

МИНИСТЕРСТВО ВНУТРЕННИХ ДЕЛ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ КАЗЕННОЕ УЧРЕЖДЕНИЕ
«ГЛАВНЫЙ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИЙ ЦЕНТР»



ОБЗОР
ОТДЕЛЬНЫХ ВОПРОСОВ
В ОБЛАСТИ БОЛЬШИХ ДАННЫХ
И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА
V ВЫПУСК
том 1

МОСКВА 2021

Под общей редакцией В.С. Овчинского,
доктора юридических наук,
генерал-майора милиции в отставке

Обзор отдельных вопросов в области
больших данных и искусственного интеллекта.

V выпуск, 1 том.

М.: ФКУ «ГИАЦ МВД России», 2021. – 479 с.

В обзоре представлены материалы, касающиеся вопросов использования больших данных и развития искусственного интеллекта в мире. Отдельно отражены актуальные вопросы использования ИИ в правоохранительной системе.

В МВД России подведены итоги первого в истории ведомства хакатона «Искусственный интеллект на службе полиции»



20 мая 2021 года в МВД России в режиме видеоконференции были подведены итоги хакатона, организатором которого выступило федеральное казенное учреждение «Главный информационно-аналитический центр Министерства внутренних дел Российской Федерации».

Основная цель конкурса – повышение эффективности оперативно-служебной деятельности подразделений министерства за счёт использования новейших технологий цифровой трансформации.

В нём приняли участие пятнадцать команд технических факультетов Московского университета МВД России им. В.Я. Кикотя, Краснодарского и Санкт-Петербургского университетов, Воронежского института МВД России, а также три команды Военно-космической академии имени А.Ф. Можайского.

Команды были сформированы из отличившихся курсантов – от 4 до 7 человек в каждой, поддерживали их опытные преподаватели образовательных организаций. За командами были закреплены кураторы из числа сотрудников Главного информационно-аналитического центра.

В течение месяца учащиеся – разработчики, проектировщики, тестировщики и дизайнеры программного обеспечения – создавали проекты

автоматизированной информационной системы, использующей технологии искусственного интеллекта для решения оперативно-служебных задач, связанных с предупреждением и раскрытием преступлений.

Следует отметить, что участники не ограничивались в выборе тем проекта. Они сами обозначали проблему, разрабатывали техническое задание и реализовывали проект, демонстрируя полученные при обучении навыки и знания в области искусственного интеллекта, машинного обучения и компьютерного зрения.

В состав жюри вошли представители ФКУ «ГИАЦ МВД России», ФКУ НПО «СТиС» МВД России, Академии управления МВД России и Московского государственного технического университета им. Н.Э. Баумана.

Более трёх часов конкурсанты представляли свои презентационные видеоролики разработанных IT-решений. В напряжённой борьбе **первое место заняла команда «Василиск.AI» Московского университет МВД России им. В.Я. Кикотя** с программой, способствующей поиску похищенного антиквариата на открытых торговых онлайн-площадках.

На **втором месте – команда Ё-КУСП Воронежского института МВД России** с разработкой, позволяющей проводить интеллектуальный анализ текстовых сообщений о преступлениях, правонарушениях и происшествиях. **Второе место** также заняла команда **«Souz Team» Военно-космической академии имени А.Ф. Можайского**, спроектировавшая сервис оперативного распознавания человеческой речи и семантического анализа текстов.

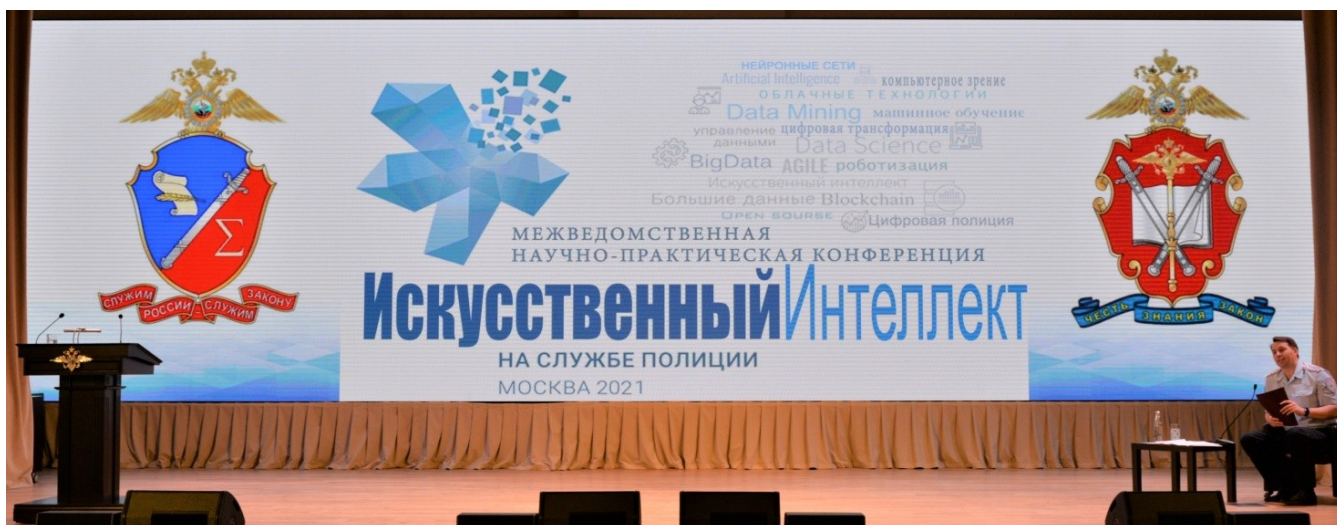
Третье место присуждено программистам из команд: **«CopIB» Московского университета МВД России им. В.Я. Кикотя**, осуществившей проект интеллектуальной системы идентификации граждан, **«Star Track» Военно-космической академии имени А.Ф. Можайского**, создавшей интеллектуальную систему поддержки принятия решения по прогнозированию и раскрытию преступлений, а также **«IT-портрет» Воронежского института МВД России**, реализовавшей проект автоматизированной системы сбора

и учёта элементов внешности для изготовления субъективных портретов разыскиваемых лиц.

Также жюри в отдельной номинации **«За актуальность идеи»** была отмечена команда **«LamerTeam» Краснодарского университета МВД России** с системой предиктивной аналитики, включающей поиск и получение данных для прогнозирования будущих криминальных событий.

Торжественное представление проектов победителей хакатона и их награждение состоится в рамках межведомственной научно-практической конференции **«Искусственный интеллект на службе полиции»** 27 мая 2021 года в Академии управления МВД России.

В МВД России обсудили вопросы внедрения искусственного интеллекта в работу полиции



В Академии управления МВД России состоялась межведомственная научно-практическая конференция «Искусственный интеллект на службе полиции», организованная совместно с Главным информационно-аналитическим центром МВД России.

Открыл мероприятие в онлайн режиме заместитель Министра внутренних дел Российской Федерации генерал-полковник полиции Виталий Шулика.

Заместитель Министра во вступительном слове отметил, что в рамках конференции планируется выработать подходы к решению задач по внедрению систем искусственного интеллекта и методов машинного обучения в процессы выявления, предупреждения и раскрытия преступлений.

«В деятельность ведомства уже активно внедряются новые программные «продукты», использующие искусственный интеллект. Эти технологии в разы повышают эффективность раскрытия преступлений», – подчеркнул Виталий Шулика.

В конференции приняли участие и выступили с докладами: начальник ФКУ «ГИАЦ МВД России» полковник внутренней службы Владимир Агеев, сотрудники центрального аппарата МВД России, руководители территориальных органов МВД России, начальник Академии управления МВД России генерал-лейтенант полиции Андрей Конев, ведомственных

научных и образовательных организаций (в режиме онлайн), а также представители общественных организаций и лидеры IT-компаний.

Собравшиеся обсудили актуальные вопросы прогнозирования раскрываемости преступлений в рамках современных возможностей анализа структурированных данных, также были рассмотрены правовые и организационные аспекты внедрения технологий искусственного интеллекта в систему МВД России. Все выразили готовность и в дальнейшем обмениваться актуальными разработками, практическим опытом внедрения интеллектуального программного обеспечения в различные сферы деятельности, в первую очередь – в работу российской полиции.

Начальник Главного информационно-аналитического центра полковник внутренней службы Владимир Агеев подвёл итоги первого в истории ведомства хакатона «Искусственный интеллект на службе полиции», который завершился 20 мая 2021 года, и анонсировал создание сообщества (клуба) – постоянно действующей площадки для обмена мнениями и опытом в области цифровизации, повышения информационной безопасности для IT-специалистов ведомства, учащихся научных и образовательных организаций.

В свою очередь начальник Академии управления МВД России генерал-лейтенант полиции Андрей Конев, обратившись к участникам мероприятия, отметил, что для защиты граждан от преступных посягательств органы внутренних дел стремятся максимально использовать передовые технологии в сфере искусственного интеллекта. Сегодня внедрение в практику подобных технологий требует значительных технологических решений.

ОГЛАВЛЕНИЕ

На пути к регулированию искусственного интеллекта.....	11
Международная перспектива регулирования систем искусственного интеллекта на основе стандартов совета Европы	15
Влияние ИИ на права человека, демократию и верховенство права	34
ИИ и уважение к человеческим ценностям	39
ИИ и свобода личности	44
ИИ и верховенство закона	52
Руководящие принципы этики ИИ: Европейские и глобальные перспективы	63
Обсуждение и нормативно-этический анализ.....	91
Анализ международных юридически обязательных документов	107
Национальные перспективы регулирования систем ИИ	186
Риски и вызовы, создаваемые ИИ в области прав человека, демократии и верховенства закона	203
Международная деятельность и сотрудничество	223
Управление искусственным интеллектом в Японии	226
Проект RIKEN Центр передовой разведки.....	230
Обсуждение ИИ в министерствах и ведомствах.....	237
«Общество 5.0» / Четвертая промышленная революция / Подключенные отрасли.....	250
ИИ-приложения в Мексике. Взгляд изнутри	263
Коллективная основа для регулирования ИИ и о регулировании ИИ	295
Недобросовестное использование и злоупотребление искусственным интеллектом	304
Вредоносное ПО ИИ	311

Злоупотребление облачными сервисами ИИ	317
Угадывание пароля с помощью ИИ	319
Подпольный инструмент для взлома CAPTCHA под названием XEvil	
Шифрование с помощью ИИ	323
Взлом с поддержкой ИИ.....	329
Социальная инженерия	332
Генерация контента.....	337
Анализ содержимого	339
Робоколлинг v2.0	342
Криминальная бизнес-разведка	343
Злоупотребление системами распознавания изображений автономных автомобилей	345
Компрометация бизнес-процессов и внедрение спутниковой телеметрии...	357
Инсайдерские атаки: банковские и торговые площадки.....	359
Заражение данных резидентным вредоносным ПО	361
Вымогатели с поддержкой ИИ.....	363
ИИ во благо:.....	365
Пример из практики: глубокое погружение в Deefakes	367
Возможные причины низкой скорости принятия Deepfakes	378
Некоторые возможные будущие угрозы Deepfakes.....	380
Противодействие Deepfakes	383
Искусственный интеллект-2021	421
Начало Начала	428

Основные положения	429
Защита Америки в эпоху искусственного интеллекта.....	432
Победа в технологическом конкурсе.	437
Концепции ИИ и технологии военного назначения.	452
Современное применение ИИ в ключевых областях	465
Границы технологии искусственного интеллекта	466
ИИ как двигатель прогресса.....	471

