Содержание

Глоссарий

Искусственный интеллект (*ИИ*) — программная система для решения различных задач с помощью антропоразмерного интеллекта, функционирующего на автоматизированной основе.

Этическая экспертиза — тестирование технологии с точки зрения ее релевантности этическим нормам человека, причастности ценностям и нормативным предписаниям, а также психологической безопасности.

Социально-надежный ИИ — искусственный интеллект, отвечающий нормам этической релевантности, социальной предсказуемости и психологической безопасности.

Агенты ИИ — различные по своему функционалу программы, которые автономно работают по определенному расписанию на основе технологий ИИ над поставленными человеком задачами; способны получать и обрабатывать данные из внешней среды с последующей выдачей рационального результата, соизмеримого поставленным задачам.

Робот — 1) программируемый исполнительный механизм с определенным уровнем автономности для выполнения перемещения, манипулирования или позиционирования (включает систему управления и интерфейс для человека); 2) автономно действующий программный модуль с понятным для человека интерфейсом, который выполняет рутинные задачи в заданной предметной области по определенному расписанию (поиск, ответы на вопросы, сбор данных).

Государственное регулирование экономики — управляющее воздействие государства на экономическую деятельность субъектов хозяйствования, которое реализуется через различные *экономические* и *административные (внеэкономические)* механизмы.

Введение[1]

Искусственный интеллект, а точнее подразумеваемые под ним технологии машинного обучения, которые позволяют компьютерам понимать речь, разбирать тексты, классифицировать данные по заранее заданным критериям, т.е. решать ряд интеллектуальных творческих задач способами, аналогичными тем, которыми пользуется человек [Barr, 1989], прочно входит в дискурс органов исполнительной власти, но также может использоваться законодательной и судебной ветвями власти [Sun, Medaglia, 2019]. В принятой в 2019 году российской национальной стратегии развития искусственного интеллекта до 2030 года под искусственным интеллектом (далее — ИИ) понимается «комплекс технологических решений, имитирующий когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и позволяющий при выполнении задач достигать результаты, как минимум сопоставимые с результатами интеллектуальной деятельности человека» [Указ Президента РФ № 490]. Комплекс технологических решений включает информационно-коммуникационную инфраструктуру, программное обеспечение, в котором в том числе используются методы машинного обучения, процессы и сервисы по обработке данных и выработке решений.

Поскольку спектр задач, решаемых ИИ, постоянно расширяется, мы перечислим только задачи, которые относятся к системе органов исполнительной власти, а также к судебной системе.

---

[1] Книга написана на основе прикладного исследования «Систематизация опыта ведущих стран мира в развитии технологий искусственного интеллекта и выработка предложений по нормативным и организационным мерам, направленным на опережающее развитие технологий искусственного интеллекта в Российской Федерации», проведенного ИГМУ НИУ ВШЭ в 2019 году в рамках программы прикладных исследований. В работе использованы результаты проекта «Трансцендентальный подход в философии: история и современность», выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2020 году.

Прежде всего, это класс задач по установлению содержательных связей между нормативно-правовыми актами или документами стратегического планирования с целью их взаимоувязки или поиска противоречий, несоответствий между целями, задачами и целевыми показателями в документах на различных уровнях управления. Известно, что нормативно-правовые акты могут содержать противоречия, ведь документы стратегического планирования в России созданы на всех уровнях управления, при этом указанные в них цели, задачи и показатели на региональном или муниципальном уровнях могут противоречить, например, ключевым документам на федеральном уровне. Работу по взаимной увязке и поиску противоречий может осуществлять программа на основе технологий искусственного интеллекта, которая в ряде документов устранит дублирование, выявит заведомо недостижимые показатели, проведет конечную классификацию типов документов стратегического планирования в зависимости от критериев классификации.

