DOI 10.37882/2500-3682.2024.05.18

ВОПРОСЫ ЭТИКИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ARTIFICIAL INTELLIGENCE ETHICS ISSUES

D. Pshychenko

Summary: An active discussion of the ethics of artificial intelligence is associated with the intensive development and dissemination of artificial intelligence technologies and with increased public interest in this issue. To further develop the areas of application of artificial intelligence, standards and recommendations that define the principles of ethical artificial intelligence have become necessary. This paper examines draft IEEE standards for the ethics of artificial intelligence. The main problems of formalizing the concept of ethics in artificial intelligence are analyzed. The main types of risks associated with the introduction of modern technologies using artificial intelligence into everyday life are considered. A critical analysis of various mathematical tools has been carried out to formalize the concept of ethics in artificial intelligence. Mechanisms for successfully solving the problem are proposed. Examples of solving practical problems are given.

Keywords: artificial intelligence, IEEE standards, verbal decision analysis, ethics.

Пшиченко Дмитрий Викторович
Доцент, Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет Высшая Школа Экономики», (г. Москва)
dpshychenko@gmail.com

Аннотация: Активное обсуждение вопросов этики искусственного интеллекта связанно с интенсивным развитием и распространением технологий искусственного интеллекта и с повышенным интересом к этому вопросу со стороны общественности. Для дальнейшего развития сфер применения искусственного интеллекта стали необходимы стандарты и рекомендации, определяющие принципы этичного искусственного интеллекта. В данной работе рассмотрены проекты стандартов IEEE этики искусственного интеллекта. Проанализированы основные проблемы формализации понятия этики в искусственном интеллекте. Рассмотрены основные типы рисков, связанные с внедрением в повседневную жизнь современных технологий, использующих искусственный интеллект. Проведен критический анализ различного математического инструментария, позволяющего формализовать понятие этики в искусственном интеллекте. Предложены механизмы для успешного решения поставленной задачи. Приведены примеры решения практических задач.

Ключевые слова: искусственный интеллект, стандарты IEEE, вербальный анализ решений, этика.

В 2016 году Институт инженеров электротехники и электроники (IEEE) в сотрудничестве с другими организациями стал инициатором глобального движения в области этики искусственного интеллекта [1, 2, 3]. Авторитет IEEE, как ведущей организации в сфере электротехники, электроники и информационных технологий, оказывает непосредственное влияние на разработчиков искусственного интеллекта. Результатом инициативы IEEE стал документ «Ethically Aligned Design» («Этически обусловленное проектирование») [3]. В документе IEEE отражены основные актуальные угрозы и риски [4], связанные с внедрением автономных систем на базе искусственного интеллекта.

Измерение риска, согласно исследованиям [4, 5], представляет собой определение опасности, исходящей от определенного источника, для индивида или группы. Существует несколько широко распространенных подходов к измерению риска. Первый подход - инженерный. В этом подходе основное внимание уделяется сбору статистических данных об отказах, авариях и выбросах вредных веществ в окружающую среду [6]. Второй подход, называемый модельным, предполагает создание моделей процессов, способных вызвать различные нежелательные последствия, такие как аварии [6]. Третий подход к измерению риска, известный как экспертный, применяется, когда возникают определенные сложности при использовании инженерного и модельного подходов [4, 5, 7]. Четвертый подход, социологический, основан на измерении восприятия риска населением и его отдельными группами [5, 8, 9].

Рассмотрим несколько небольших примеров, связанных с оценкой риска при внедрении в повседневную жизнь различных технологий искусственного интеллекта. По причине перераспределения рабочей силы на постсоветском пространстве после распада СССР в Москве сосредоточено несколько сотен тысяч водителей автотранспорта (таксисты, водители маршруток, троллейбусов, автобусов и т.п.). Соответственно, если использование беспилотных автомобилей будет набирать современные темпы, то уже через пару лет значительная часть водителей – выходцев из стран ближнего зарубежья, останется без работы. Большинство этих людей другими профессиями быстро овладеть не смогут. Можно предположить, что для такого мегаполиса как Москва (как, впрочем, и для любого другого) такое развитие ситуации может легко привести к социальному взрыву, если предварительно не проделать глубокий анализ всех возможных последствий от внедрения тех или иных технологий искусственного интеллекта.

В июле 2017 года была образована Российская рабочая группа IEEE по вопросам этики искусственного ин-

теллекта в составе: Готовцев П.М. (руководитель), Карпов В.Э., Овсянникова Е.Е. (секретарь) и Ройзензон Г.В. [10]. Главные цели группы – транслировать предложения и мнения российских ученых, которые участвуют в работах над документами IEEE: Ethically Alligned Design, информировать российских ученых о результатах деятельности рабочей группы IEEE по созданию документа IEEE: Ethically Alligned Design, привлекать российских ученых к исследованиям тематики «Этики систем искусственного интеллекта».