НА ПУТИ К РЕГУЛИРОВАНИЮ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Глобальные перспективы развития правовой базы в области систем
искусственного интеллекта на основе стандартов Совета Европы

Декабрь 2020 года

Эта публикация была профинансирована за счет добровольного взноса Японии

Системы искусственного интеллекта (ИИ) все чаще используются практически во всех видах человеческой деятельности. Преимущества этой технологии признаны и уже являются частью нашей повседневной жизни. В контексте борьбы с Covid-19 были развернуты многочисленные приложения для ускорения исследований, улучшения выявления случаев заболевания и измерения пандемии. Однако развитие этой технологии вызывает озабоченность общественности, и государства несут ответственность за обеспечение того, чтобы в эту новую технологическую эпоху права человека, демократия и верховенство права по-прежнему были полностью защищены. Совет Европы является ведущей правозащитной организацией континента с 47 государствами-членами и защищает и продвигает права человека, демократию и верховенство права на общеевропейском уровне. Таким образом, Совет Европы играет четкую роль в решении вопроса о развитии и использовании искусственного интеллекта. Организация уже разработала новаторские глобальные юридически обязательные стандарты, касающиеся сложных технологических вопросов, таких как защита персональных данных, биоэтика и киберпреступность, согласование инноваций и защита прав человека. Кроме того, Совет Европы разработал множество документов, касающихся воздействия систем искусственного интеллекта на права человека, демократию и верховенство права. Был сделан еще один шаг с созданием специального комитета по искусственному интеллекту (САНАИ), который отвечает за изучение на основе широких консультаций с участием многих

заинтересованных сторон осуществимости и потенциальных элементов правовой основы для разработки, разработки и применения искусственного интеллекта на основе стандартов Совета Европы в области прав человека, демократии и верховенства права. Комитет имеет уникальный состав, объединяющий государства-члены и государства-наблюдатели, а также наблюдателей от гражданского общества, научных кругов и частного сектора. Будучи убежденным в важности глобальной рефлексии и объединения усилий в этой области, САНАІ работает в тесном сотрудничестве с другими международными учреждениями, такими как ЮНЕСКО, ОЭСР и Европейский союз.

Вступление

Эта публикация направлена на поддержку текущих размышлений в рамках САНАИ по анализу проблем, возникающих в связи с системами ИИ, и возможных ответных мер со стороны регулирующих органов. Во-первых, в нем ставится цель проинформировать читателя о ходе работы САНАИ и представлено несколько исследований, подготовленных под эгидой САНАИ, о влиянии систем ИИ на права человека, верховенство права и демократию, а также на существующие международные юридически обязательные документы и этические руководящие принципы в области ИИ. Во-вторых, он учитывает национальные точки зрения различных государств, с тем чтобы поддержать разработку международной правовой базы по использованию определенных систем искусственного интеллекта на основе стандартов, установленных Советом Европы.

Мы уверены, что эта публикация даст ценный взгляд на текущие дебаты по регулированию систем искусственного интеллекта и подчеркнет уникальный вклад, который вносят Совет Европы и САНАИ.

1. Международная перспектива

В предварительной главе представлен настоящий доклад, представленный САНАИ Комитету министров, и подробно описан прогресс, достигнутый на сегодняшний день, с учетом воздействия мер по борьбе с пандемией COVID-19. Он также включает в себя размышления о методах работы, синергии и взаимодополняемости с другими соответствующими заинтересованными сторонами и предложения о дальнейших действиях со стороны САНАИ с помощью надежной и четкой дорожной карты.

В главе 1 описывается влияние ИИ на права человека, демократию и верховенство права. В нем определяются те права человека, закрепленные в Европейской конвенции о правах человека («ЕСПЧ»), Протоколах к ней и

Европейской социальной хартии («ЕСК»), которые в настоящее время в наибольшей степени затронуты или, вероятно, будут затронуты ИИ.

В главе 2 представлен соответствующий свод документов мягкого права и других этико-правовых рамок, разработанных правительственными и неправительственными организациями по всему миру с двойной целью. Во-первых, мы хотим отслеживать этот постоянно развивающийся спектр необязательных инструментов управления. Во-вторых, мы хотим перспективно оценить влияние ИИ на этические принципы, права человека, верховенство закона и демократию.

Глава 3 направлена на то, чтобы внести вклад в разработку будущего регулирования ИИ, опираясь на существующие обязательные документы, контекстуализируя их принципы и обеспечивая ключевые нормативные руководящие принципы для будущей правовой базы с целью сохранения гармонизации существующей правовой базы в области прав человека, демократии и верховенства права.

Национальные перспективы

В главе 1 излагается нынешнее положение дел в разработке политики Израиля в отношении возможностей и проблем, связанных с искусственным интеллектом (ИИ) в области прав человека и этики. Он в значительной степени основан на докладе Национальной инициативы Израиля по созданию защищенных интеллектуальных систем, который недавно был представлен израильскому правительству.

В главе 2 кратко описывается характер управления ИИ и особенности дискуссий в Японии. Поскольку этот отчет в основном посвящен попыткам управления и дискуссиям до 2018 года, легко представить, что детали, содержащиеся в нем, будут меняться с будущими технологическими разработками и изменениями в социальных условиях. Однако организация дискуссий в фиксированной точке полезна для будущих дискуссий по

управлению ИИ и для сравнительных исследований с другими передовыми технологиями управления.

В главе 3 представлены некоторые из наиболее распространенных и расширенных видов использования Приложений искусственного интеллекта (ИИ-приложений) в Мексике, а также нормативная база, применимая к ИИ-приложениям. Мы стремимся с высокой степенью точности представить контекст, в котором работает каждая система искусственного интеллекта. Контекст, а также краткое описание системы, как мы надеемся, предоставят читателю достаточно информации, чтобы провести сравнение с другими юрисдикциями и пролить некоторый свет на сложности, связанные с потенциальным регулированием применения ИИ в Мексике и в других странах.

Международная перспектива регулирования систем искусственного интеллекта на основе стандартов совета Европы

I. Доклад о ходе работы Специального комитета по искусственному интеллекту (САНАИ)

11 сентября 2019 года Комитет министров утвердил круг ведения специального комитета по искусственному интеллекту (САНАИ), в соответствии с которым Комитету было поручено изучить на основе широких консультаций с участием многих заинтересованных сторон осуществимость и потенциальные элементы правовой основы для разработки и применения искусственного интеллекта на основе стандартов Совета Европы в области прав человека, демократии и верховенства права. Настоящий доклад представляется САНАИ Комитету министров и содержит подробную информацию о прогрессе, достигнутом на сегодняшний день, с учетом воздействия мер по борьбе с пандемией COVID-19. Он также включает в себя размышления о методах работы, синергии и взаимодополняемости с другими соответствующими

заинтересованными сторонами, а также предложения по дальнейшим действиям САНАИ с помощью дорожной карты.

В докладе подчеркивается, что Совет Европы должен сыграть решающую роль в обеспечении того, чтобы Искусственный интеллект (ИИ) соответствовал стандартам Организации в области прав человека, демократии и верховенства права. САНАИ сосредоточил свою работу на составлении соответствующих международных и национальных правовых рамок и этических руководящих принципов, а также на анализе рисков и возможностей, связанных с искусственным интеллектом, в частности их влияния на права человека, верховенство закона и демократию.

Предварительный анализ, проведенный до настоящего времени, подтверждает важность углубления размышлений о целесообразности и разработке элементов «горизонтальной», сквозной правовой базы для регулирования использования и воздействия приложений ИИ, которая опиралась бы на уникальный опыт Организации и работу, проводимую на «вертикальном», отраслевом уровне. Его очевидная актуальность также была подтверждена и усилена недавней пандемией COVID-19. Такая правовая база могла бы консолидировать существующие стандарты в этой области или разработать дополнительные стандарты, необходимые в цифровую эпоху. Он мог бы основываться на подходе к оценке воздействия на права человека и содержать практические оперативные механизмы. Наконец, он мог бы также послужить основой для других инициатив и инструментов, которые по-прежнему необходимы для всестороннего решения проблем, связанных с применением ИИ в соответствующих областях деятельности Совета Европы.

Работа, проводимая САНАИ, также предоставляет возможность внести вклад и дополнить другие международные инициативы в этой области (например, ОЭСР, Европейским союзом – в частности, Европейской комиссией – ЮНЕСКО и Организацией Объединенных Наций в целом) путем принятия конкретного документа, основанного на правах человека, в рамках глобального правового механизма регулирования цифровых технологий, области, в которой

Совет Европы может принести реальную добавленную стоимость. Тесная координация и взаимодействие с работой этих организаций в области искусственного интеллекта будут продолжаться на регулярной основе.

Наконец, отчет включает в себя рабочую версию на сегодняшний день оглавления будущего технико-экономического обоснования и четкую дорожную карту с указанием ключевых сроков процесса. Были рассмотрены новые методы работы, включая создание трех тематических рабочих групп, отвечающих за конкретные задачи, связанные с технико-экономическим обоснованием САНАИ, а именно подготовку предложений по конкретным аспектам политики, подготовку консультаций с участием многих заинтересованных сторон и разработку правовых рамок, которые должны обеспечить достижение существенного прогресса в выполнении мандата САНАИ в ближайшие месяцы и до декабря 2021 года.

II. Предыстория и мандат САНАИ

В течение нескольких лет Совет Европы оценивает и прогнозирует влияние цифровых технологий на права человека, демократию и верховенство права. Он разрабатывает соответствующие правовые инструменты и две стратегии управления Интернетом, чтобы обеспечить, чтобы наше общество и отдельные лица в полной мере пользовались преимуществами инновационной практики. Среди этого технологии, основанные на искусственном интеллекте (ИИ), выделяются среди традиционных компьютерных приложений своей автономностью. Все чаще используемые в широком спектре государственных и частных услуг, эти конкретные технологии открывают большие возможности для развития, но также поднимают важные и сложные этические, правовые, политические и экономические вопросы.

Совет Европы взял на себя обязательство определить их масштабы и последствия в большинстве своих специализированных областей деятельности, таких как правосудие, защита данных, равенство и недискриминация. Влияние ИИ на отдельных лиц и общество также рассматривалось с различных точек

зрения Комитетом министров, Парламентской ассамблеей и Комиссаром по правам человека. В настоящее время Организация применяет дополнительный подход к инициативам, реализованным на сегодняшний день, приняв сквозной стратегический подход, основанный на изучении возможности создания обязательных и необязательных рамок с целью обеспечения того, чтобы разработка и применение ИИ соответствовали европейским стандартам в области прав человека, демократии и верховенства права. Конференция высокого уровня, организованная финским председательством в Комитете министров и Совете Европы в Хельсинки (Финляндия) 26-27 февраля 2019 года, отметила важность и срочность политических мер реагирования на воздействие ИИ на права человека, демократию и верховенство права и придала необходимый импульс созданию Специального комитета по искусственному интеллекту (САНАИ), круг ведения которого был утвержден Комитетом министров 11 сентября 2019 года.

САНАИ отвечает за изучение осуществимости и потенциальных элементов правовой основы для разработки и применения ИИ на основе стандартов Совета Европы в области прав человека, демократии и верховенства права. В рамках его мандата будут обеспечены широкие консультации с участием многих заинтересованных сторон и тесная координация с другими международными организациями.