Другим классом задач для ИИ является предиктивная аналитика на основе массивов данных, которые описывают объект в отрасли, например, пациента по истории болезни в медицинской карте, школьника или студента по портфолио достижений и успеваемости, совершившего или подозреваемого в правонарушении или преступлении по материалам уголовного дела. На основе анализа аналогичных объектов одного класса рекомендательная система может классифицировать каждый новый объект относительно набора признаков, которые также могут быть априори заданы. Таким образом, можно с некоторой вероятностью получить предсказания о склонности заключенного совершить повторное преступление, о намерении ученика получить определенную профессию, о возникновении у пациента некоторой болезни или, например, послеоперационных осложнений. Другой целью рекомендательной системы может стать выработка предписаний о том, какое может быть назначено наказание подозреваемому, какие курсы в дальнейшем предпочтительно слушать школьнику или студенту, какие препараты принимать пациенту и процедуры проходить, чтобы предотвратить негативный сценарий развития болезни [ЦНТИ МФТИ, 2020].

Также у органов власти формируются данные для профиля гражданина в той роли, в которой он выступает объектом их соответствующих полномочий: налогоплательщик, подозреваемый в уголовном или административном правонарушении, предприниматель, пенсионер и т.д. На основе системы предиктивной аналитики по профилю гражданина можно выстраивать так называемые проактивные услуги, о которых гражданин мог не знать и не инициировать их самостоятельно. При этом соответствующие возможности получения поддержки, обеспечения и реализации прав были предусмотрены государством и предложены гражданину со стороны обеспечивающих органов власти в инициативном порядке. Например, проактивными услугами можно назвать предварительный расчет пенсии, начисление социальных пособий, подбор университета или места работы [Добролюбова, 2018].

Искусственный интеллект помогает обеспечивать безопасность граждан, например, сопоставляя лица граждан, попавших в камеры наблюдения, с лицами разыскиваемых нарушителей закона, тем самым идентифицируя их положение и перемещение в случае совпадения. Таким образом, органы внутренних дел получают мощный инструмент для отслеживания разыскиваемых лиц, совершивших правонарушения, выявления лиц, которые отличаются подозрительным или противоправным поведением [Faggella, 2019]. Аналогичные инструменты компьютерного зрения применяются на дорогах для выявления нарушителей правил дорожного движения. Большой потенциал для функционала компьютерного зрения наблюдается в системах «Умный дом» — это совокупность камер, датчиков и иных управляющих элементов, которые круглосуточно накапливают и предоставляют уполномоченному наблюдателю видеоданные о состоянии квартир, домов, придомовых территорий. Таким образом, для построения «умных городов», согласно соответствующей концепции Минстроя России (https://russiasmartcity.ru/), система интеллектуального видеонаблюдения является ее неотъемлемым элементом.

Отдельное направление не только в бизнесе, но и в государстве — замена рутинного труда человека на программу, функцио-

нирующую на основе технологий машинного обучения для выполнения рутинных операций. Примером такой программы является чат-бот, который может отвечать на достаточно простые вопросы граждан, связанные с государственными информационными ресурсами и заданные ему в режиме реального времени в текстовом виде, например, через официальные информационные ресурсы органов власти или судов. Вопросы могут затрагивать целый ряд тем: разъяснение отдельных положений законодательства, диагностику статусов граждан для получения социальной поддержки, консультации по получению государственных услуг. С помощью чат-ботов можно реализовать простые инструменты диагностики по различным направлениям. В основу положен механизм диалога между пользователем и чат-ботом, в котором посредством задания вопросов пользователем в виде текста или голосом может быть выявлена или решена некоторая проблема. Например, с помощью простых вопросов чат-бот может диагностировать наличие некоторого заболевания (в частности, COVID-19), определить, есть ли формальные предпосылки у кандидата пройти по конкурсу на некоторую должность на государственной службе, есть ли право у заявителя получить социальную льготу или субсидию, к примеру, в сельском хозяйстве.

Таким образом, программы на основе машинного обучения позволяют заменить труд человека при решении рутинных задач или хотя бы при их выполнении снизить нагрузку на государственных служащих, переключив усилия сотрудников на более сложные и неоднозначные проблемы.