В рамках инициативы IEEE предполагается разработка серии стандартов, применение которых, скорее всего, станет обязательным для всех специалистов и организаций, занятых в создании различных продуктов, в той или иной степени использующих технологии искусственного интеллекта. Кроме того, со своей стороны, организации и специалисты, использующие технологии искусственного интеллекта, должны в самое ближайшее время сформулировать и внести свои предложения по выработке дополнительных условий соответствия упомянутым стандартам. Таким образом, в процедуре выработки стандартов должны принять участие несколько сторон (например, IEEE, разработчики и научное сообщество).

Основная цель данной работы — обзор существующего математического инструментария, который может быть использован для формализации понятия этики в искусственном интеллекте. В частности, особый интерес представляют подходы, позволяющие оценивать на соответствие определенным требованиям (этическим нормам, критериям, стандартам и т.п.) те или иные технологии, использующие искусственный интеллект.

Прежде чем рассматривать конкретные математические подходы для решения поставленной задачи важно изучить основные определения для данных терминов.

Этика – философская дисциплина, которая исследует вопросы нравственности и морали.

Относительно определения научного направления искусственного интеллекта дело обстоит несколько сложнее. Нужно признать, что какого-то одного устоявшегося и единодушно принятого научным сообществом определения к настоящему моменту не выработано. Разработано огромное количество различных определений искусственного интеллекта. Под искусственным интеллектом понимается группа методов и подходов, которые ориентированы на решение слабоструктурированных задач.

Нужно упомянуть классические фундаментальные работы по этике как отечественных, так и зарубежных специалистов, которые оказали существенное влияние на современные представления. В частности, можно отметить работы: Апресяна Р.Г. [11]; Гусейнова А.А. [12]; Дробницкого О.Г. [13]; Кропоткина П.А. [14]; Фролова И.Т. и Юдина Б.Г. [15]; Швейцера А.[16]; Шпемана Р. [17] и ряда других.

В истории науки можно привести несколько ярких примеров того, как развитие новых технологий и вопросы этики приводили к весьма существенным противоречиям и столкновениям мнений политиков, общественных организаций, ведущих мировых ученых и т.п. При этом, важно обратить внимание на то, что в вопросах этики в рамках НТП (научно-технического прогресса) инициатива исходила зачастую именно от самих ученых. В 1955 году Б. Рассел озвучил одну из первых таких инициатив (знаменитый манифест Рассела-Эйнштейна). Инициатива Б. Рассела положила начало широко известному теперь Пагуошскому движению за мир и разоружение [15], поскольку к середине 1950-х годов угроза всеобщей ядерной катастрофы стала очевидной и появилась необходимость мобилизовать авторитетнейших ученых (Ф. Жолио-Кюри, Л. Полинг, А. Эйнштейн и др.) для преодоления сложившейся критической ситуации. Еще одним ярким примером может служить инициатива американского генетика П. Берга, который в 1974 году предложил наложить мораторий на эксперименты с рекомбинантной ДНК для того, чтобы иметь возможность оценить все риски и последствия использования этой новой технологии [15].

Основные выводы, которые можно сделать из подобных инициатив ведущих мировых ученых, заключаются в том, что развитие и неразумное использование тех или иных технологий может угрожать существованию человечества. Повсеместное внедрение различных технологий искусственного интеллекта также сопряжено с определенными опасностями и рисками [4], что во многом и стало причиной инициативы по разработке проектов специальных стандартов в области этики искусственного интеллекта со стороны IEEE (см. Табл. 1).

Еще один важный вывод, который можно сделать, заключается в том, что, с одной стороны, развитие тех или иных опасных технологий без учета вопросов этики, может привести человечество к совершенно катастрофическому результату. С другой стороны, если «загнать» развитие современных технологий в слишком жесткие рамки это станет причиной замедления темпов научно-технического прогресса [15].

Кроме того, есть еще один важный аспект, связанный с инициативой IEEE по разработке стандартов. К настоящему моменту на рынке появилось огромное количество продуктов, которые претендуют на то, что в них в той или иной степени используются технологии искусственного интеллекта. В действительности при тщательном изучении оказывается, что представленные продукты во-

обще не имеют никакого отношения к рассматриваемой предметной области. Таким образом, предстоящая возможная достаточно жесткая сертификация продуктов, использующих технологии искусственного интеллекта, заставит многих недобросовестных производителей задуматься, прежде чем осуществлять различные маркетинговые действия. Соответственно разработка рассматриваемых стандартов может послужить определенным фильтром, так как многие производители продукции много раз подумают, стоит ли им заявлять, что их продукция использует технологии искусственного интеллекта, зная, что им предстоит обязательная непростая процедура сертификации.

Таблица 1.

Проекты стандартов IEEE.