Экономические и другие выгоды, вытекающие из применения ИИ, а также важность ИИ для Европы, чтобы идти в ногу с остальным миром, относятся к числу вопросов, которые САНАИ мог бы рассмотреть в рамках своей работы.

III. Ход работы

На сегодняшний день председатель провел свое первое пленарное заседание 18, 19 и 20 ноября 2019, а второе пленарное заседание состоялось 6-8 июля 2020 года. На своем первом пленарном заседании Комитет избрал своего Председателя (Грегора Строжина, Словения), заместителя Председателя (Пегги

Вальке, Бельгия) и свое Бюро (в состав которого, помимо Председателя и заместителя Председателя, вошли пять членов от Эстонии, Франции, Германии, Италии и Швейцарии). На сегодняшний день Бюро провело три совещания, соответственно 19 ноября 2019 года, 23 и 24 января 2020 года и онлайн совещание 27 марта 2020 года. Комитет также назначил докладчика по вопросам гендерного равенства (Яна Новоградска, Словацкая Республика) на своем первом пленарном заседании.

На своем первом пленарном заседании САНАИ уточнил, что технико-экономическое обоснование должно включать в себя сопоставление национальных и международных правовых документов (как Совета Европы, так и других международных организаций) и этических рамок, связанных с применением ИИ, а также сопоставление рисков и возможностей, возникающих в связи с разработкой, проектированием и применением ИИ, включая его влияние на права человека, демократию и верховенство права. Выводы, полученные в результате сопоставления, должны быть рассмотрены для принятия решения о целесообразности определения ИИ и для определения подходящей правовой основы для разработки и применения ИИ на основе стандартов Совета Европы. Трем независимым экспертам было поручено подготовить технико-экономическое обоснование с целью оказания поддержки САНАИ в принятии решений по этому вопросу, и они представили свои доклады САНАИ на втором пленарном заседании.

Что касается составления карт международно-правовых документов, применимых к ИИ, то выбранный подход заключается в выявлении во многих секторах деятельности Совета Европы и при поддержке соответствующих межправительственных комитетов соответствующих международных документов, руководящих правовых принципов и основных ценностей, вытекающих из этих документов, а также их потенциальных пробелов. В частности, следует рассмотреть возможность появления общих специфичных для ИИ «горизонтальных» и «поперечных» принципов, которые являются

основой для специализированных «вертикальных» принципов, определенных (или определяемых) в различных областях деятельности Совета Европы.

Результаты онлайн – консультаций, проведенных между государствами-членами Совета Европы в ноябре 2019 года – феврале 2020 года, используются для составления карт национальных инициатив в эфире.

В картографировании этических и необязательных рамок анализируется обширная разработка хартий, деклараций и принципов в области ИИ, разработанных в последние годы частными, научными и гражданскими организациями, с целью выявления тех, которые могли бы способствовать созданию возможной правовой базы в области ИИ в соответствии со стандартами Совета Европы.

Картирование рисков и возможностей, возникающих в результате разработки и применения ИИ, включая воздействие ИИ на права человека, демократию и верховенство права, проводится конкретно в связи с правами, защищаемыми Конвенцией о защите прав человека и основных свобод (далее: Европейская конвенция о правах человека), и направлено на определение того, в какой степени на их осуществление могут повлиять определенные виды применения ИИ и какие стратегии следует принять для решения этой проблемы. Ответы, полученные в рамках онлайн-консультаций, могут также оказать некоторую помощь в отношении стратегий, действующих в государствах-членах для смягчения воздействия ИИ на права, защищаемые Конвенцией.

Эти сопоставления должны быть направлены на определение применимых правовых рамок в области прав человека, которые важны в цифровую эпоху. Тем не менее, проведенный до сих пор предварительный анализ указывает, с одной стороны, на ограничения существующих правовых режимов, появившихся до разработки приложений ИИ, которые снижают их эффективность в адекватном реагировании на текущие вызовы; и, с другой стороны, на значительное влияние ИИ на ряд стандартов, определенных в документах Совета Европы, и в частности на права, защищаемые Европейской

конвенцией о правах человека. В этой связи следует продолжить изучение вопроса о роли и обязанностях всех соответствующих заинтересованных сторон, включая государства, предприятия, научно-исследовательские учреждения и гражданское общество, в разработке и внедрении технологий искусственного интеллекта, обеспечивающих уважение прав человека.

Столкнувшись с этими проблемами, Совет Европы должен сыграть ключевую роль в обеспечении того, чтобы разработка, проектирование и применение технологий искусственного интеллекта соответствовали стандартам Организации. Поэтому важно углубить размышления о разработке элементов правовой базы для регулирования использования и воздействия приложений ИИ, опираясь на уникальный опыт Организации и основываясь на подходе к оценке воздействия на права человека. Кроме того, в этой правовой базе можно было бы рассмотреть целесообразность консолидации существующих стандартов в этой области (таких как стандарты Европейской конвенции по правам человека или Конвенции 108+) или разработки дополнительных стандартов, необходимых в цифровую эпоху. Наконец, он мог бы также послужить основой для ряда отраслевых инициатив и инструментов, которые по-прежнему необходимы для комплексного решения проблем, связанных с применением ИИ в соответствующих областях деятельности Совета Европы.

Это также возможность поддержать другие международные инициативы (ОЭСР, Европейский союз – в частности, Европейская комиссия – ЮНЕСКО и Организация Объединенных Наций в целом) путем принятия конкретного документа, основанного на правах человека, в рамках глобального правового механизма регулирования цифровых технологий, в котором Совет Европы имеет реальную добавленную стоимость.

Кроме того, было бы важно также изучить в рамках технико-экономического обоснования механизмы и меры, которые могут быть приняты для обеспечения эффективности предлагаемой правовой базы как на уровне Совета Европы, так и на уровне государств-членов. Это будет включать, в

частности, вопросы, касающиеся создания механизмов совместного права – создание (в частности, использование в некоторых конкретных случаях «песочницы») и предварительная проверка и/или сертификация приложений ИИ или механизмов контроля независимыми органами, а также целесообразность регулирования профессий экспертов по мегаданным (например, возможность создания профессионального заказа для специалистов по данным, разработка деонтологических хартий, основанных на этических принципах, «Клятва Гиппократ» для специалистов по ИИ или создание «Водительских прав» для ИИ и т.д.).

В процессе подготовки технико-экономического обоснования, в том числе в рамках деятельности рабочих групп и в рамках консультаций с участием многих заинтересованных сторон, САНАИ будет стремиться интегрировать перспективу гендерного равенства, способствовать созданию сплоченных обществ, поощрению и защите прав инвалидов.

В ходе второго пленарного заседания САНАИ государства-члены высказали свои мнения по содержанию различных пунктов проекта оглавления технико-экономического обоснования, которое включено в Приложение I к настоящему докладу.

IV. Влияние пандемии COVID-19 на деятельность САНАИ

ИИ является одним из многих инструментов, используемых государствами для сдерживания и борьбы с пандемией COVID-19. Было разработано несколько применений ИИ в таких областях, как поддержка исследований по разработке вакцин, формулирование диагностики для поддержки медицинского персонала и разработка прогнозных моделей для возможной эволюции пандемии. ИИ использовался для облегчения анализа тысяч научных работ, опубликованных по пандемии, и для обеспечения более эффективного обмена научными знаниями, а также для целей скрининга заболеваний. Кроме того, ИИ может быть полезен для борьбы с

дезинформацией о коронавирусе, если права на защиту данных и свободу слова должным образом защищены.

Были также определены другие виды использования, направленные на контроль эпидемических рисков: например, распознавание лиц и биометрические устройства, устройства геолокации и беспилотные летательные аппараты использовались для обеспечения соблюдения мер сдерживания лицами, находящимися в группе риска, или инфицированными лицами или группами населения. Также были разработаны приложения, предупреждающие пользователей о том, что они контактировали с людьми, потенциально несущими вирус. Этот вид использования оказывает значительное влияние на некоторые права и свободы, защищаемые международными договорами по правам человека, включая Европейскую конвенцию о правах человека, такие как право на уважение частной и семейной жизни (в частности, право на неприкосновенность частной жизни) и свободу передвижения, и должен быть рассмотрен САНАИ. Существующее или планируемое использование приложений ИИ в некоторых странах для мониторинга и борьбы с эпидемиями, а также для предотвращения новых вспышек заболеваний может привести к упрощению использования массового эпиднадзора за населением в отсутствие адекватных гарантий.

Меры по блокированию COVID-19 привели к непреднамеренной дискриминации детей и молодых людей, которые не имеют доступа к ноутбукам и не посещают школы, предлагающие онлайн-курсы. Это привело к перерыву в их образовании и препятствию для развития их цифровых навыков, включая искусственный интеллект как межсекторальную область изучения. Во времена экономического кризиса, который может возникнуть в результате кризиса COVID-19, существует историческая модель дискриминации девочек и молодых женщин, когда семьи принимают решения о том, инвестировать ли средства в высшее или техническое образование детей. Девочки и молодые женщины часто находятся в неблагоприятном положении, и они склоняются в пользу мальчиков и молодых мужчин. Это еще больше усугубит статус-кво

недопредставленности женщин в ролях по развитию ИИ и связанных с ИИ ролях принятия решений. САНАІ через свою Группу по разработке политики (САНАІ-PDG) могла бы обратить внимание на этот вопрос и сосредоточиться на мерах, направленных на смягчение этой исторической модели дискриминации девочек и молодых женщин во время экономического кризиса. Можно было бы рассмотреть вопрос о применении временных позитивных мер по учету гендерной проблематики в рамках политики ИИ для защиты и поддержки женщин и молодых девушек, с тем чтобы они не оставались в стороне и имели доступ к роли принятия решений в экосистеме ИИ.

В этом контексте повышается актуальность правовой базы по разработке, разработке и применению ИИ, основанной на принципах и ценностях Совета Европы. САНАІ, возможно, пожелает рассмотреть последствия и, в соответствующих случаях, необходимость создания правовой базы для учета таких ситуаций, как использование приложений искусственного интеллекта в конкретных ситуациях, как это имеет место в случае других договоров Совета Европы, и в соответствии с принципами и ценностями Совета Европы.

V. Методы работы, документы и дорожная карта

На своем первом пленарном заседании председатель поручил своему Бюро в соответствии с резолюцией CM/Res(2011)241 выполнить задачу по составлению необходимых карт для технико-экономического обоснования. САНАІ, приняв к сведению юридическое заключение Юрисконсультанта Совета Европы, также санкционировал специальное участие других государств-членов в работе Бюро в соответствии с условиями, определенными САНАІ и его Бюро.

В целях обеспечения более активного участия государств-членов в САНАІ и его транспарентного, и сбалансированного подхода к принятию решений САНАІ провел обмен мнениями по предложению Российской Федерации об увеличении числа членов Бюро на шесть, в общей сложности до тринадцати. Было предложено, чтобы новые члены могли участвовать в заседаниях Бюро без покрытия их расходов Советом Европы.