Вообще говоря, помимо анализа и сопоставления данных, программы на основе искусственного интеллекта могут не только выдавать рекомендации, но и принимать решения в автоматическом режиме. Вопрос состоит в готовности ответственных управленцев контролировать каждое решение либо полагаться на решение, предложенное программой. Ярким примером может служить система государственного контроля и надзора, в которой применяется риск-ориентированный подход [Кнутов, Плаксин, 2019]. Данные о проверках объектов, требующих надзора (заводы, пред-

приятия, учреждения сферы образования, здравоохранения, сфера общественного питания), могут быть автоматически проанализированы и сопоставлены с нормативами, определяющими риски непроведения своевременных проверок. Таким образом, программа в состоянии присваивать риски поднадзорным объектам, автоматически устанавливая классы этих рисков, определяя в дальнейшем частоту проверок. При подобном подходе подразумевается только контроль со стороны человека, а выработка решения может оставаться за программой.

Аналогичным образом программа на основе технологий искусственного интеллекта может взять на себя функционал оценки регулирующего воздействия (ОРВ), которая проводится для целей государственного регулирования, определения возможных вариантов достижения целей, а также оценки связанных с ними позитивных и негативных эффектов [Клименко, Минченко, 2016]. При обеспечении сбора подробных цифровых данных об отрасли, подвергающейся регулированию, ИИ может просчитывать текущие и прогнозные ключевые экономические показатели государственной политики в отдельно взятой отрасли. При совершенствовании выбора решений на основе машинного обучения полученные прогнозы могут быть точнее, чем достаточно субъективные подходы, которые предлагаются отраслевыми экспертами. Тем не менее с учетом вероятностного расчета значений показателей с помощью ИИ полученные результаты также требуют контроля со стороны отраслевых экспертов и ответственных государственных служащих. Проведение ОРВ с привлечением функционала искусственного интеллекта существенно убыстряет процесс выработки и оценки альтернатив, которые возможны для совершенствования нормативно-правовой базы и основных показателей государственной политики в некоторой отрасли, а значит ИИ остается перспективной технологией для проведения ОРВ.

Осуществление государственной бюджетной и налоговой политики для стимулирования бизнеса также можно свести к задаче определения налогового режима и объема поддержки для различных компаний в зависимости от их финансово-экономических по-

казателей, сравнимых с установленными эталонными. Обученные нейронные сети могут не только классифицировать компании по критериям для определения налогового режима и бюджетной поддержки, но позволяют уточнить эти критерии на основе проанализированного множества собираемых показателей о деятельности компаний.

Абсолютно аналогично решаются задачи осуществления лицензионной и разрешительной деятельности органов власти, которые сравнимы с задачей кредитного скоринга клиента в банке. Организация, деятельность которой подлежит лицензированию или требует получения разрешений, может быть в автоматическом режиме оценена нейросетью на основе ранее изученных аналогичных данных. В таком случае организация может быть автоматически классифицирована под положительное либо отрицательное решение о выдаче лицензии или разрешения. Окончательное решение может быть принято сотрудником-специалистом, однако подавляющая часть предварительных расчетов для определения параметров выдачи лицензии или разрешения может быть проведена программой на основе технологий ИИ.

Приведенный выше обзор возможностей ИИ, призванных помочь в решении разных классов задач в государственном управлении и в судебной системе, показывает неизбежность внедрения инновационных решений на основе машинного обучения в ближайшем будущем. Искусственный интеллект дает возможность сократить издержки при осуществлении государственных функций, увеличить скорость отклика на запросы граждан, повысить качество результатов взаимодействия органов власти с внешними акторами, а также перераспределить нагрузку на государственных служащих, избавив их от решения рутинных задач. При этом изложенные выше возможности ИИ демонстрируют позитивный эффект от его внедрения в деятельность органов власти. Однако деятельность органов государственной власти сопряжена с пристальным вниманием общественности, требованиями соблюдать прозрачность и подотчетность в принятии решений и представлении результатов. Если опорой деятельности для органов власти становится ИИ, то госу-

дарству следует обеспечить важнейший аспект внедрения ИИ в операционную и стратегическую деятельность — этический.