Код проекта стандарта IEEE	Перевод
P7000	Проект стандарта модельного процесса для решения этических проблем при проектировании систем
P7001	Проект стандарта прозрачности автономных систем (АС)
P7002	Проект стандарта обеспечения конфиденциальности данных
P7003	Проект стандарта учета необъективности алгоритма
P7004	Проект стандарта управления данными детей и студентов
P7005	Проект стандарта для прозрачного управления данными работодателя
P7006	Проект стандарта для интеллектуального агента управления персональными данными
P7007	Проект онтологического стандарта робототехники и систем автоматизации, управляемых на основе этики
P7008	Проект стандарта для учета этических принципов в робототехнике, интеллектуальных и автоматизированных системах
P7009	Проект стандарта проектирования безотказных автономных и полуавтономных систем
P7010	Проект стандарта метрик благосостояния для систем искусственного интеллекта и автономных систем, действующих на основе этических принципов
P7011	Проект стандарта определения и оценки надежности новостных источников
P7012	Проект стандарта для машиночитаемых правил конфиденциальности личных данных
P7013	Проект стандарта включения и применения технологии автоматизированного анализа состояния лица

Анализ представленных классических работ по этике, а также некоторые выводы авторов работы [1] позволяют констатировать, что вопросы соответствия этики и искусственного интеллекта отличаются от того, что подразумевается под проблемами этики генных технологий и т.п. Это отличие обусловливается тем, что в искусственном интеллекте вопросы этики ближе к пониманию этики в философском смысле, и сопряжены эти этические аспекты с тем, что они относятся к вопросам поведения и принятия решений. Соответственно данный аспект существенно влияет на выбор математического инструментария для формализации понятия этики в искусственном интеллекте.

1. Краткое описание проектов стандартов IEEE этики ИИ

Дадим более подробную характеристику целей проекта этических стандартов искусственного интеллекта, представленных в табл. 1.

P7000. Инженерам, технологам и другим участникам проекта нужен метод выявления, анализа и согласования этических проблем конечного пользователя на ранних этапах жизненного цикла системы и программного обеспечения. Целью настоящего стандарта является обеспечение практического применения этого типа метода проектирования, основанного на ценностях, который демонстрирует, что концептуальный анализ ценностей и обширный технико-экономический анализ могут помочь улучшить этические стандарты в системах и жизненных циклах программного обеспечения. Этот стандарт предоставляет важный инструмент для руководства процессами управления инновациями, подходами к разработке информационных систем и практикой разработки программного обеспечения для снижения этического риска для организаций, заинтересованных сторон и конечных пользователей.

P7001. Одной из важнейших проблем автономных систем (АС) является необходимость обеспечения их работы таким образом, чтобы она была понятна и доступна широкому кругу заинтересованных лиц по нескольким причинам. Прозрачность играет ключевую роль для пользователей, так как она способствует формированию доверия к системе, предоставляя простой способ понимания того, что и зачем система делает. Например, в случае робота, ухаживающего за пожилыми или больными людьми, прозрачность означает, что пользователь может быстро понять, какие функции выполняет робот в различных ситуациях или, если робот принимает неожиданное решение, пользователь имеет право спросить у робота «почему это было сделано именно так?». Важность валидации и сертификации прозрачности автономных систем проявляется в том, что это позволяет раскрывать внутренние процессы системы для проверки. Кроме того, в случае происшествий или аварий АС должна обеспечить прозрачность в ходе расследования, чтобы легко выявить внутренние процессы, приведшие к инциденту. Для адвокатов или других экспертов-свидетелей прозрачность становится важной после несчастных случаев, так как они могут требовать предоставления доказательств. Наконец, для передовых технологий,

таких как беспилотные автомобили, важно обеспечить определенный уровень прозрачности для широкой общественности с целью повышения доверия к этим технологиям. Стандарты, разработанные для дизайнеров, предоставляют руководство по оценке прозрачности при разработке и предоставляют инструменты для улучшения прозрачности, например, обязательство обеспечения защиты данных датчиков и информации о внутреннем состоянии, аналогично «черному ящику» или регистратору данных полета.

P7002. Данный стандарт направлен на создание общего методологического подхода, который устанавливает методы управления вопросами конфиденциальности в жизненном цикле систем и программ.

P7003. Данный стандарт разработан для того, чтобы обеспечить отдельным личностям или компаниям, занимающимся созданием алгоритмов, в основном ориентированных на автономные или искусственные интеллектуальные системы, четкую методологию сертификации, ясность отчетности относительно целей работы алгоритмов, процедуры оценки и их воздействия на пользователей и заинтересованные стороны. Прохождение сертификации согласно данному стандарту поможет разработчикам алгоритмов информировать пользователей и регулирующие органы о том, что при разработке, тестировании и оценке алгоритма использовались передовые современные методы, что способствует избежанию нежелательных дифференцированных последствий для пользователей.

P7004. Данный стандарт разработан с целью предоставить организациям, работающим с информацией о детях и учащихся, процессы управления и сертификацию, обеспечивающие прозрачность и ответственность за их действия в области безопасности и благополучия детей, их родителей, учебных заведений, в которых дети обучаются, а также общественных сообществ, где они проводят время, как офлайн, так и онлайн. Стандарт также создан с целью помочь родителям и учителям осознать, что многие люди могут не обладать достаточными техническими знаниями, чтобы понимать основные вопросы в области обработки данных, но при этом должны быть правильно проинформированы о безопасности информации о своих детях и иметь доступ к инструментам и услугам, предоставляющим адекватные возможности выбирать, основываясь на контенте и полученной ранее информации.