После переноса совещания, запланированного на 11-13 марта 2020 года в связи с пандемией COVID-19, Бюро обсудило конкретные предложения по адаптации методов работы, с тем, чтобы САНАИ могла продолжать свою работу эффективно и в соответствии с поставленными целями, если только это не окажется сложным, и до следующего пленарного заседания, в частности, в условиях, которые гарантировали бы безопасность и участие членов и наблюдателей. Бюро САНАИ поручило Секретариату организовать письменные консультации членов САНАИ по ряду рабочих документов и вопросов, включая принятие настоящего доклада о ходе работы и рассмотрение новых методов работы САНАИ, таких как создание небольших рабочих групп. На основе этих консультаций члены САНАИ договорились создать три рабочие группы, занимающиеся соответственно разработкой политики, правовой базой, консультациями и информационно-пропагандистской деятельностью (в связи с консультацией с участием многих заинтересованных сторон), круг ведения которых содержится в Приложении III к настоящему документу. Кроме того, на своем втором пленарном заседании САНАИ принял документ о методах работы и условиях функционирования рабочих групп.

Как отмечалось выше, технико-экономическое обоснование и потенциальные элементы будущей правовой базы должны основываться на широких консультациях с участием многих заинтересованных сторон. В этой связи члены САНАИ, представив письменные замечания в ходе первой онлайн-консультации после первого пленарного заседания, указали, что этот вопрос следует в приоритетном порядке адресовать представителям частного сектора, гражданского общества и научного сообщества, а также другим международным организациям, включая организации, устанавливающие технические стандарты. Дорожная карта САНАИ предусматривает, что предварительный проект технико-экономического обоснования, включающий основные элементы будущей правовой базы, будет рассмотрен САНАИ на его третьем пленарном заседании (30 ноября-2 декабря 2020 года). Это обеспечит первый результат, который может быть обсужден в открытой и транспарентной

форме различными заинтересованными сторонами и впоследствии станет основой для консультаций с участием многих заинтересованных сторон, которые начнутся в 2021 году. 2021 год также, возможно, будет посвящен доработке элементов вышеупомянутой правовой базы.

Синергия и взаимодополняемость работы САНАИ с работой других международных организаций

Активно предпринимались усилия по координации с существующей или текущей работой по ИИ в других международных организациях, в частности в Европейском союзе (например, в рамках текущего процесса, инициированного Европейской комиссией по Белой книге по ИИ) и ОЭСР, в целях содействия синергизму и избежания дублирования. В дополнение к участию в совещаниях САНАИ и регулярным контактам между учреждениями Председатель САНАИ и Секретариат приняли участие в запуске Обсерватории политики ОЭСР в области ИИ 27 февраля 2020 года. Кроме того, в ходе пленарных заседаний САНАИ планируется регулярно проводить обмен мнениями с представителями ОЭСР, Европейского союза, МСЭ, ИСО и ЮНЕСКО о текущей работе в рамках каждой организации. Цель состоит в том, чтобы содействовать синергизму в разработке правовой базы по ИИ, в которой конкретный вклад и опыт каждой организации (в случае САНАИ, опыт в области прав человека, демократии и верховенства права, в частности) могут быть выделены и дополнять друг друга.

VI. Бюджетные последствия

Бюджет Председателя, утвержденный Комитетом министров, охватывает только проведение пленарных заседаний и заседаний Бюро и частично проведение ограниченного числа заседаний рабочей группы. Выделенный бюджет не предусматривает дополнительных мер, таких как усиление аспекта консультаций с участием многих заинтересованных сторон, вовлечение других экспертов высокого уровня и/или организация мероприятий, которые могут повысить значимость работы Совета Европы в этой области.

Поэтому государствам-членам и государствам-наблюдателям предлагается рассмотреть вопрос о внесении финансового вклада в укрепление деятельности САНАІ посредством добровольных взносов или любых других средств, которые они сочтут целесообразными.

VIII. Выводы и конкретные практические предложения

Сегодня Совет Европы должен сыграть решающую роль в обеспечении того, чтобы приложения ИИ соответствовали защите прав человека и соответствовали стандартам Организации, когда они существуют. Таким образом, инициатива САНАІ остается уникальной и дополняет инициативы, предпринимаемые другими международными организациями. Важно, чтобы работа по разработке потенциальных элементов правовой базы могла начаться уже в январе 2021 года и чтобы САНАІ были предоставлены дополнительные финансовые средства для осуществления своего мандата.

Комитету министров предлагается поручить председателю предпринять следующие действия: добиться существенного прогресса в разработке технико-экономического обоснования правовой основы к ноябрю 2020 года с целью начала в январе 2021 года анализа элементов правовой основы, которые станут предметом широких консультаций с участием многих заинтересованных сторон; эта правовая основа могла бы регулировать разработку и применение ИИ, оказывающих значительное влияние на права человека, демократию и верховенство права. Она могла бы также рассмотреть вопрос о желательности консолидации существующих стандартов путем толкования норм, принципов и ценностей, уже принятых в этой области, или разработки новых стандартов, необходимых для цифровой эпохи. Наконец, это заложило бы основу для дальнейшего развития ряда инициатив и инструментов в различных секторах деятельности Совета Европы, что по-прежнему необходимо для комплексного решения проблем, связанных с применением ИИ в соответствующих областях деятельности Совета Европы.

Одновременно предложить дополнительные меры по введению в действие вышеупомянутой правовой базы: в частности, можно было бы сослаться на предшествующую процедуру оценки воздействия на права человека, средства проверки или сертификации алгоритмов и систем искусственного интеллекта или подготовку и организацию определенных профессий, связанных с применением инструментов искусственного интеллекта.

Комитету министров также предлагается принять к сведению дорожную карту САНАІ и предложить государствам-членам или государствам-наблюдателям, которые желают этого, внести финансовый вклад в укрепление деятельности САНАІ посредством добровольных взносов или любых других средств, которые они сочтут полезными.

Приложение I. Проект содержания технико-экономического обоснования (рабочая версия: 16 июня 2020 года)

Общее введение

Сфера применения правовой базы Совета Европы в области искусственного интеллекта

Возможности и риски, связанные с разработкой, развитием и применением искусственного интеллекта в области прав человека, верховенства права и демократии. «Зеленая» и «красная» области» – означают соответственно позитивные и проблемные примеры применения искусственного интеллекта с точки зрения прав человека, верховенства права и демократии, с учетом контекстно-зависимой среды для проектирования, разработки и применения искусственного интеллекта в Европе и событий на глобальном уровне.

Работа Совета Европы по искусственному интеллекту на сегодняшний день

Отображение инструментов, применимых к искусственному интеллекту

Международно-правовые документы, применимые к искусственному интеллекту

Этические принципы, применимые к искусственному интеллекту

Обзор национальных инструментов, политики и стратегий, связанных с искусственным интеллектом

Преимущества, недостатки и ограничения существующих международных и национальных документов и этических руководящих принципов в области искусственного интеллекта

Международные правовые документы, этические принципы и частные субъекты

Основные выводы консультаций с участием многих заинтересованных сторон

План технико-экономического обоснования

Основные выводы о типе и содержании правовой основы для разработки, разработки и применения искусственного интеллекта, основанной на стандартах Совета Европы в области прав человека, демократии и верховенства права Основные элементы правовой основы для разработки, разработки и применения искусственного интеллекта

III. Ключевые ценности, права и принципы, вытекающие – в восходящей перспективе – из отраслевых подходов и этических принципов; в нисходящей перспективе – из прав человека, демократии и требований верховенства права.

Роль и обязанности государств-членов и частных субъектов в разработке приложений, соответствующих таким требованиям.

Ответственность за ущерб, причиненный искусственным интеллектом.

8. Возможные варианты правовой базы Совета Европы по разработке, развитию и применению искусственного интеллекта на основе прав человека, верховенства права и демократии (для каждого варианта: содержание, адресаты, добавленная стоимость, роль частных субъектов, ожидания государств-членов, вытекающие из представленных письменных комментариев).

I. Обновление существующих юридически обязательных документов

Соглашение;

Основополагающая конвенция;

Инструмент(ы) мягкого права

Другие виды поддержки государств-членов, такие как выявление передовой практики.

Возможная взаимодополняемость между горизонтальными и сквозными элементами, которые могли бы стать частью инструмента обычного типа, и вертикальной и секторальной работой, которая могла бы привести к созданию конкретных инструментов иного характера.

Возможные практические механизмы обеспечения соблюдения и эффективности правовой базы (например, создание механизма предварительной проверки и/или сертификации, надзор со стороны независимых органов, песочница и т.д.).

Приложение III. Круг ведения Рабочих групп

В состав САНАИ-PDG входят до 20 экспертов, назначенных государствами-членами, а также участники и наблюдатели, которые выражают заинтересованность в участии в этой рабочей группе и вносят профессиональный и постоянный вклад в ее работу.

Задачи САНАИ – PDG должен:

1) Внести вклад в разработку технико-экономического обоснования правовой основы применения искусственного интеллекта, опираясь на картографическую работу, уже проделанную САНАИ, и уделяя особое внимание, в частности, следующим вопросам:

а. Выявлять и анализировать, как применение ИИ государствами и частными субъектами влияет на права человека, верховенство права и демократию, включая риски, угрозы и возможности, создаваемые или создаваемые ИИ, в том числе, в соответствующих случаях, путем оценки предложений по регулированию ИИ, внесенных государствами-членами и

другими заинтересованными сторонами; и принимая во внимание важность того, чтобы идти в ногу с развитием ИИ во всем мире;

b. На основе выводов картографирования подготовить ключевые выводы и предложения по политике и другим мерам, обеспечить актуальность и эффективность международных стандартов и международно-правовых документов в этой области, а также подготовить предложения по конкретному правовому документу, регулирующему искусственный интеллект;

c. Выявлять приложения ИИ с высоким риском и высокими возможностями, рассматривать и разрабатывать политику и другие меры по устранению рисков, связанных с ними, и обеспечивать защиту прав человека, верховенства права и демократии;

d. В связи с работой, проводимой в соответствии с пунктами b. и c. подготовить предложения по содействию осуществлению соответствующих международных и европейских обязательств в области прав человека и направленные на поддержку осуществления эффективных оценок рисков в области прав человека и любых других необходимых действий для реагирования по мере необходимости на значительные новые и возникающие возможности, угрозы и риски.

2) Разработать предложения по взаимодействию и консультированию с соответствующими заинтересованными сторонами по вышеупомянутым элементам в тесной координации и консультациях с Консультативной и информационно-пропагандистской группой (CAHAI-COG).

Методы работы

CAHAI-PDG проведет 2 заседания в 2020 году и, при условии принятия пленарного решения на этот счет, любые другие дополнительные заседания, необходимые в 2021 году. Он будет проводить свою работу с помощью технологий, таких как электронная почта или другие аналогичные средства электронной связи, виртуальные встречи и/или физические встречи. Группа назначает Председателя и Сопредседателя из числа своих членов на своем первом совещании.

Бюджет Совета Европы будет покрывать расходы на участие председателей (для физических совещаний) и расходы на организацию совещаний с помощью технических средств, с тем чтобы рабочая группа могла выполнять свои функции и обязанности посредством онлайн-обменов.

Участие членов в физических собраниях в принципе будет осуществляться направляющими учреждениями и организациями без покрытия расходов.

Июль 2020-Декабрь 2021

Консультативно-пропагандистская группа (САНАІ-COG)

Структура

В состав САНАІ – COG входят до 20 экспертов, назначенных государствами-членами, а также участники и наблюдатели, которые выражают заинтересованность в участии в этой рабочей группе и вносят профессиональный и постоянный вклад в ее работу.