Анализ зарубежных и российской национальной стратегии развития искусственного интеллекта демонстрирует, что в подавляющем большинстве в них включен раздел по обеспечению этичности использования ИИ в деятельности органов власти. Как правило, в стратегиях определяются этические принципы внедрения ИИ (конфиденциальность, прозрачность, подотчетность, защита гражданских прав, ответственность, доверие, устойчивое развитие, справедливость, приоритет конечного решения за человеком), декларируется необходимость защиты персональных данных граждан, принятие решений с помощью технологий ИИ согласно важным для общества ценностям и убеждениям. Важную роль в принятии этических принципов развития ИИ сыграли ОЭСР и ЕС[2]. Этические механизмы в национальных стратегиях развития ИИ – это один из механизмов контроля над результатами деятельности на основе использования ИИ. Например, декларируется право человека принимать конечные решения и вмешиваться при необходимости в деятельность, реализованную на технологиях ИИ. Использование ИИ должно вызывать доверие у общества: должны быть созданы механизмы верификации действий ИИ, механизмы защиты граждан от ошибок ИИ, которые могут нанести урон человеку, привести к дискриминации одних групп граждан перед другими при принятии решений, ущемлять моральные, религиозные ценности и достоинство граждан при реализации управленческих механизмов на основе ИИ. Отдельно государства декларируют создание консультативных органов, которые будут дискуссионными площадками для выработки политики по реализации этических принципов в продуктах и услугах на основе ИИ.

---

[2] Forty-two countries adopt new OECD Principles on Artificial Intelligence (URL: https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm); Ethics Guidelines for Trustworthy AI (URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai).

В Российской Федерации принята национальная стратегия развития искусственного интеллекта на период до 2030 года, создан теоретический и технологический задел в области ИИ, сформирован рынок цифровых продуктов и услуг на основе ИИ. Государственная политика России предусматривает содействие повсеместному отраслевому внедрению ИИ, включая сектор государственного управления. В качестве прикладных областей исследования мы выбрали очень чувствительные для общества сферы, в которых вопросы прозрачности, подотчетности, справедливости, соблюдения прав граждан не менее важны, чем инвестиции, внедрение цифровых технологий, создание стандартов. Это — судебная система, отрасли образования, здравоохранения, жилищно-коммунального хозяйства.

Цель данной книги заключается в разработке основанного на этике подхода к внедрению технологий ИИ в работу органов государственной власти (с акцентом на примеры из практики в Российской Федерации). Сопутствующая задача в этом исследовании — это обнаружение условий формирования, внедрения и адаптации возможностей этичного ИИ для нужд государственного управления. Отдельная задача состоит в прогнозировании и обнаружении таких нужд. Важнейшей областью участия государства в вопросах развития ИИ является разработка этического кодекса ИИ, а также контроль за его использованием. Бурное развитие ИИ в современном мире сопровождается повышением числа этических трудностей, дилемм и коллизий на пути внедрения и применения ИИ. В то же время для разработчиков и конечных пользователей существует понимание важности того, что внедрение ИИ должно соответствовать этическим нормам и ожиданиям людей, которые требуют детального рассмотрения и уточнения.

Этически ориентированный, или так называемый надежный (reliable), ИИ включает три взаимосвязанных аспекта, каждый из которых предполагает отдельный уровень экспертной оценки.

Во-первых, его разработка и применение должны соответствовать существующему законодательству и правоприменительным практикам, не нарушать прав и обязанностей граждан, не мешать

их исполнению. Успешная реализация настоящего требования предполагает проведение юридической экспертизы с привлечением советующих квалифицированных специалистов.

Во-вторых, разработка надежного ИИ предполагает соблюдение и воспроизводство ключевых этических норм, принципов и ценностей, присущих определенной культуре или региону. Для их соблюдения также необходимо привлечение соответствующих экспертов, а именно тех, кто обладает достаточными компетенциями для проведения самой этической экспертизы ИИ. В число таких экспертов могут входить люди с междисциплинарной подготовкой и разносторонней квалификацией: философы, психологи, социологи, управленцы и др. [Russell, Norvig, 2009].

