвращать ненадлежащее использование ИИ и уменьшать риск причинения вреда;
- устойчивость: разработанная модель ИИ, соответствующая спросу нынешнего поколения, не должна помешать удовлетворению потребностей будущих поколений;
- конфиденциальность: ИИ должен быть спроектирован и создан согласно регламенту «Privacy by Design» («Конфиденциальность проектирования»);
- защищенность: необходимо разработать систему, которая будет обеспечивать безопасность и уменьшать риск несанкционированного доступа;
- беспристрастность: в своих решениях ИИ не должен принимать во внимание такие факторы, как наделенность властью и демографическая категория;
- равноправие: ИИ должен не только относиться ко всем одинаково с точки зрения прав, выгоды и обязанностей, но и поощрять такое отношение;

• толерантность: ИИ не должен лишать пользователей права участия в принятии затрагивающих их решений из-за расовых, половых, социально-экономических и подобных признаков.

Выводы. Организация ISO еще не представила окончательный вариант своего видения универсальной модели этичного ИИ, однако в их работе над данным вопросом прослеживается определенная схожесть с представлениями ЮНЕСКО и национального кодекса ИИ: на данный момент этические принципы стандарта кардинально не выделяются среди остальных предложенных вариантов. Многие формулировки, безусловно, режут слух и выглядят неуклюже и непривычно, но это – общая тенденция для документов такого рода.

ГОСТ. В России на данный момент ведется активная стандартизация технологий ИИ. Сам ИИ рассматривается как подкласс информационных технологий (ИТ), поэтому, по мнению разработчиков, к нему можно применить общие стандарты и рекомендации в области ИТ, в том числе касающиеся этических вопросов [Перспективная… 2021]. Однако, учитывая процесс стандартизации ИИ, можно утверждать, что разработка нормативных документов, регулирующих этические проблемы, будет происходить схожим образом, а именно: разрабатываемые документы будут «гармонизировать» с международными, в том числе созданными подкомитетом SC42 “Artificial Technologies” Объединенного международного технического комитета Международной организации по стандартизации и Международной электротехнической комиссии ISO/IEC JTC 1 “Information Technologies”.

Выводы. Сейчас еще рано обсуждать отечественное представление о сути этичного ИИ. Однако можно предположить, что, скорее всего, его основой станут стандарты ISO/IEC AWI TR 24 368 «Искусственный интеллект. Этические и социальные проблемы. Общие положения» и ISO/IECTR 24027 «Искусственный интеллект. Предвзятость в системах искусственного интеллекта и процессе принятия решений искусственным интеллектом» [Там же]. Вопрос заключается в том, появятся ли в ГОСТах кардинальные отличия или новые дополнения к представлению об этичности ИИ от ISO.

3. Сравнение подходов

Между рассмотренными стандартами и рекомендациями имеется определенное сходство. Во-первых, каждая организация при разработке этических норм считает благополучие человека (что бы под этим ни понималось) самым важным аспектом. Во-вторых, если рассматривать стандарт ISO, рекомендации ЮНЕСКО и кодекс этики ИИ, требования (этические принципы) к ИИ во многом схожи между собой или повторяют друг друга. Например, во всех трех стандартах присутствуют следующие принципы: уважение и защита свободы и прав человека, непричинение вреда, обеспечение безопасности и защищенности, справедливость, отказ от дискриминации, подконтрольность и подчиненность человеку, прозрачность, объяснимость, подотчетность и т. п., вплоть до принципа благополучия окружающей среды (ЮНЕСКО).

На фоне остальных нормативных документов достаточно сильно выделяется стандарт IEEE P7000. Он не занимается определением этичности И/АС и не определяет универсальные этические принципы, а предлагает конкретные алгоритмы и нормы для внедрения их в И/АС. Стандарт помогает организациям внедрить в свои системы этические нормы, определенные самой организацией в соответствии с ее ценностями.

Возможно, в будущем, когда организации опубликуют более предметные принципы и методы регулирования вопросов этики, станет проще определить различия между их представлениями об этичном ИИ.