ЗАДАЧИ

САНАІ – COG должна подвести итоги анализа, проведенного секретариатом ответов на первую онлайн-консультацию, и разработать для рассмотрения и утверждения САНАІ:

1) Анализ и картирование заинтересованных сторон (срок проведения: сентябрь 2020 года);

2) На основе информационно-пропагандистской работы со странами, выразившими заинтересованность в проведении консультаций в странах, методология и шаблон(ы) для использования государствами-членами в ходе консультаций в странах (срок: ноябрь 2020 года);

3) Расписание тематических консультаций (срок: ноябрь 2020 года) и первоначальный консультативный документ (срок: декабрь 2020 года);

4) Анализ текущих событий и отчетов, которые имеют непосредственное отношение к задачам рабочих групп САНАІ (срок: текущий – октябрь 2020

года), а также анализ вклада респондентов в онлайн-консультации для рассмотрения САНАИ (срок: 2021 год).

Методы работы

Ожидается, что САНАИ-COG будет работать в основном в режиме онлайн и будет поддерживаться секретариатом Председателя. Группа назначает Председателя и Сопредседателя из числа своих членов на своем первом совещании.

Продолжительность Июль 2020-Январь 2021.

Группа по правовым основам (САНАИ – LFG)

Структура

Продолжительность

В состав САНАИ – LFG входят до 20 экспертов, назначенных государствами-членами, а также участники и наблюдатели, которые выражают заинтересованность в участии в этой рабочей группе и в существенном вкладе в ее работу.

Задачи

САНАИ – LFG должна:

1) Подготовить основные выводы и предложения по возможным элементам и положениям правовой базы с целью разработки, в соответствующих случаях, проекта правового документа (документов) для рассмотрения и утверждения САНАИ с учетом обзора существующих правовых документов, применимых к искусственному интеллекту, и вариантов политики, изложенных в технико-экономическом обосновании, утвержденном САНАИ;

2) Разработать конкретные предложения по регулированию разработки, проектирования и применения ИИ в областях, определенных государствами-членами и другими заинтересованными сторонами как рискованные, с учетом подходов государств-членов к регулированию.

Методы работы

В 2021 году фонд САНАІ проведет 2 заседания, и при условии принятия пленарного решения на этот счет любые другие дополнительные заседания по мере необходимости. Он будет проводить свою работу с помощью технологий, таких как электронная почта или другие аналогичные средства электронной связи, виртуальные встречи и/или физические встречи. Группа назначает Председателя и Сопредседателя из числа своих членов на своем первом совещании.

Бюджет Совета Европы будет покрывать расходы на участие председателей (для физических совещаний) и расходы на организацию совещаний с помощью технических средств, с тем чтобы рабочая группа могла выполнять свои функции и обязанности посредством онлайн-обменов.

Участие членов в физических собраниях в принципе будет осуществляться направляющими учреждениями и организациями без покрытия расходов.

Продолжительность Январь 2021 – Декабрь 2021

Влияние ИИ на права человека, демократию и верховенство права

Введение

ИИ, как технология общего назначения, оказывает влияние на всю структуру общества. В 2017 году Европейский экономический и социальный комитет в том, что широко считается «начальным докладом» о более широком социальном воздействии ИИ, определил наиболее важные области социального воздействия, включая: безопасность; этику; законы и регулирование; демократию; прозрачность; конфиденциальность; работу; образование и равенство. Это означает, что ИИ оказывает влияние на наши права человека, демократию и верховенство закона – основные элементы, на которых строятся наши европейские общества.

В 2019 году Группа экспертов ИИ представила Руководящие принципы этики для ИИ.¹ Эти руководящие принципы определяют надежный ИИ как законный, этичный и социально – технически надежный. Что касается этического элемента заслуживающего доверия ИИ, то в руководящих принципах прямо рассматриваются основные права в качестве основы этики ИИ.² Хотя эти руководящие принципы содержат элементы, которые непосредственно вытекают из существующих (прав человека), сами по себе они еще не являются юридически обязательными. В последнее время все громче звучит призыв к более сильным (существующим или новым) юридически обязательным инструментам для ИИ. Европейская комиссия объявила о потенциальных элементах законодательной базы в своей Белой книге по ИИ³ и подчеркивает важность того, чтобы ИИ соответствовал основным правам ЕС и законам, которые обеспечивают эти права.

В этом документе описывается влияние ИИ на права человека, демократию и верховенство права. В нем определяются те права человека, закрепленные в Европейской конвенции о правах человека («ЕСПЧ»), Протоколах к ней и Европейской социальной хартии («ЕСК»), которые в настоящее время в наибольшей степени затронуты или, вероятно, будут затронуты ИИ (Глава II). В главах III и IV он стремится представить ряд возможных стратегий, которые при необходимости могут быть реализованы одновременно. В главе III рассматривается вопрос о воздействии в рамках существующих рамок прав человека, демократии и верховенства права. В главе IV рассматриваются стратегии на случай, если существующие рамки не смогут должным образом защитить нас. Поскольку технологии и общество быстро развиваются, этот документ не может быть исчерпывающим, но определяет

1 Руководящие принципы этики для надежного ИИ, Экспертная группа высокого уровня по ИИ при Европейской комиссии, 2019 г.

2 Далее в руководстве описываются 7 требований для надежного (т.е. законного, этичного и надежного) ИИ: 1) Человеческое управление и надзор, 2) Техническая надежность и безопасность, 3) Конфиденциальность и управление данными, 4) Прозрачность, 5) Многообразие, недискриминация и справедливость, 6) Социальное и экологическое благополучие, 7) Подотчетность.

3 Whitepaper по искусственному интеллекту, CQM(2020)

приоритеты наиболее релевантных воздействий в той степени, в какой они могут быть определены сегодня.

Понятие ИИ

ИИ имеет множество приложений, которые уже были внедрены в общество: биометрическое (включая распознавание лиц) распознавание объектов, прогнозирование рисков и успеха, алгоритмическое принятие или поддержка решений, автоматический перевод, рекомендательные системы и так далее. Эти приложения нашли свое применение в таких секторах, как правоохранительные органы, правосудие, управление человеческими ресурсами, финансовые услуги, транспорт, здравоохранение, общественные услуги и т.д. ИИ остается по существу спорной концепцией, поскольку не существует общепринятого определения. Тем не менее, определения в целом можно разделить на два лагеря: рационалистические и ориентированные на человека определения. Наиболее известное рационалистическое определение определяет ИИ как «агент, созданный людьми, который принимает решения и выполняет действия, основанные на его восприятии».⁴ Наиболее известным примером определения, ориентированного на человека, является тест Тьюринга, который проходит компьютер, как только он выполняет задачу, которая в противном случае потребовала бы человеческого интеллекта. Группа экспертов по ИИ представила определение ИИ в 2019 году⁵: Системы искусственного интеллекта (ИИ) – это программные (и, возможно, также аппаратные) системы, разработанные людьми, которые, учитывая сложную цель, действуют в физическом или цифровом измерении, воспринимая окружающую среду посредством сбора данных, интерпретируя собранные структурированные или неструктурированные данные, рассуждая о знаниях или обрабатывая информацию, полученную из этих данных, и принимая

4 Рассел, С. Дж., Норвиг, П., и Дэвис, Е. (2010). Искусственный интеллект: Современный подход (3-е изд.). Prentice Hall.

5 Экспертная группа высокого уровня ЕС по ИИ. Определение ИИ, основные возможности и научные дисциплины, 2019.

решение о наилучших действиях для достижения данной цели. Системы искусственного интеллекта могут либо использовать символические правила, либо изучать числовую модель, а также адаптировать свое поведение, анализируя, как на окружающую среду влияют их предыдущие действия.

Часто ИИ описывается как совокупность технологий, сочетающих данные, алгоритмы и вычислительные мощности. Хотя это верно для наиболее широко используемых в настоящее время систем ИИ, это лишь очень ограниченное описание того, что такое ИИ. ИИ – это термин – контейнер для многих компьютерных приложений, некоторые из которых объединяют данные и алгоритмы, но существуют и другие, не основанные на данных подходы к ИИ, например экспертные системы, рассуждения и представления знаний, реактивное планирование, аргументация и другие. Большинство систем ИИ, которые в последнее время проникают в наши общества, действительно являются примерами ИИ, управляемого данными, с особым влиянием на демократию в области прав человека и верховенство закона. В нижеследующем определяются наиболее релевантные термины:

Узконаправленный ИИ: системы ИИ, которые могут выполнять только очень специфические «узкие» задачи;

Большие Данные: системы искусственного интеллекта, которым для хорошей работы требуется много данных. Качество, объем и содержание данных влияют на работу этих систем искусственного интеллекта и часто приводят к повторению и усилению ошибок, пробелов и искажений в данных;

Корреляция: многие системы ИИ только ищут связи в данных, которые не устанавливают (или «видят») причинно-следственную связь между «случаем» и решением, а просто делают прогноз, основанный на общих характеристиках с другими «случаями»;

Черные ящики: многие системы искусственного интеллекта часто являются так называемыми черными ящиками, в которых происходят процессы (принятия решений), которые не могут быть полностью объяснены человеческими терминами;

Здравый смысл: системы искусственного интеллекта не обладают здравым смыслом, что означает, что, хотя система может распознать кошку или раковую клетку, она не имеет представления о том, что такое кошка или раковая клетка. Он просто предоставляет ярлык для определенного шаблона. Он также не может использовать информацию о кошке или раковой клетке для идентификации собаки или головной боли.

Все эти характеристики могут сделать текущий ИИ уязвимым, нестабильным и непредсказуемым, но также востребованным и широко применяемым.

Самое главное, системы ИИ – это нечто большее, чем просто сумма их программных компонентов. Системы ИИ также включают в себя окружающую их социотехническую систему. При рассмотрении вопросов управления следует сосредоточиться не только на технологии, но и на окружающих ее социальных структурах: организациях, людях и учреждениях, которые ее создают, разрабатывают, внедряют, используют и контролируют, а также на людях, на которых она влияет, таких как граждане в их отношении к правительствам, потребители, работники или даже все общество.

Определение ИИ для целей регулирования.

Осложняющим фактором является то, что юридические определения отличаются от чисто научных, поскольку они должны отвечать ряду требований⁶ (таких как всеохватность, точность, полнота, практичность, постоянство), некоторые из которых являются юридически обязательными, а некоторые считаются хорошей нормативной практикой.⁷

6 Юридическое определение ИИ – Йонас Шютт Университет Гете Франкфурт-на-Майне 4 сентября 2019 г. (Юридические определения должны быть: (I) инклюзивными: цели регулирования не должны быть чрезмерно или недостаточно инклюзивными. (Джулия Блэк. Правила и регуляторы. Oxford University Press, 1997. [32] Robert Baldwin, Martin Cave, and Martin Lodge. Understanding Regulation: Теория, стратегия и практика. Oxford University Press, end edition, 2012.); (II) точным: должно быть ясно, какое дело подпадает под определение, а какое нет; (III) всеобъемлющим: определение должно быть понятно тем, кто регулирует; (IV) практичным: специалисты в области права должны иметь возможность легко определить, подпадает ли дело под определение; (V) постоянным: следует избегать необходимости постоянного обновления законодательства.

7 Всеохватность может быть выведена из принципа пропорциональности в праве ЕС (ст. 5(4) Договора о Европейском Союзе. Критерии точности и всесторонности основаны на принципе правовой определенности в праве ЕС. Критерии практичности и постоянства считаются хорошей законодательной практикой.

Влияние искусственного интеллекта на права человека, демократию и верховенство закона.