P7005. Данный стандарт разработан с целью обеспечить организациям четкие рекомендации и сертификаты, подтверждающие, что они обеспечивают сохранность, защиту и этичное использование информации о сотрудниках. Его задача также заключается в помощи работодателям в осознании того, что несмотря на возможное отсутствие у сотрудников технических навыков для полного понимания вопросов обработки данных, они должны быть должным образом осведомлены о безопасности персональных данных своих сотрудников, а также иметь доступ к инструментам и сервисам, обеспечивающим должные возможности выбора информации, которую они обменивают на своем рабочем месте, на основе контекста и предварительно полученной информации. Данный стандарт, разработанный в соответствии с Законодательством ЕС о защите персональных данных GDPR (Общее Регулирование по Защите Данных), будет выстроен как своеобразная «GDPR для работников», обеспечивая, что работники, сталкивающиеся с широко распространенными проблемами автоматизации, связанными с потенциальной утратой рабочих мест, будут иметь контроль и управление своей персональной информацией, являющейся важным активом их личности и жизни, независимо от того, была ли она получена в результате мониторинга рабочих процессов или из хранилищ персональных данных.

P7006. С развитием и использованием искусственного интеллекта возникает риск принятия решений на основе данных, которые могут быть непрозрачными для людей, что делает их аналогией «черного ящика». Для обеспечения этических принципов в создании и использовании искусственного интеллекта необходимо разработать средства управления значениями, правилами и данными, которые будут направлять развитие персонализированных алгоритмов. Такие средства должны представлять интересы индивидуумов в контексте общественных норм, этики и прав человека, а также предвидеть этические последствия обработки данных и помогать уменьшать их. Этот подход позволит людям безопасно делиться своей личной информацией на машинном уровне и использовать персонализированный искусственный интеллект в качестве своего представителя при принятии решений машинами. Важной целью создания такого стандарта является информирование государственных и коммерческих организаций о том, что более предпочтительным является создание механизмов, направленных на обучение персональных агентов искусственного интеллекта для преодоления асимметрии и гармонизации использования персональных данных в будущем.

P7007. Данный стандарт определяет набор терминов и их взаимосвязей, который способствует развитию робототехники и систем автоматизации в соответствии с мировыми теориями этики и морали. Его целью является приведение инженерных сообществ к осознанию методов разработки таких систем и их гармоничного внедрения. Эти термины обеспечивают точное взаимопонимание экспертов из различных областей, включая робототехнику, автоматизацию и этику. Использование онтологий для представления знаний в данном контек-

сте имеет несколько преимуществ, таких как: а) формальное определение понятий в определенной области, независимо от языка программирования, с возможностью реализации на конкретном языке программирования; б) инструменты для анализа понятий и их взаимосвязей для выявления несогласованностей, неполноты и избыточности; в) язык для коммуникации между роботами разных производителей и т.д.

P7008. Стандарт предназначен для этически направленных роботизированных, интеллектуальных и автономных систем. Он определяет набор терминов, функций и их взаимосвязей, учитывая преимущества, зависящие от культурных особенностей пользователей, таких как благосостояние и здоровье. Целью данного стандарта является развитие робототехники, интеллектуальных и автономных систем в соответствии с мировыми теориями этики и морали. Он направлен на то, чтобы помочь инженерным сообществам понять, как разрабатывать такие системы рационально и гармонично их внедрять. Вместе с определениями, данный стандарт обеспечивает точное взаимопонимание мировых экспертов в различных областях, включая робототехнику, автоматизацию и этику.

P7009. Данный стандарт устанавливает технические основы для разработки, внедрения и использования эффективных и надежных механизмов в автономных и полуавтономных системах. В его состав входят методологии и инструменты, обеспечивающие практическую реализацию этих целей. Стандарт предусматривает четкие процедуры для измерения, тестирования и сертификации работоспособности системы при различных нагрузках, а также указания по ее усовершенствованию в случае недостаточной эффективности. Он является основой как для разработчиков, так и для пользователей и регулирующих органов, чтобы обеспечить создание надежных, прозрачных и ответственных механизмов, устойчивых к отказам.

P7010. Стандарт метрик благополучия для этического искусственного интеллекта и автономных систем предоставляет программистам, инженерам и технологам возможность более эффективного анализа того, как их продукты и услуги могут способствовать улучшению благосостояния людей. Он основан на широком наборе показателей, включая не только экономический рост и производительность, но и другие аспекты. Сегодня оценка систем с датчиками распознавания эмоций часто происходит преимущественно с точки зрения их экономической ценности на рынке, не учитывая их потенциальные влияния на области, такие как психология и другие. В то время как некоторые могут опасаться, что этические аспекты могут затруднить инновации из-за возможного введения нежелательного регулирования, отсутствие метрик для оценки психического и эмоционального благополучия, как на уровне индивидуума, так и общества, делает невозможной количественную оценку инноваций. Внедрение и использование таких метрик для программистов и технологов позволяет учитывать благосостояние человека как дополнительный фактор, помимо экономического роста, что может способствовать его улучшению.