Различаются рассмотренные нормативные документы лишь степенью принуждения выполнения этих самых стандартов и возможностью осуществления сертификации ИИ. Идеология IEEE заключается в том, что стандарты носят лишь рекомендательный характер, а их соблюдение не определяет этичность разрабатываемых систем. Соответственно, на основе этого стандарта нельзя производить сертификацию систем. Такой же позиции придерживается и организация ISO. ЮНЕСКО, в свою очередь, не только дает рекомендации относительно определения свойств этичного ИИ, но и призывает закреплять разработанные ими этические принципы на законодательном уровне все заинтересованные государства – члены ООН. Национальный кодекс этики аналогично ЮНЕСКО по-

ощряет добровольное присоединение и дополнительно одобряет введение добровольной сертификации разработанных ИИ-систем.

Вопрос доверия к системам ИИ

Особое место в области этики И/АС занимает вопрос доверия к системам ИИ. Не вдаваясь в рассуждения о том, что такое доверие, насколько метафорично понимание этого термина и к чему вообще применимо это понятие, отметим одну интересную особенность. Вопросы доверия к системам ИИ поднимаются в этических рекомендациях и таких «общих» стандартах, как ISO и ГОСТ. В то же время в «конкретном» стандарте IEEE доверие к И/АС в явном виде не регламентируется, а упоминается через призму требований к надежности, объяснимости и т. п. Например, в стандартах ISO есть документ «Information Technology – Artificial Intelligence – Overview of Trust Worth in Assign Artificial Intelligence» [ISO… 2020]. Имеется аналогичный ГОСТ: «Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения» [ГОСТ… 2021].

Очевидно, что когда в технических регламентах пытаются определить степень доверия, речь идет прежде всего о реализации некоторой процедуры верификации И/АС, то есть проведении неких объективных измерений. Причем речь идет не о доверии к И/АС как к «стандартному» продукту информационных технологий или к некоторой технической системе (погрешность, надежность, наработка на отказ и т. п.). Нас интересует определение степени доверия к И/АС как системе, принимающей решения, основанные на представлениях об их этичности. Это значит, что здесь возникает ряд серьезных и нерешенных проблем. Верификация этичности подразумевает наличие некоторой системы тестов, программ и методик испытаний. А это автоматически означает необходимость наличия оценочного аппарата – формализованных понятий о моральности.

В этом смысле примечательно Постановление президиума РАН «Искусственный интеллект в контексте информационной безопасности» № 165 от 23.11.2021, в котором, в частности, говорится: «Научному совету РАН по методологии искусственного интеллекта и когнитивных исследований до 1 июня 2022 г. разработать <…> методику оценки доверия к “искусственному интеллекту”», а также

предписывается «подготовить обращение в Федеральное агентство по техническому регулированию и метрологии (Росстандарт) и технический комитет № 164 по стандартизации ИИ с предложением по разработке проекта стандарта оценки доверия к системам ИИ» [Постановление… 2021: 4–5].

Заключение

Работы по стандартизации и регуляции процессов создания и применения И/АС с точки зрения этичности ведутся уже не один год, однако движение в этом направлении выглядит не совсем «поступательным» и равномерным. С одной стороны, работы по техническому регулированию типа стандартов IEEE стартовали еще в 2016 г., а глобальный и абстрактный этический стандарт ИИ – рекомендация ЮНЕСКО – появляется лишь в конце 2021 г. С другой стороны, до этого этические принципы уже существовали в локальных нормах отдельных крупных компаний, например у Google и «Сбербанка» [Artificial… 2022; Принципы… 2022].

Почти все организации, связанные с разработкой нормативных документов, сейчас работают над стандартизацией вопросов этики, а некоторые уже успели опубликовать ряд своих работ. Многие придерживаются идеи начать с создания основы стандарта этики, чтобы в будущем страны, учитывая свои ценности и традиции, смогли сформулировать необходимые им стандарты или рекомендации и заняться разработкой и эксплуатацией ИИ. Благодаря этому разработанные принципы урегулирования этических вопросов не слишком сильно отличаются друг от друга, декларируя лишь самые общие благие пожелания.