Применение подхода «жизненного цикла ИИ» важно для того, чтобы учитывать не только стадию разработки ИИ, но и стадии внедрения и использования. Еще один элемент, о котором следует помнить, заключается в том, что большинство используемых в настоящее время приложений ИИ могут закреплять, усугублять и усиливать воздействие на права человека, демократию и верховенство закона в масштабе, затрагивая более широкие слои общества и большее количество людей одновременно.

Четыре «основы прав человека» в рамках ЕКПЧ, протоколов к ней и ЕКПЧ подвергаются воздействию ИИ:

Уважение человеческого достоинства, свобода личности, равенство, недискриминация и солидарность, социальные и экономические права.

Кроме того, ИИ оказывает значительное влияние на: демократию, верховенство закона.

Важно отметить, что многие системы ИИ или их использование могут одновременно влиять на различные права человека, демократию и верховенство закона, или негативно влиять на права человека одного человека, но при этом позитивно влиять на права другого.

ИИ и уважение к человеческим ценностям

Уважение к человеческой ценности отражено в ЕКПЧ в различных правах, таких как право на свободу и безопасность (статья 5), право на справедливое судебное разбирательство (статья 6), право на отсутствие наказания без закона (статья 7) и право на частную жизнь и физическую и психическую неприкосновенность (статья 8).

Свобода и безопасность, справедливое судебное разбирательство, отсутствие наказания без закона (ст. 5, 6, 7 ЕКПЧ)

Тот факт, что ИИ может увековечить или усилить существующие предубеждения, особенно актуален при использовании в правоохранительных и

судебных органах. В ситуациях, когда на карту поставлена физическая свобода или личная безопасность, например, при предиктивной работе полиции, определении риска рецидивизма и вынесении приговора, право на свободу и безопасность в сочетании с правом на справедливый суд оказываются уязвимыми. Когда ИИ-система используется для прогнозирования рецидивизма или вынесения приговора, она может иметь необъективные результаты. Когда это «черный ящик», для специалистов в области права, таких как судьи, адвокаты и окружные прокуроры, становится невозможным понять обоснование результатов работы системы, что усложняет мотивировку и обжалование приговора.

Менее очевидным является влияние ИИ на право на обоснованное подозрение и запрет произвольного ареста. Приложения ИИ, используемые для предиктивной работы полиции, просто ищут взаимосвязи на основе общих характеристик с другими «случаями». Подозрение в таких случаях основано не на фактическом подозрении в совершении преступления или проступка конкретным подозреваемым, а лишь на общих характеристиках подозреваемого с другими лицами (таких как адрес, доход, гражданство, долги, занятость, поведение, поведение друзей и членов семьи и так далее). Более того, фактические характеристики, используемые в ИИ-системе, и «веса», придаваемые этим характеристикам, часто неизвестны.

Однако при ответственном подходе некоторые виды или способы использования ИИ могут также повысить безопасность, например, приложения ИИ, которые могут «состарить» пропавших людей на фотографиях, чтобы повысить шансы на их поиск, или распознавание объектов на основе ИИ, которое может сканировать багаж в аэропорту на предмет подозрительного содержимого.

Частная и семейная жизнь, физическая, психологическая и моральная неприкосновенность (статья 8 ЕКПЧ)

Многие ИИ-системы и способы их использования оказывают широкое и глубокое влияние на право на неприкосновенность частной жизни. Дискуссии о

неприкосновенности частной жизни вокруг ИИ в настоящее время, как правило, сосредоточены в основном на конфиденциальности данных и неизбирательной обработке персональных (и неперсональных) данных. Однако следует отметить, что, хотя конфиденциальность данных действительно является важным элементом, влияние ИИ на нашу частную жизнь выходит далеко за рамки наших данных. Ст. 8 ЕКПЧ включает в себя защиту широкого спектра элементов нашей частной жизни, которые можно сгруппировать в три широкие категории, а именно: (I) частная (общая) жизнь человека, (II) физическая, психологическая или моральная неприкосновенность человека и (III) личность и автономия человека.⁸ Различные виды применения и использования ИИ могут оказывать влияние на эти категории, и до сих пор им уделялось мало внимания.

Массовое наблюдение с помощью ИИ, например, с помощью распознавания лиц, включает в себя сбор, хранение и обработку персональных (биометрических) данных (наши лица)⁹, но оно также затрагивает нашу «общую» частную жизнь, идентичность и автономию таким образом, что создает ситуацию, когда за нами (постоянно) наблюдают, следят и идентифицируют. В качестве психологического «охлаждающего» эффекта, люди могут почувствовать склонность адаптировать свое поведение к определенной норме, что смещает баланс сил между государством или частной организацией, использующей распознавание лиц, и человеком.¹⁰ В правовой доктрине и прецедентах охлаждающий эффект слежки может представлять собой нарушение личного пространства, необходимого для развития личности и демократического обсуждения.¹¹ Даже если наши лица немедленно удаляются

8 Руководство к ст. 8 ЕКПЧ, Совет Европы.

9 Практика Европейского суда по правам человека (ЕСПЧ) ясно показывает, что сбор, хранение и обработка такой информации, даже кратковременная, влияет на ст. 8 ЕКПЧ.

10 Исследованные жизни: Информационная приватность и субъект как объект, Джули Э. Коэн, 2000.

11 Охлаждающий эффект описывает торможение или препятствование законному осуществлению права. Доказано, что если люди знают, что за ними ведется наблюдение, они начинают вести себя и развиваться по-другому.

Штабен, Й. (2016). Der Abschreckungseffekt auf die Grundrechtsausübung: Strukturen eines verfassungsrechtlichen Arguments. Mohr Siebeck.

после съемки, технология все равно нарушает нашу психологическую целостность.

И хотя в случае с распознаванием лиц воздействие на наше «общее» право на частную жизнь и нашу психологическую целостность может быть более очевидным, можно утверждать, что неизбирательное отслеживание всех аспектов нашей жизни в режиме онлайн и офлайн (через наше поведение в сети, данные о местоположении, данные IoT от умных часов, медицинских трекеров, умных колонок, термостатов, автомобилей и т.д.) может иметь такое же воздействие на наше право на частную жизнь, включая нашу психологическую целостность.

Другие формы биометрического распознавания, управляемого ИИ, оказывают еще большее влияние на нашу психологическую целостность. Распознавание микровыражений, ворот, (тембра) голоса, пульса, температуры и т.д. в настоящее время используется для оценки или даже прогнозирования нашего поведения, психического состояния и эмоций.

Следует сразу отметить, что не существует надежных научных доказательств того, что внутренние эмоции или психическое состояние человека можно точно «прочитать» по его лицу, воротам, пульсу, тону голоса или температуре, не говоря уже о том, что по ним можно предсказать будущее поведение. В недавнем мета-исследовании группа ученых пришла к выводу, что распознавание эмоций с помощью ИИ может, максимум, распознать, как человек субъективно интерпретирует определенный биометрический признак другого человека. Интерпретация не совпадает с тем, что человек чувствует на самом деле, а ИИ просто маркирует эту интерпретацию, которая сильно зависит от контекста и культуры. Надуманные заявления о том, что ИИ может, например, определить, будет ли человек успешен в работе, основываясь на микровыражениях или тоне голоса, просто лишены научной основы.

Более того, широкое использование подобных методов ИИ, например, при приеме на работу, в правоохранительных органах, в школах, влияет на

физическую, психологическую или моральную целостность человека и, следовательно, на элементы его частной жизни.

Следует отметить, что GDPR ограничивает обработку биометрических данных лишь в некоторой степени. Биометрические данные, согласно GDPR, – это «персональные данные, полученные в результате специальной технической обработки, относящиеся к физическим, физиологическим или поведенческим характеристикам физического лица, которые позволяют или подтверждают уникальную идентификацию этого физического лица». Последняя часть предложения имеет решающее значение, поскольку если биометрическое распознавание не направлено на идентификацию (а, например, на категоризацию, профилирование или распознавание аффектов), оно может не подпадать под определение GDPR. На самом деле, в статье 51 GDPR говорится, что «обработка фотографий [считается] биометрическими данными только в том случае, если она осуществляется с помощью специальных технических средств, позволяющих однозначно идентифицировать или аутентифицировать физическое лицо».

Многие технологии биометрического распознавания не направлены на обработку биометрических данных для уникальной идентификации человека, а просто для оценки поведения человека (например, в классе) или для категоризации людей (например, для определения их страховой премии на основе статистической распространенности проблем со здоровьем). Эти виды использования могут не подпадать под определение биометрических данных (обработки) в соответствии с GDPR.

Возвращаясь к вопросу конфиденциальности данных, следует отметить, что персональные и неперсональные данные используются не только для обучения систем искусственного интеллекта, но и для профилирования и оценки людей в различных целях, таких как предиктивная полиция, принятие страховки, выплата социальных пособий, прогнозирование производительности в процессах найма и увольнения. Более того, огромное количество «точек данных» о том, как мы ведем повседневную жизнь, используется не только для

рассылки нам целевой рекламы, но и для того, чтобы подтолкнуть/влиять/побудить/подтолкнуть нас к определенной информации и таким образом повлиять на наш выбор, нарушая нашу моральную целостность.

ИИ и свобода личности

Свобода личности отражена в ЕКПЧ в различных правах, таких как свобода выражения мнения (статья 10) и свобода собраний и объединений (статья 11). ИИ может оказывать «охлаждающее» воздействие и на эти свободы.

Свобода выражения мнения (ст. 10 ЕКПЧ)

Ст. 10 ЕКПЧ предусматривает право на свободу выражения мнения и информации, включая свободу придерживаться своего мнения, получать информацию и идеи. Использование ИИ для составления профиля, опроса, отслеживания и идентификации людей, отбора, определения, сортировки и влияния или подталкивания поведения не только потенциально влияет на право на моральную неприкосновенность, как описано выше, но и может оказать охлаждающее воздействие на эти конкретные свободы.

Использование системы распознавания лиц в общественных местах может нарушить свободу мнения и самовыражения человека просто потому, что защита «групповой анонимности» больше не существует, если каждый в группе может быть потенциально узнан. Это может привести к тому, что люди изменят свое поведение, например, перестанут участвовать в мирных демонстрациях.¹²

То же самое относится и к ситуации, когда все наши данные используются для подсчета, оценки и оценки эффективности с помощью искусственного интеллекта (например, для получения кредита, ипотеки, займа, работы, повышения по службе и т.д.). Люди могут стать более нерешительными в том, чтобы открыто выражать определенное мнение, читать

¹² Отчет об оценке воздействия на частную жизнь при использовании технологий распознавания лиц для идентификации субъектов в полевых условиях, 30 июня 2011 г., стр. 18.

определенные книги или газеты онлайн или смотреть определенные онлайн-медиа.

Что касается права на получение и распространение информации и идей, ИИ, используемый в СМИ и новостях, приносящий все более «персонализированный» онлайн-контент и новости людям, вызывает беспокойство. Поисковые системы, системы рекомендаций видео и новостные агрегаторы часто непрозрачны, как в отношении данных, которые они используют для отбора или определения приоритетов контента, так и в отношении цели конкретного отбора или определения приоритетов.¹³ Многие бизнес-модели основаны на доходах от рекламы в Интернете. Для того чтобы люди проводили как можно больше времени на платформе или сайте, они могут выбирать и отдавать предпочтение контенту, который будет делать только это: удерживать людей на своей платформе, независимо от того, является ли этот контент объективным, фактологически достоверным, разнообразным или даже актуальным.