P7011. Целью этого стандарта является предотвращение негативных последствий распространения ложных новостей путем создания открытой системы рейтингов, которая легко понимается. Он также направлен на восстановление доверия к определенным новостным поставщикам и правильную дискредитацию других, что помогает сдерживать распространение ложных новостей и поощрять улучшения у поставщиков новостей. Этот стандарт нацелен на создание репрезентативного набора новостных сообщений, который будет использоваться для формирования объективной и точной шкалы рейтингов.

P7012. Целью данного стандарта является предоставление людям инструментов для определения собственных условий в отношении личной конфиденциальности таким образом, чтобы их можно было прочитать, признать и подтвердить машинами, управляемыми в цифровом мире. Формально цель стандарта состоит в том, чтобы дать индивидуумам возможность быть главными в соглашениях с другими сторонами, преимущественно с компаниями, выступающими в роли второй стороны. Важно отметить, что цель данного стандарта не сводится к рассмотрению политик конфиденциальности, так как они предполагают наличие только одной стороны и не требуют соглашения. Соблюдение условий требует наличия соглашения, в то время как соблюдение политики конфиденциальности не требует его.

P7013. Исследования продолжают подтверждать, что искусственный интеллект, применяемый для автоматизированного анализа лиц, может быть предвзятым, что может привести к усилению человеческих предрассудков и систематической дискриминации людей по признакам пола, этнической принадлежности, возраста и других факторов. Цель данного стандарта заключается в предоставлении руководящих принципов для разработки и сравнительной оценки автоматизированных технологий анализа лиц с целью смягчения демографической и фенотипической предвзятости и дискриминации. Разделы и протоколы отчетности, установленные в данном стандарте, направлены на повышение прозрачности в использовании такой автоматизированной технологии, чтобы разработчики и лица, принимающие решения, могли сравнивать доступные варианты и выбирать наиболее подходящую технологию на основе целевых групп населения и ожидаемых сценариев использования. Учитывая важность биометрических данных, получаемых

из анализа лица, стандарт также обсуждает правильное и неправильное использование автоматизированного анализа лица на основе общепринятых ценностей, установленных мировым сообществом.

2. Формальные методы

Попытка формализовать этические стандарты решает две важные задачи. Первым шагом является создание моделей представления норм, а вторым — определение правильных математических методов реализации этих моделей: анализ, измерение, сравнение и другие. Уровень сложности и практичности используемых ими технологий обычно находится в области нечеткой, многозначной или вероятностной логики, которые являются высокоразвитыми областями. Качество представления параметров систем искусственного интеллекта и этических стандартов имеет первостепенное значение, которое должно определяться на качественной основе. Необходимо подчеркнуть, что задача формализации этических норм тесно переплетается с более широкой задачей - формализацией гуманитарного знания [9, 18, 19, 20].

Кроме того, необходимо разработать новые установки, а именно сингулярность (вопрос регулирования над «умными» системами), гуманность (влияние машин на наше поведение), безопасность [21] и т. п. Поэтому соответствие этим нормам не всегда может быть сведено к простым «да» или «нет». В этой связи актуальным становится использование неклассических логик, таких как темпоральные и многозначные логики [22], механизмы многокритериальной систематизации, нечеткая и вероятностная логики, теория решеток [19] и т.п. Давайте подробнее рассмотрим некоторые из этих подходов.

2.1. Булева алгебра

В последнее время наблюдается бурное развитие концепции формализации различных этических аспектов. Книга В.А. Лефевра «Алгебра совести» заслуживает особого упоминания как пионерская работа в данном направлении [18]. В ней содержится глава, посвященная этике и потенциальным критериям формализации этого направления. В данном контексте широко используется математический аппарат булевой алгебры. Стоит отметить как положительные, так и отрицательные стороны его использования. К положительным аспектам относится развитая структура булевой алгебры, наличие большого количества приложений и программных библиотек для различных инструментальных средств. Однако, следует отметить, что проблемы этики, включая те, что связаны с искусственным интеллектом, не всегда могут быть четко разделены на «белые» и «черные» согласно булевой логике [23]. В работе Д.А. Поспелова [23] предложено понятие «кольцевых» шкал как способ преодоления этой проблемы. Этот подход представляется оригинальным и перспективным для решения задачи формализации этики в области искусственного интеллекта.

2.2. Многозначные логики

В рамках развития различных нетрадиционных подходов в исследованиях искусственного интеллекта, включая методы многозначных логик, значимыми являются работы отечественных ученых, таких как А. С. Карпенко [24], В.К. Финна [9], О.П. Кузнецова [25], В.Б. Тарасова [26], В.Н. Вагина [27] и др. Применение многозначных логик для формализации этики в области искусственного интеллекта также встречает некоторые трудности. Например, переход к четырехзначной логике от трехзначной может потребовать значительного изменения математических конструкций, что, в сущности, предполагает переосмысление задачи на новом уровне.