В-третьих, разработка и внедрение ИИ должны быть надежными с социальной точки зрения. Под этим параметром подразумевается этическая безопасность применения ИИ, когда объектом воздействия является большое количество людей или общество в целом. Даже при успешном прохождении этической экспертизы системы искусственного интеллекта могут причинить непреднамеренный или незаметный вред в краткосрочной перспективе. Для устранения таких рисков применение ИИ должно поддерживаться не только технологиями конечного внедрения, но и процедурами долгосрочного мониторинга.

В оптимальной перспективе все три компонента должны работать слаженно и поддерживать друг друга. Если на практике между этими компонентами возникает конфликт или несогласованность экспертных сообществ, группы разработчиков и государство должны стремиться к их устранению.

Мы сосредоточимся главным образом на втором и частично на третьем принципах разработки надежного ИИ. Нашей задачей является анализ этических аспектов разработки и внедрения ИИ, а также анализ его социального и антропоразмерного потенциала. Настоящая работа преследует три цели: *описательную, аналитическую и рекомендательную.*

В рамках первой цели мы предлагаем схематичный обзор возможных применений ИИ в судебной системе и определенном от-

раслевом сегменте (образование, здравоохранение, ЖКХ). Данный обзор ни в коем случае не претендует на полный охват существующих и разрабатываемых технологий ИИ. Он нацелен на подбор иллюстративного материала для ситуаций, с которыми могут столкнуться управленцы-практики в выбранных нами отраслях, анализ которого поможет решить текущие и спрогнозировать дальнейшие этические затруднения при внедрении технологий ИИ.

Вторая цель, аналитическая, является центральной для данного руководства и указывает собственно на этически проблемные стороны внедрения и применения ИИ в определенных отраслевых сегментах и судебной системе. Согласно аналитическому замыслу раздела мы не ограничиваемся одной лишь констатацией возможных проблем, но показываем логику их формирования, что должно помочь обнаружению и прогнозированию аналогичных этических трудностей работы технологий ИИ и в других отраслевых сегментах, которые не вошли в данную книгу.

Наконец, рекомендательная цель книги заключается в описании критериев улучшения работы ИИ в случае возникновения этических затруднений и проблем при его внедрении. Реализация этой цели не предполагает формирования полного и детализированного перечня управленческих практик, в которых ИИ-технологии способствуют совершенствованию отраслевой деятельности, так как постоянные инновации в разработках ИИ быстро делают такую работу неактуальной. Нам важнее указать на общие типы проблем при этической оценке ИИ, которые можно в дальнейшем распространять на новые ситуации.

В свою очередь, для реализации рекомендательной цели на более системном уровне требуется отдельный документ, подготовленный на основе данной работы соответствующей группой специалистов. Настоящая книга состоит из четырех частей, в каждой из которых описан соответствующий блок отраслевых инноваций ИИ, дан аналитический разбор потенциальных несовершенств этического характера и пул возможных рекомендаций по их устранению.

Книга построена так, что разбор ключевых сложностей внедрения ИИ в различных отраслевых сегментах дан поблочно (на мате-

риале конкретных кейсов). В начале книги приведен общий анализ прикладных проблем, связанных с этикой разработки и внедрения искусственного интеллекта, затем изложены руководящие принципы и ценности этически ориентированного искусственного интеллекта и далее анализ этических проблем внедрения по секторам. Каждая глава посвящена отдельной отрасли и делится на две части: в первой дан собственно обзор этических аспектов применения и внедрения ИИ и, во второй, пул рекомендаций по возможному их устранению или минимизации. Таким образом, текст настоящей работы имеет выраженно прикладное и рекомендательное назначение. В книге выделяются четыре основные сферы, где использование ИИ требует решения ряда этических проблем:

1. *Судебная система. Конкретная правовая и судебная практика.*

2. *Образование. Отслеживание успехов учащихся и целевая помощь в планировании образования и трудовой карьеры.*

3. *Здравоохранение. Медицинская помощь гражданам, комплекс диагностических процедур, системы мониторинга здоровья и терапии.*

4. *ЖКХ. Системы «Умный дом».*

Мы надеемся, что помимо студентов, аспирантов, представителей академического сообщества наша книга будет полезна государственным управленцам, экспертам, которые на практике сталкиваются с различными барьерами при отраслевом внедрении технологий ИИ, а также осуществлении государственного отраслевого регулирования согласно этическим принципам и верховенству закона.