Помимо коммерческих мотивов, политические или иные мотивы могут привести к тому, что системы ИИ будут оптимизированы для выбора или приоритезации определенного контента в попытке принудить и повлиять на людей в пользу определенных точек зрения, например, во время избирательных процессов.¹⁴

Более того, ИИ становится все более способным производить медиаматериалы (видео, аудио, изображения), напоминающие внешность и/или голос реальных людей (также известные как «глубокие подделки»), что позволяет использовать обманные практики в различных целях.

Все это может привести к возникновению «пузырей фильтров» и распространению фальшивых новостей, дезинформации и пропаганды, а также влияет на способность людей формировать и развивать мнения, получать и

13 Баррелл, Дж. (2016). Как машина «думает»: Понимание непрозрачности в алгоритмах машинного обучения. *Big Data & Society*, 3(1)

14 Cambridge Analytica, документальный фильм Netflix: Великий взлом

передавать информацию и идеи и, таким образом, влияет на нашу свободу выражения мнений.¹⁵

Свобода собраний и объединений (статья 11 ЕКПЧ)

Интернет и социальные медиа оказались полезными инструментами для осуществления людьми своего права на мирные собрания и ассоциации. Однако в то же время использование ИИ может поставить под угрозу эти права, когда люди или группы людей автоматически отслеживаются и идентифицируются и, возможно, даже «исключаются» из демонстраций или протестов.¹⁶

Как уже упоминалось, использование системы распознавания лиц в общественных местах может отбить у людей желание посещать демонстрации и участвовать в мирных собраниях, что является одним из важнейших элементов демократического общества. Примеры этого уже наблюдались в Гонконге, когда протестующие начали носить маски и использовать лазеры, чтобы избежать «попадания» в камеры распознавания лиц. ИИ и равенство, нет дискриминации и солидарность.

Запрет дискриминации (статья 14 ЕКПЧ, Протокол 12)

Одним из наиболее известных воздействий ИИ на права человека является воздействие на запрет дискриминации и право на равное обращение. Как отмечалось ранее, во многих случаях, как показывает практика, ИИ увековечивает, усиливает и, возможно, закрепляет дискриминационные или иным образом неприемлемые предубеждения. Кроме того, ИИ может расширить группу людей, на которых оказывается воздействие, когда он группирует их на основе общих характеристик.¹⁷ Более того, эти системы,

¹⁵ Доклад Специального докладчика ООН по вопросу о поощрении и защите права на свободу мнений и их свободное выражение, A/73/348.

¹⁶ Алгоритмы и права человека, Исследование о правозащитных аспектах методов автоматизированной обработки данных и возможных регулятивных последствиях, Совет Европы, 2018 г.

¹⁷ Аффинити-профайлинг и дискриминация по ассоциации в поведенческой онлайн-рекламе, Wachter 2020.

основанные на данных, скрывают существование предубеждений, оттесняя на второй план механизмы социального контроля, управляющие поведением человека.

Например, ИИ компании Amazon при подборе персонала отдавал предпочтение мужчинам, а не женщинам, поскольку был обучен на профилях успешных сотрудников Amazon, которыми оказались мужчины. ИИ-система не просто отсеивала женщин, она рассматривала характеристики успешных сотрудников, такие как типичные формулировки и формулировки, и отсеивала резюме, которые не соответствовали этим характеристикам.¹⁸

Если вернуться к работе современного ИИ, где системы просто ищут корреляции на основе общих характеристик с другими «случаями», то легко могут всплыть все виды неприемлемых предубеждений. Проблема с этими системами заключается в том, что даже если бы они прекрасно справлялись с выявлением закономерностей, например, типичных фраз, используемых успешными сотрудниками, система не понимает смысла этих фраз, не говоря уже о том, что она сможет понять значение успеха или даже понять, что такое сотрудник. Она сможет лишь дать ярлык определенному шаблону.

Вопреки распространенному мнению, не все предубеждения являются результатом некачественных данных. Проектирование любого артефакта само по себе является накоплением необъективных решений, начиная от рассматриваемых входных данных и заканчивая целями, поставленными для оптимизации; оптимизируется ли система для чистой эффективности, или же она учитывает влияние на работников и окружающую среду? Является ли целью системы найти как можно больше потенциальных мошенников, или она избегает выявления невинных людей? Все эти решения так или иначе обусловлены предвзятостью, присущей человеку (людям), который их принимает. Короче говоря, предполагать, что мы сможем устранить все

18 Дастин, Дж. (2018, 10 октября). Amazon отменила секретный инструмент ИИ для подбора персонала, который показал предвзятое отношение к женщинам. Reuters.

предубеждения в ИИ (или даже с его помощью), – значит принимать желаемое за действительное.¹⁹

ИИ и социальные и экономические права

ИИ может принести большую пользу при использовании для выполнения опасной, тяжелой, изнурительной, грязной, неприятной, повторяющейся или скучной работы. Однако системы ИИ также все чаще используются для мониторинга и отслеживания работников, распределения работы без вмешательства человека, а также для оценки и прогнозирования потенциала и производительности работников в ситуациях найма и увольнения. Такое применение ИИ может поставить под угрозу право на справедливые условия труда, безопасные и здоровые условия труда, достоинство на работе, а также право на организацию (ст. 2 и 3 ESC). Если за работниками постоянно следят работодатели, они могут стать более нерешительными в вопросах организации (статья 5). ИИ-системы, оценивающие и прогнозирующие производительность работников, могут поставить под угрозу право на равные возможности и равное обращение в вопросах труда и занятий без дискриминации по признаку пола (статья 20 ESC), особенно если эти системы закрепляют предубеждения в данных или в их создателе.

Существует риск утраты необходимых навыков, когда все больше работы и решений, которые раньше выполнялись или принимались людьми, переходят к ИИ-системам. Это может не только привести к снижению квалификации рабочей силы, но и повысить риск системного сбоя, когда только несколько человек способны работать с ИИ-системами и реагировать на события, когда эти системы дают сбой.

Хотя неизвестно, будут ли потеряны или приобретены рабочие места в результате внедрения ИИ, и если да, то сколько, в период разрушительных

19 Первый анализ Белой книги ЕС по ИИ, Вирджиния Дигнум, Кателин Мюллер, Андреас Теодору, 2020.

преобразований несоответствие между уязвимыми трудовыми ресурсами и требуемыми навыками может привести к технологической безработице.²⁰

Искусственный интеллект и демократия

ИИ может оказывать (и, вероятно, уже оказывает) негативное влияние на демократию, в частности, когда речь идет о: (I) социально-политическом дискурсе, доступе к информации и влиянии на избирателей, (II) неравенстве и сегрегации и (III) системных сбоях или нарушениях.

Социально-политический и общественный дискурс, доступ к информации и влияние на избирателей

Хорошо функционирующие демократии требуют наличия хорошо информированных граждан, открытого социального и политического дискурса и отсутствия непрозрачного влияния на избирателей.

Для этого требуется хорошо информированный гражданин. В информационных обществах граждане могут выбрать для потребления лишь небольшое количество всей доступной информации. Поисковые системы, ленты социальных сетей, рекомендательные системы и многие новостные сайты используют ИИ для определения того, какой контент создается и показывается пользователям (персонализация информации). При правильном подходе это может помочь гражданам лучше ориентироваться в потоке доступной информации и повысить их демократическую компетентность, например, предоставив им доступ к ресурсам на других языках с помощью инструментов перевода.²¹ Однако если ИИ определяет, какая информация показывается и потребляется, какие вопросы подавляются в потоке онлайн-информации, а какие получают вирусное распространение, это также несет риски предвзятости и неравного представления мнений и голосов.

Персонализация информации с помощью ИИ обеспечивается постоянным мониторингом и профилированием каждого человека. Движимая

20 Бриниолфссон, Э., и МакАфи, А. (2014). Второй машинный век: Работа, прогресс и процветание в эпоху блестящих технологий. W. W. Norton & Company.

21 Шредер, Р. (2018). Социальная теория после Интернета. UCL Press.

коммерческими или политическими мотивами, эта технологически оснащенная информационная инфраструктура наших обществ может усилить гиперпартийный контент, с которым человек, скорее всего, согласится, и предоставить беспрецедентно мощный инструмент для индивидуального влияния.²² Как следствие, это может подорвать общее понимание, взаимное уважение и социальную сплоченность, необходимые для процветания демократии. Если персональные прогнозы ИИ станут очень мощными и эффективными, они могут даже угрожать подорвать человеческую самостоятельность и автономию, необходимые для принятия избирателями значимых решений.²³

В-третьих, ИИ может подорвать справедливый избирательный процесс. Политические кампании или иностранные субъекты могут использовать (и уже используют) персонализированную рекламу для передачи различных сообщений отдельным группам избирателей без публичной отчетности на агоре.²⁴ Однако следует отметить, что остается неясным, насколько влиятельна микро-таргетированная реклама.²⁵ ИИ также может использоваться для создания и распространения дезинформации и глубоких подделок в виде текста, изображений, аудио или видео. Поскольку их трудно идентифицировать гражданам, журналистам или государственным учреждениям, вводить в заблуждение и манипулировать общественностью становится проще, а уровень правдивости и доверия к СМИ и демократическому дискурсу может ухудшиться.

Неравенство и сегрегация

Широко ожидается, что ИИ повысит производительность экономики. Тем не менее, ожидается, что этот рост производительности будет распределяться

22 Wu, T. (2016). Эпоха капитализма наблюдения: Борьба за будущее на новом рубеже власти.

23 Taddeo, M., & Floridi, L. (2018b). Как искусственный интеллект может стать силой добра. Наука, 361(6404), 751-752.

24 Брэдшоу, С., и Ховард, П. (2019). Социальные медиа и демократия в условиях кризиса. Oxford University Press.

25 Бонд, Р.М., Фарисс, К.Дж., Джонс, Дж.Дж., Крамер, А.Д.И., Марлоу, К., Сеттл, Дж.Е., и Фаулер, Дж.Х. (2012). Эксперимент по социальному влиянию и политической мобилизации на 61 миллион человек. Nature, 489(7415), 295-298.

неравномерно, и большинство выгод получают обеспеченные люди. Аналогичным образом, выбор данных и дизайна в сочетании с недостаточной прозрачностью алгоритмов «черного ящика», как показывает практика, может привести к сохранению несправедливого предубеждения в отношении и без того уязвимых групп общества, таких как женщины и этнические меньшинства.²⁶ ИИ может привести к неравенству и сегрегации и тем самым поставить под угрозу необходимый уровень экономического и социального равенства, требуемый для процветающей демократии.

Системные риски

Решения ИИ, которые ранее могли принимать только люди, создают новые проблемы для безопасности и устойчивости общественных систем. В частности, если решения, которые ранее принимались многими децентрализованными субъектами, заменяются несколькими централизованными системами, управляемыми ИИ, системные риски возрастают, когда для потенциально катастрофических результатов достаточно сбоя лишь нескольких централизованных систем.