2.3. Теория вероятностей и нечеткая логика

Использование нечеткой логики и вероятностного аппарата [28] для формализации понятий этики в сфере искусственного интеллекта представляет собой интересный подход [29]. Нечеткую логику позволено рассматривать как обобщение многозначной логики [26]. Однако есть некоторая характерная специфика применения нечеткой логики, включая проблемы при создании функций принадлежности: различные методы построения таких функций могут привести к различным результатам, что подчеркивает нестабильность методов нечеткой логики относительно начальных данных.

2.4. Вербальный анализ решений

Еще одним потенциальным методом для описания этики искусственного интеллекта является применение методов вербального анализа решений (ВАР) [30].

Методы вербального анализа решений (ВАР) опираются на достижения различных научных областей: когнитивной психологии (включая операции измерения и получения информации при формировании правил принятия решений), прикладной математики (для обоснования выбора правил принятия решений и методов проверки информации на согласованность), теории организаций (для предоставления объяснений), а также компьютерных наук (для разработки человеко-компьютерного взаимодействия). Созданные в рамках этого подхода методы принятия решений дают возможность анализировать варианты сложных решений, гармонично сочетая качественную и количественную информацию о альтернативах, экспертные оценки и предпочтения принимающих решения, а также объективные и субъективные аспекты, характерные для конкретной проблемной ситуации.

Например, в контексте задачи формализации этики в искусственном интеллекте с использованием методов BAP, можно поставить следующую задачу. Предположим, что разрабатываются нормы для этики искусственного интеллекта. Тогда есть возможность рассматривать оценку соответствия каждой нормы как задачу многокритериальной порядковой классификации [7, 30]. Следовательно, посредством анализа этических норм для искусственного интеллекта необходимо принять решение о том, соблюдаются ли нормы, есть ли незначительные нарушения, или прослеживается существенное отклонение от установленных норм и т. д. Следовательно, требуется классифицировать набор оценок по каждой норме в соответствующую категорию решений.

Позитивные аспекты использования методов BAP включают в себя отсутствие необходимости преобразовывать исходные данные в числовую форму. Известно, что такие преобразования вербальных оценок в числовые могут быть субъективными и лишены строгого математического обоснования. К тому же, методы BAP обеспечивают возможность объяснения принятых решений с точки зрения предметной области, в данном случае – описания норм этики в искусственном интеллекте. Однако среди негативных моментов применения методов BAP стоит отметить значительные затраты времени лица, принимающего решения, при работе с многомерным пространством признаков. В таких случаях требуется применение различных методов для сокращения размерности пространства признаков [31, 32].

Следовательно, на данный момент существует множество инструментальных ресурсов, которые основаны на разнообразных математических методах и позволяют эффективно решать задачу формализации этических норм в области искусственного интеллекта.

Пример оценки надежности новостных источников и проект стандарта P7011

В течение последнего десятилетия наблюдается рост масштабов информационных войн, достигших рекордных уровней. Проект стандарта IEEE по этике искусственного интеллекта P7011, который затрагивает оценку надежности новостных источников, направлен на борьбу с негативными последствиями неуправляемого распространения поддельных новостей, известных как «фейковые» новости, путем создания открытой системы оценок. Учитывая, что «фейковые» новости стали одним из ключевых инструментов информационных войн, любые усилия по восстановлению здорового информационного пространства представляют значительный интерес. Если в самое ближайшее время не предпринять соответствующих усилий по предотвращению распространения «фейковых» новостей, это может привести к практически полной потере доверия конечных потребителей новостного контента к масс-медиа (телевидение, радио, интернет-ресурсы и т.п.). Потеря аудитории, в свою очередь, приведет к оттоку рекламодателей от производителей медиаконтента. А результатом этого будет фактическое разорение и банкротство большинства медиа-холдингов. Поэтому во внедрении стандарта этики искусственного интеллекта P7011 в первую очередь заинтересованы сами производители медиаконтента. Для решения данной задачи можно применять методы лингвистической семантики и анализа текстов (с использованием семантических технологий веба). Тем не менее, с учетом современных условий распространения «фейковых» новостей необходимо также анализировать другие компоненты, такие как видео, фотографии, звук и другие, поскольку они также могут быть подвержены фальсификации. Поэтому разработка специальной системы критериев для оценки достоверности новостных источников может также рассматриваться как часть системы этических норм в рамках стандарта этики искусственного интеллекта P7011.Можно предложить следующую модельную систему критериев: 0. «Кричащий» заголовок (оценки: 0. Заголовок новости является «Кричащим»; 1. Нормальный заголовок); 1. Характеристика источника (оценки: 0. Подозрительный источник (много рекламы, странный дизайн, URL и т.п.); 1. Нормальный источник); 2. Неверная дата публикации (оценки: 0. Дата публикации очевидно не соответствует появлению новостного сообщения; 1. Правильная дата); 3. Подложные фотографии (оценки: 0. Очевидно поддельные фотографии; 1. Достоверность фотографий вызывает определенные сомнения; 2. Нормальные фотографии); 4. Обилие ошибок (оценки: 0. Много орфографических и синтаксических ошибок; 1. Есть некоторое количество орфографических и синтаксических ошибок; 2. Ошибок нет); 5. Давление на жалость (оценки: 0. Присутствует давление на жалость; 1. Давления на жалость нет). Классы решений: A. Определенно фейковая новость. B. Есть подозрение, что новость является фейковой. C. Новость не является фейковой.