Руководящие принципы и ценности этически-ориентированного ИИ

Индустрия ИИ и робототехники в современном мире стремительно развивается, хотя многие вопросы взаимодействия ИИ с обществом остаются либо принципиально нерешаемыми, либо неуспевающими за технологическим развитием. Революция робототехники обещает массу преимуществ, но, как и в случае с другими новыми технологиями, она сопряжена с рисками и новыми вопросами, с которыми общество должно столкнуться. Использование роботов должно соответствовать законодательным нормам и этическим принципам, но сами эти нормы и принципы могут быть не до конца выяснены в области взаимодействия с ИИ. Социальная рефлексия нередко медленно «догоняет» технологические инновации, что приводит к образованию «этического вакуума» [Moor, 1985]. Такой эффект означает, что ИИ-технологии развиваются по своим собственным законам, и это приводит к ряду этических коллизий и противоречий. Чтобы их избежать, необходимо провести ревизию существующих этических лакун в области использования ИИ и наметить пути их устранения. Искусственный интеллект все еще находится в зачаточном состоянии с точки зрения сопутствующих его развитию этических исследований.

Существует *две взаимосвязанные проблемы* использования ИИ с точки зрения этики. *Первая* — согласование работы ИИ с существующими в обществе ценностными установками. *Вторая* — формализация данных ценностных установок. Большинство программ ИИ ориентированы на нейтральный анализ данных. Однако для ряда задач, связанных с оценкой человеческой деятельности, это невозможно, противоречит законодательству или неэтично. Отсутствие ценностной составляющей лишает смысла работу программы для человеческих нужд. Одним из естественных способов минимизации риска причинения вреда программами ИИ (например, роботами) является их программирование в соответствии с нашими законами или этическим кодексом. Однако здесь перед

нами встает ряд трудностей: законы могут быть расплывчатыми и контекстно-зависимыми, а сами роботы могут оказаться достаточно сложными для программирования в соответствии со всеми правовыми и этическими нормами. Более того, в силу ограниченности ценностных выборов конкретной ситуацией, бесконечность вариантов решений, с которым работает программа ИИ, может всерьез затруднить выработку этического решения.

Значительное число этических вопросов посвящено рискам дискриминационной практики алгоритмов, которые воспроизводят или даже усиливают враждебные настроения в обществе. Основное внимание должно быть уделено понятию дискриминации, как индивидуальной, так и коллективной, чтобы заложить основу измерения дискриминационной предвзятости, инструментов ее выявления и возможной коррекции. Например, постановление Европейского парламента 2016/679 от 27 апреля 2016 года [Regulation (EU), 2016/679] строго регулирует сбор персональных данных (религиозного, политического, сексуального, этнического характера и т.д.) и запрещает лицам, ответственным за алгоритмические решения, принимать их во внимание при автоматизированной обработке. Следует также уделить особое внимание уязвимым группам населения, таким как дети и инвалиды, а также другим лицам, которые исторически находились в неблагоприятном положении, подвергались риску изоляции или оказывались в ситуациях, характеризующихся асимметрией власти или информации (между работодателями и работниками, производителями и потребителями).

Для такого класса задач характерна проблема формализации собственно человеческих решений. Люди не являются полностью рациональными агентами. Вместо того чтобы рассчитывать свои действия по максимизации суммы функций полезности, мы зачастую имеем конкурирующие намерения. Иногда эмоциональные реакции не позволяют действовать рационально. Не все человеческие решения оказываются безупречными при их этической оценке. Следовательно, возникает более сложный вопрос: как выработать и формализовать этические принципы для ИИ. Некоторые

исследователи занимались вопросом их формализации [Hibbard, 2001; Yudkowsky, 2004; Muehlhauser, Helm, 2012]. В данных исследованиях ключевым вопросом было то, какие можно найти пути для преодоления противоречий между четкостью вычислительных алгоритмов ИИ и неоднозначным, непоследовательным, субъективным разнообразием человеческих ценностей. Например, иногда одни исследователи предлагают основывать системы искусственного интеллекта на нравственности, но при этом не дают точного объяснения того, как агент по искусственному интеллекту должен выбирать действия, последовательно базирующиеся на ней [Haidt, Kesebir, 2010].