Причина «беззубости» и неконкретности рекомендаций очевидна. Задача рекомендаций и кодексов – обозначить проблемные места, сделать некоторые отметки на будущее, чтобы далее все это превращалось в некие нормативные акты и технические документы. Проблема заключается в ином. Все эти кодексы ориентированы на производителей, владельцев и пользователей И/АС, никоим образом не принимая во внимание профессиональное сообщество, тех, кто создает базис такого рода систем, – ученых (как технического, так и гуманитарного плана). Стимулируется ли при этом пресловутая безответственность ученого за создаваемые им технологии, или, напротив, ограждаются ли таким образом исследовате-

ли от дополнительных ограничений в своем творчестве, – это отдельный вопрос, выходящий за рамки данной работы.

На основе рассмотренных нормативных документов можно сделать вывод, что начало процесса стандартизации этики ИИ как на мировом, так и на национальном уровне положено: определено примерное понимание универсальной модели этичного ИИ, которого в будущем при разработке и эксплуатации ИИ-систем будет придерживаться большинство стран. При этом возникает множество вопросов о реализации мер, описанных в стандартах и рекомендациях. В том числе о возможности контроля добровольного соблюдения заявленного списка этических принципов; о последствиях для стран и частных организаций при несоблюдении ими стандартов; о том, как противостоять несоблюдению разработанных нормативных документов. Очевидно, что с течением времени, когда разработанные нормы начнут применять на практике, документы будут дорабатываться, однако предугадать характер этих изменений довольно сложно. Крайне важно в этом случае избежать чересчур сильного влияния со стороны современных социальных конъюнктур, спекуляций и заигрываний с толерантностью.

Литература

ГОСТ Р 59276-2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения (Artificial Intelligence Systems. Methods for Ensuring Trust. General). 2021 [Электронный ресурс]. URL: https://docs.cntd.ru/document/1200177291.

Карпов В. Э., Готовцев П. М., Ройзензон Г. В. К вопросу об этике и системах искусственного интеллекта // Философия и общество. 2018. Т. 87. № 2. С. 84–105.

Кодекс этики в сфере искусственного интеллекта. 2021 [Электронный ресурс]. URL: https://a-ai.ru/code-of-ethics/.

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

Перспективная программа стандартизации по приоритетному направлению «Искусственный интеллект» на период 2021–2024. [Электронный ресурс] : ТК-164. 2021. URL: https://www.tc164.ru/.

Постановление президиума РАН «Искусственный интеллект в контексте информационной безопасности» № 165 от 23.11.2021.

Принципы этики искусственного интеллекта Сбера [Электронный ресурс] : ПАО Сбербанк. 2022. URL: https://www.sberbank.com/ru/about/ethics (дата обращения: 19.01.2022).

Рекомендации I Форума «Этика искусственного интеллекта: начало доверия». 2021 [Электронный ресурс]. URL: https://www.aiethic.ru (дата обращения: 19.01.2022).

СМИ: в Китае искусственный интеллект взял на себя функции прокурора [Электронный ресурс] : ТАСС. 2021. 27 декабря. URL: https://tass.ru/ekonomika/13306521 (дата обращения: 19.01.2022).

Цифровая экономика: международная повестка // AI Cent. MGIMO. 2021. № 8. С. 1–17.

ЮНЕСКО. Рекомендации об этических аспектах искусственного интеллекта // 2021. С. 1–39.

Artificial Intelligence at Google: Our Principles [Электронный ресурс] : Google. 2022. URL: https://ai.google/responsibilities/ (дата обращения: 25.01.2022).

Chatila R., Havens J. C. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems // Intell. Syst. Control Autom. Sci. Eng. 2019. Т. 95. Pp. 11–16.

IEEE. Ethically Aligned Design: First Edition. 2019 [Электронный ресурс]. URL: https://standards.ieee.org/industry-connections/ec/ead-v1/ (дата обращения: 19.01.2022).

IEEE 7000. IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being // IEEE Std 7010-2020. 2020. Pp. 1–96.

IEEE 7000. IEEE Approved Draft Model Process for Addressing Ethical Concerns During System Design // IEEE Std 7000-2021. 2021. Pp. 1–82.

ISO. ISO/IEC DTR 24368 Information Technology – Artificial Intelligence – Overview: Aspects of Ethics and Societal Concerns. 2022.

ISO. ISO/IEC TR 24028:2020. Information Technology – Artificial Intelligence – Overview of Trustworthiness in Artificial Intelligence. 2020 [Электронный ресурс]. URL: https://www.iso.org/standard/77608.html.