Рассмотрим более подробно применение метода BAP ЦИКЛ [7, 30], предназначенного для построения многокритериальной порядковой классификации с привлечением ЛПР или эксперта, применительно к проблеме формализации понятия этики искусственного интеллекта. В результате работы метода формируются верхние и нижние границы классов решений (фактически это наборы решающих правил), которые позволяют классифицировать любые наборы оценок по критериям (этических норм).

Построенная классификация выглядит следующим образом: Класс A (верхняя граница): (000000); Класс A (нижняя граница): (100010) (001120) (000220) (001201) (001111) (000211) (001021) (000121); Класс B (верхняя граница): (010000) (101000) (100100) (001210) (100020)

(100001) (001121) (000221); Класс B (нижняя граница): (111200) (101210) (111120) (100220) (111101) (111021) (110121) (101121) (011221); Класс C (верхняя граница): (110210) (101220) (100201) (111111); Класс C (нижняя граница): (111221).

Если бы в исходной системе критериев использовались только бинарные оценки на шкалах критериев, и новости нужно было разделить только на два класса (категории), то можно использовать, например, математический аппарат булевой алгебры, т.е. применить подход, предложенный В. А. Лефевром [18] или методы расшифровки булевых функций [33], а также более эффективный с вычислительной точки зрения, по сравнению с методом ВАР ЦИКЛ, метод ВАР ДИФКЛАСС [34] для решения задачи формализации понятия этики искусственного интеллекта.

Заключение

В статье рассмотрена инициатива IEEE по разработке проектов стандартов этики искусственного интеллекта, предложены возможные определения этики и искусственного интеллекта, проанализирована особенность применения некоторых норм этики применительно к разработке и использованию современных технологий искусственного интеллекта, представлен перечень проектов стандартов этики искусственного интеллекта IEEE (P7000 – P7013), что говорит о достаточно широком спектре проблем, с которыми в самое ближайшее время столкнутся разработчики различных систем искусственного интеллекта. Фактически это означает, что разработчики систем искусственного интеллекта уже сейчас должны начать подготовку к прохождению различных процедур сертификации, т.е. соответствия их продукции, использующей технологии искусственного интеллекта, указанным стандартам искусственного интеллекта. В этой связи критический анализ различного математического инструментария, позволяющего формализовать понятия этики искусственного интеллекта, будет способствовать разработке понятных и прозрачных «правил игры». К средствам формального анализа, которые активно применяются для поставленной задачи, относятся: булева алгебра, многозначные логики, нечеткая логика, теория вероятностей, а также методы вербального анализа решений (ВАР). Благодаря развитию всех этих концепций к настоящему моменту, можно быть оптимистичным относительно успешного решения задачи формализации этики в области искусственного интеллекта. Для примера можно рассмотреть формализацию этики для проекта стандарта P7011 (оценка достоверности новостных источников) с использованием метода ВАР ЦИКЛ. Из нерешенных задач необходимо отметить, что для каждого из проектов стандартов IEEE необходима разработка индивидуальных систем критериев и соответствующих шкал оценок, по которым можно будет принять решение о степени соответствия той или иной технологии искусственного интеллекта этическим стандартам.

ЛИТЕРАТУРА

1. Карпов, В.Э. К вопросу об этике и системах искусственного интеллекта / В.Э. Карпов, П. М. Готовцев, Г. В. Ройзензон // Философия и общество. — 2018. — № 2(87). — С. 84–105. — DOI: 10.30884/jfio/2018.02.07.
2. Ройзензон, Г.В. Проблемы формализации понятия этики в искусственном интеллекте / Г.В. Ройзензон // Шестнадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2018). Труды конференции. В 2-х томах. — Т. 2. — М.: РКП, 2018. — С. 245–252.
3. IEEE. Ethically Aligned Design. —2016. — Режимдоступа: https://ethicsinaction.ieee.org.
4. Бритков, В.Б. Междисциплинарный подход к анализу риска / В.Б. Бритков, Г.В. Ройзензон // IX Московская международная конференция по исследованию операций (ORM2018). Труды. В двух томах / Под ред. Ф.И. Ерешко. — Т. 2. — М.: МАКС Пресс, 2018. — С. 340–345.
5. Ларичев, О.И. Проблемы принятия решений с учетом факторов риска и безопасности / О.И. Ларичев // Вестник АН СССР. — 1987. — Т. 57, № 11. — С. 38–45.
6. Интеллектуальные системы поддержки принятия решений в нештатных ситуациях с использованием информации о состоянии природной среды / В.А. Геловани, А.А. Башлыков, В.Б. Бритков, Е.Д. Вязилов. — М.: Эдиториал УРСС, 2001.
7. Метод многокритериальной классификации ЦИКЛ и его применение для анализа кредитного риска / А.А. Асанов, О.И. Ларичев, Г.В. Ройзензон и др. // Экон. и мат. методы. — 2001. — Т. 37, № 2. — С. 14–21.
8. Бек, У. Общество риска. На пути к другому модерну / У. Бек. — М.: Прогресс-Традиция, 2000.
9. Финн, В.К. Интеллектуальные системы и общество: Сборник статей / В.К. Финн. — М.: КомКнига, 2006.
10. Российская рабочая группа IEEE по вопросам этики ИИ. — 2017. — Режим доступа: http://ecai.raai.org/.
11. Апресян, Р.Г. Этика: учебник / Р.Г. Апресян. — М.: КноРус, 2017.
12. Гусейнов, А.А. Античная этика / А.А. Гусейнов. — М.: URSS, 2017.
13. Дробницкий, О.Г. Моральная философия: Избранные труды / О.Г. Дробницкий. — М.: Гардарики, 2002
14. Кропоткин, П.А. Этика: Избранные труды / П.А. Кропоткин. — М.: Политиздат, 1991.
15. Фролов, И.Т. Этика науки: Проблемы и дискуссии / И.Т. Фролов, Б.Г. Юдин. — М.: URSS, 2016.
16. Швейцер, А. Культура и этика / А. Швейцер. — М.: Прогресс, 1973.
17. Шпеман, Р. Основные понятия морали / Р. Шпеман. — М.: Московский философский фонд, 1993.
18. Лефевр, В.А. Алгебра совести / В.А. Лефевр. — М.: «Когито-Центр», 2003.