Существует целый класс этических проблем, связанных с этикой предсказуемости. Множество программ ИИ пишется для решения прогностических задач. Для ряда программ (например, в области медицины) уровень их эффективности достаточно высок и продолжает расти. Однако при непосредственном взаимодействии человека и программы на стороне человека могут возникать коллизии этического характера. В частности, не ясно, в какой степени пациенты должны и желают знать исход протекающего заболевания, риски возникновения новых заболеваний, характер новых болезней и проч. Также не вполне ясно, как должно регулироваться поведение лечащих врачей в ситуации такой осведомленности и широких возможностей прогнозирования, которая обеспечивается ИИ.

Точно так же программы по тестированию психологической совместимости (например, супружеских пар) могут иметь высокую предсказательную силу. Агент ИИ на основе уже известной информации о человеке может моделировать ценности и поведение людей, которые он наблюдает в течение достаточно длительного периода времени и лучше человека может предсказывать результаты выбора разных опций. Но последствия такого взаимодействия между ИИ и человеком представляют собой этическую трудность.

Также различные программы тестирования способностей студентов или планирования профессиональных и карьерных портфолио для абитуриентов сталкиваются с проблемой готовности пользователя программы ознакомиться с результатом. В последу-

ющем могут возникать различные виды субъективного дискомфорта, связанного с «программированием» выбора или эффектами снижения мотивации и проч.

Решение данных вопросов осложняется тем, что формализовать человеческий запрос не всегда возможно. Ряд авторов [Muehlhauser, Helm, 2012] пришли к выводу, что люди не могут точно описать свои собственные ценности. Часто мы можем наблюдать у людей противоречия в их ценностной системе или отсутствие однозначных ответов на этические вопросы. Решения, принятые быстро, могут разительно отличаться от решений, которые долго продумывались или обсуждались. Ценности также могут изменяться в зависимости от пережитого опыта, под влиянием ближайшей среды и т.д.

В целом «ценностные» установки большинства программ ИИ на сегодняшний день — оптимизационные (или утилитарные). Таким образом, этический Кодекс ИИ — прагматический.

В связи с этими установками возникает масса проблем. Например, при выполнении инструкций по максимизации прибыли «Максимально увеличить прибыль, но никому не навредить» — для ИИ остается неясным, что имеется в виду под вредом. Идет ли речь о том, что не стоит заключать жесткие сделки, или не следует продавать никому то, что им не нужно? «Размышляя» над этими вопросами, ИИ может прийти к выводу, что в случае соблюдения всех этих правил, акционеры, сотрудничавшие с компанией, могут потерять свои деньги, что, действительно, причинит им вред [Raphael, 2009].

Есть и другой аспект. Поскольку рациональный агент выбирает действия для максимизации ожидаемой суммы функций полезности, он может выбирать действия, которые не увеличивают полезность в краткосрочном периоде, но в целом увеличивают его способность повысить полезность в будущем. Такие действия могут включать в себя самозащиту, увеличение собственных ресурсов и возможностей агента.

Если углубляться в детали этих компромиссов, то придется пересматривать этическую модель ИИ. Однако проблема не только в том, что недостаточно провести изменения в инструкции высокого

уровня (такой как «максимизация прибыли»). Не совсем ясно, как сформулировать инструкции для ИИ так, чтобы сбалансировать двусмысленность максимизации прибыли с минимизацией вреда. Прибыль — это ценность, связанная с результатами, но существуют и другие способы определения ценности результатов, которые в более общем плане влияют на благосостояние человека [Arkin, 2013].

Процесс человеческой оценки тех или иных ситуаций зависит от слишком многих правил, условий и влияния среды, чтобы его можно было сформулировать через перечень конкретных аналитических высказываний. Существует мнение, что в силу отсутствия таких правил или точных способов их поиска [Tasioulas, 2017], более успешной стратегией является использование статистических данных для решения различных дилемм (как люди и с какой частотой поступают в конкретных ситуациях). Это было бы более эффективным способом взаимодействия с ИИ, однако потребовалась бы огромнейшая база данных и сложные алгоритмы ИИ для такого масштабного исследования человеческих ценностей статистическим путем. Это создает проблему курицы и яйца для этического ИИ: изучение человеческих ценностей требует мощного ИИ, но этический ИИ требует знания человеческих ценностей.

Таким образом, в масштабах проблемы этической оценки ИИ появляется дилемма. Если мы могли бы запрограммировать кодекс этики для регулирования поведения роботов, какую этическую теорию мы использовали бы? Или же роботы должны рассматриваться исключительно инструментально (как оружие, компьютеры и т.д.) и регулироваться соответствующим образом?

На эти вопросы предстоит ответить как широкой общественности, так и узким специалистам, занимающимся разработкой ИИ. Важную роль в экспертной подготовке решений поставленных проблем должны играть специальные комитеты по этической экспертизе разработки и внедрения того или иного продукта ИИ, куда будут входить специалисты из междисциплинарных областей. В ходе открытого обсуждения и свободной дискуссии ими будут вырабатываться решения по тем или иным этическим вопросам внедрения и успешной реализации ИИ.

Ключевые этические рекомендации, лежащие в основе настоящего документа, выстроены на основе определенных принципов и ценностей, которые мы последовательно артикулируем в этой части. Таким образом, настоящий документ основан на следующих этических принципах и ценностях.

1. *Социальные принципы*:

а. Уважение автономии человека на основе равенства прав и доступа к реализации потребностей в рамках закона.

б. Справедливость в распределении благ, социальная и правовая справедливость.

в. Право знать, как устроены те или иные социальные механизмы, в том числе включающие ИИ-технологии.

г. Отсутствие любых форм дискриминации или ущемления прав граждан, внимание к наиболее уязвимым группам граждан.

2. *Экологические принципы*:

Сохранение природы, разумное использование ресурсов, техническая надежность и безопасность.

3. *Правовые принципы*:

Верховенство закона, неприкосновенность частной жизни, надзор за законностью применения новых технологий и использованием старых.

4. *Информационные принципы*:

Прозрачность, объяснимость, доступность и открытость информации.

Системы ИИ должны функционировать в соответствии с этическими ценностями обществ, в которых они будут реализованы. Несовпадение этих значений между человеком и машиной может препятствовать их эффективному взаимодействию.

Вопрос о том, какие ценности машины должны использовать и как внедрить эти ценности в научном поле, до сих пор остается дискуссионным. В настоящее время рассматривается несколько влиятельных этических теорий, регулирующих взаимодействие человека и ИИ (деонтологические, утилитарные, этики добродетели и т.д.).

На этом уровне мы находим два типа взаимосвязанных проблем:

1. Каким образом решать этические проблемы, возникающие при использовании ИИ?

2. Как формализовать выработанные этические стратегии для самого ИИ?

Способствовать решению данных проблем может разработка руководства ИИ по этике, которое может быть адаптировано к конкретным профессиям и реальным сценариям. В нем будут изложены основные принципы и ценности, которые должны функционировать в системах искусственного интеллекта. Они могут основываться на его способности усваивать примеры и динамически адаптироваться к реальным ситуациям, с которыми придется столкнуться в определенной области.

Отдельно обозначим фундаментальные этические принципы работы ИИ:

1. ИИ не должен вредить людям своими действиями или бездействием.

2. ИИ обязан выполнять приказы людей, только если это не вступает в противоречие с п. 1.

3. ИИ должен стремиться продолжать свое существование и поддерживать его, кроме тех случаев, когда это противоречит п. 1 или п. 2 [Pei, 2018].