19. Таран, Т.А. Булевы модели рефлексивного управления в ситуации выбора / Т.А. Таран // Автоматика и телемеханика. — 2001. — № 10. — С. 103–117.
20. Фоминых, И.Б. О формализации гуманитарного знания / И.Б. Фоминых // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2008). Труды конференции. — Т. 1. — М.: Ленанд, 2008. — С. 133–141.
21. Цыгичко, В.Н. Безопасность критических инфраструктур / В.Н. Цыгичко, Д.С. Черешкин, Г.Л. Смолян. — М.: URSS, 2019.
22. Еремеев, А.П. Темпоральные модели на основе логики ветвящегося времени в интеллектуальных системах / А.П. Еремеев, И.Е. Куриленко // Искус. интеллект и принятие решений. — 2011. — № 1. — С. 14–26.
23. Поспелов, Д.А. «Серые» и/или «черно-белые» / Д.А. Поспелов // Прикладная эргономика. Специальный выпуск «Рефлексивные процессы». — 1994. — № 1. — С. 29–33.
24. Карпенко, А.С. Развитие многозначной логики / А.С. Карпенко. — 3-е изд. — М.: URSS, 2016.
25. Кузнецов, О.П. Неклассические парадигмы в искусственном интеллекте / О.П. Кузнецов // Известия РАН. Теория и системы управления. — 1995. — № 5. — С. 3–23.
26. Тарасов, В.Б. От многоагентных систем к интеллектуальным организациям: философия, психология, информатика / В.Б. Тарасов. — М.: Эдиториал УРСС, 2002.
27. Достоверный и правдоподобный вывод в интеллектуальных системах / В.Н. Вагин, Е.Ю. Головина, А.А. Загорянская, М. В. Фомина; Под ред. В.Н. Вагина, Д.А. Поспелова. — М.: Физматлит, 2008.
28. Zadeh, L.A. Fuzzy sets / L.A. Zadeh // Information and Control. — 1965. — Vol. 8, no. 3. — P. 338–353.
29. Шрейдер, Ю.А. Проблема неполного добра в модели ценностной рефлексии по В.А. Лефевру / Ю.А. Шрейдер, Н.Л. Мусхелишвили // Системные исследования. Методологические проблемы. Ежегодник / Под ред. Д.М. Гвишиани, В.Н. Садовского. — № 25. 1997. М.: УРСС, 1997. — С. 213–224.
30. Ларичев, О.И. Вербальный анализ решений / О.И. Ларичев. — М.: Наука, 2006.
31. Ройзензон, Г.В. Способы снижения размерности признакового пространства для описания сложных систем в задачах принятия решений / Г.В. Ройзензон // Новости искус. интеллекта. — 2005. — № 1. — С. 18–28.
32. Ройзензон, Г.В. Синергетический эффект в принятии решений / Г.В. Ройзензон // Системные исследования. Методологические проблемы. Ежегодник / Под ред. Ю.С. Попкова, В.Н. Садовского, В.И. Тищенко. — №36. 2011-2012. М.: УРСС, 2012. — С. 248–272.
33. Соколов, Н.А. Оптимальная расшифровка монотонных булевых функций / Н.А. Соколов // Журнал вычислительной математики и математической физики. — 1987. — Т. 27, № 12. — С. 1878–1887.
34. Ларичев, О.И. Система ДИФКЛАСС: построение полных и непротиворечивых баз экспертных знаний в задачах дифференциальной классификации / О.И. Ларичев, А.А. Болотов // Научно-техническая информация. Серия 2. Информационные процессы и системы. — 1996. — № 9.— С. 9–15.

Журнал «Современная наука: актуальные проблемы теории и практики»