Финансовые рынки демонстрируют, как возникают новые системные риски, когда различные агенты ИИ взаимодействуют на сверхчеловеческих скоростях, как это продемонстрировал рост числа финансовых флеш-катастроф.²⁷ Когда критически важные энергетические инфраструктуры, транспортные системы и больницы все больше зависят от автоматизированных решений искусственных агентов, это создает новые уязвимости в виде единой точки отказа с широкомасштабными последствиями. После внедрения эффективных систем критической инфраструктуры их сложнее заменить или создать резервные копии в случае поломки.²⁸

26 Юбэнкс, В. (2017). Automating inequality: How high-tech tools profile, police, and punish the poor (First Edition). St. Martin's Press; O'Neil, C. (2016). Оружие математического разрушения: Как большие данные увеличивают неравенство и угрожают демократии (Первое издание).

27 Веллман, М. П., и Раджан, У. (2017). Этические вопросы для автономных торговых агентов. Minds and Machines, 27(4), 609-624.

28 Брундаж, М., Авин, С., Кларк, Ж., Тонер, Х., Экерсли, П., Гарфинкель, Б., Дафо, А., Шарпе, П., Зейтцхофф, Т., Филар, Б., Андерсон, Х., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Eigartaigh, S. O., Beard, S.,

Особая опасность для международной безопасности и мира заключается в том, что развитие ИИ рассматривается как конкурентная гонка. ИИ не только приведет к нежелательным побочным эффектам, но и расширит возможности злонамеренных субъектов – от киберпреступников до тоталитарных государств в их стремлении контролировать население.

Концентрация цифровой власти

Многие приложения ИИ разрабатываются и внедряются лишь несколькими крупными частными компаниями, которых иногда называют «Большой пятеркой», GAFAM или даже «Страшной пятеркой».²⁹ Если слишком много политической власти сосредоточено в руках нескольких частных компаний, которые отдают предпочтение акционерной стоимости, а не общему благу, это может угрожать авторитету демократических государств.

ИИ и верховенство закона

К государственным учреждениям предъявляются более высокие требования, когда речь идет об их поведении в отношении отдельных лиц и общества, что отражено в таких принципах, как обоснованность, пропорциональность и равенство. ИИ может повысить эффективность институтов, но с другой стороны, он также может подорвать процедурную легитимность и доверие к демократическим институтам и авторитету закона. Суды, правоохранительные органы и государственное управление могли бы стать более эффективными, но ценой большей непрозрачности и меньшей автономии и контроля со стороны человеческих агентов³⁰.

Аналогичным образом, если раньше только суды определяли, что считается незаконным языком вражды, то сегодня в основном частные системы

Belfield, H., Farquhar, S., Amodei, D. (2018). Злонамеренное использование искусственного интеллекта: Прогнозирование, предотвращение и смягчение последствий.

29 Google, Facebook, Microsoft, Apple и Amazon. См: Немитц, П. (2018). Конституционная демократия и технология в эпоху искусственного интеллекта. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133), 20180089. Вебб, А. (2019). The Big Nine: How the tech titans and their thinking machines could warp humanity.

30 Данахер, Дж. (2016). Угроза аллократии: Реальность, сопротивление и приспособление. *Философия и технология*, 29(3), 245–268.

Вейценбаум, Дж. (1976). Мощность компьютера и человеческий разум: From Judgment to Calculation. W. H. Freeman.

искусственного интеллекта определяют, будут ли удалены высказывания платформами социальных сетей. Эти системы ИИ де-факто конкурируют за полномочия с судьями и законом. В целом, ИИ может способствовать развитию судебных систем, которые действуют вне границ и защиты верховенства закона.

Автоматизированные онлайн-решения споров, предоставляемые частными компаниями, регулируются условиями предоставления услуг, а не законом, который не предоставляет потребителям те же права и процессуальную защиту, что и государственные суды³¹.

Европейская комиссия по эффективности правосудия уже в 2018 году изложила 5 принципов использования ИИ в судебной системе в «Европейской этической хартии по использованию ИИ в судебных системах и их окружении». Группа экспертов высокого уровня по ИИ призвала государственные органы соблюдать 7 требований к надежному ИИ при разработке, закупке или использовании ИИ. Аналогичные принципы и требования должны быть предъявлены и к правоохранительным органам.

Однако ИИ может не только угрожать верховенству закона, но и укреплять его.³² При ответственном подходе к разработке и использованию, он может расширить возможности ведомств по выявлению коррупции в государстве.³³ Аналогичным образом, ИИ может быть использован для обнаружения и защиты от кибератак.³⁴

31 Сасскинд, Дж. (2018). Политика будущего: Совместная жизнь в мире, преобразованном технологиями.

32 Винуэса, Р., Азизпур, Х., Лейте, И., Балаам, М., Дигнум, В., Домиш, С., Фелландер, А., Лангханс, С. Д., Тегмарк, М., и Нерини, Ф. Ф. (2020). Роль искусственного интеллекта в достижении Целей устойчивого развития. *Nature Communications*, 11(1), 1-10.

33 Вест, Дж. и Бхаттачария, М. (2016). Интеллектуальное обнаружение финансового мошенничества: Всесторонний обзор. *Computers & Security*, 57, 47-66. Hajek, P., & Henriques, R. (2017). Добыча корпоративных годовых отчетов для интеллектуального обнаружения мошенничества в финансовой отчетности – сравнительное исследование методов машинного обучения. *Knowledge-Based Systems*, 128, 139- 152.

34 Таддео, М., и Флориди, Л. (2018а). Регулирование искусственного интеллекта для предотвращения гонки кибервооружений. *Nature*, 556(7701), 296- 298.

Как решить проблему воздействия ИИ на права человека, демократию и верховенство закона?

Влияние искусственного интеллекта на права человека, демократию и верховенство закона в последнее время привлекает все больше внимания, что наиболее заметно в недавней «Белой книге по искусственному интеллекту» Европейской комиссии. Однако, как бороться с этим воздействием, до сих пор остается неопределенной территорией. В этой главе описаны возможные стратегии, которые могут быть реализованы в существующих рамках прав человека, демократии и верховенства закона. Эти стратегии не являются исчерпывающими и должны помочь продвинуть дискуссию к следующему этапу.

Включение прав человека в контекст ИИ

Многие разработчики, создатели и пользователи ИИ (государственные и частные), похоже, не знают о (потенциальном) воздействии ИИ на права человека. В качестве первого шага очень полезно провести итерацию или (ре)артикуляцию, в ходе которой существующие права человека ЕКПЧ будут «переведены» в контекст ИИ, и это можно сделать с помощью Рамочной конвенции.

Меры по соблюдению, подотчетности и возмещению ущерба

Для надлежащего решения проблемы воздействия ИИ на существующие права человека, демократию и верховенство закона можно было бы доработать некоторые существующие механизмы соблюдения, подотчетности и возмещения ущерба, а также внедрить новые механизмы. Однако важно то, что использование ИИ часто скрыто или неизвестно, что затрудняет или делает невозможным определить, есть ли воздействие на права человека, демократию и верховенство закона в первую очередь. Таким образом, меры по обеспечению соответствия, подотчетности и возмещению ущерба должны начинаться с обязательства обеспечить прозрачность использования систем ИИ, которые

могут повлиять на права человека, демократию или верховенство закона. Это включает в себя реестр ИИ, в котором затем указывается класс риска и требуемый объем прозрачности и подотчетности для конкретного применения.

Соблюдение норм может начаться с того, что недавно было названо новой культурой «прав человека, демократии и верховенства закона по замыслу».³⁵ В такой культуре разработчики, внедрители и пользователи ИИ с самого начала будут думать о том, как технология может повлиять на права человека, демократию и верховенство закона, и соответствующим образом корректировать технологию или ее использование. Это может быть подкреплено (юридическим) обязательством проводить оценку воздействия ИИ на права человека, демократию и верховенство закона.

Такая новая культура должна включать в себя обязательство учитывать не только соответствующую структуру, которая будет создана, но и результаты оценки воздействия ИИ на права человека, демократию и верховенство закона, а также решения по разработке и управлению на ее основе.

Возмещение ущерба в свете воздействия ИИ на права человека подразумевает доступ к правосудию и эффективным средствам правовой защиты. Что касается доступа к правосудию, то, возможно, еще слишком рано определять, достаточно ли он гарантирован, когда речь идет о воздействии ИИ на права человека. Только недавно мы увидели первые несколько решений национальных судов о (потенциальном) воздействии ИИ на одно конкретное право человека – право на неприкосновенность частной жизни, предусмотренное ст. 8 ЕКПЧ.³⁶

Однако еще важнее то, что доступ к правосудию затруднен, когда многие приложения ИИ разрабатываются и внедряются лишь горсткой крупных частных компаний. Эти компании доминируют как в разработке ИИ, так и в

35 Немитц, П. (2018). Конституционная демократия и технологии в эпоху искусственного интеллекта.

36 Голландский суд рассмотрел закон, позволяющий государственным учреждениям потенциально использовать ИИ для предсказания мошенничества с социальными пособиями в нарушение права на частную жизнь ст. 8 ЕКПЧ, Высокий суд Великобритании в Кардиффе признал, что распознавание лиц затрагивает ст. 8 ЕКПЧ, поскольку позволяет извлекать «внутренне частную» информацию, но счел использование законным по соображениям пропорциональности. Французский суд счел, что использование системы распознавания лиц в школах нарушает ст. 8 ЕКПЧ.

(эко)системах, в которых работает ИИ. Хотя государства обязаны защищать отдельных лиц и группы лиц от нарушений прав человека, совершаемых другими субъектами, оценка влияния негосударственных субъектов на права человека неуклонно растет.³⁷

Поскольку эти крупные технологические компании теперь стали операторами, способными определять и, возможно, изменять наши социальные и даже демократические структуры, влияние их ИИ (-использования) на права человека становится более значимым. В этом отношении ИИ может послужить хорошей возможностью и подумать о структуре, которая юридически обязывала бы частных субъектов соблюдать права человека и предоставлять доступ к правосудию, если они этого не делают.³⁸ Основной вопрос заключается в том, что делать: а) принять частную власть компаний ИИ и убедиться, что они используют ее ответственно, или б) оспорить ее и попытаться вновь утвердить власть государства. Когда речь идет об эффективном средстве правовой защиты, ИИ – это тема, где, как заметил Шелдон, средства правовой защиты – это «не только возмещение ущерба жертве; они выражают осуждение правонарушителя с точки зрения общества в целом» и, таким образом, «подтверждают, усиливают и подтверждают фундаментальные ценности общества».³⁹ Европейский суд по правам человека в своем решении по делу Броневского подчеркнул, что международное право требует, чтобы «индивидуальное и общее возмещение (...) шли рука об руку».⁴⁰ Чтобы определить эффективное средство правовой защиты в случае нарушения прав человека в результате использования ИИ, необходимо рассмотреть как индивидуальные, так и общие средства правовой защиты. Более того, поскольку ИИ имеет огромное количество применений, начиная от наблюдения и идентификации, заканчивая профилированием, подталкиванием и принятием

37 Бизнес и права человека, руководство для юристов-практиков, Клэр Метвен О'Брайен, Совет Европы.

38 Это означает выход за рамки простой ссылки на Рекомендацию CM/Rec(2016)3 о правах человека и бизнесе Комитета министров Совета Европы (и Руководящие принципы ООН в области бизнеса и прав человека).

39 Дайна Шелтон, «Право на возмещение ущерба за акты пыток: Какое право, какие средства правовой защиты?», 17(2) Torture 96 (2007), at 96.

40 Броневский против Польши, ЕСПЧ.

решений, средства правовой защиты должны быть адаптированы к этим различным применениям. Надлежащие средства правовой защиты должны включать в себя прекращение противоправного поведения и гарантии неповторения, когда государства, например, могут быть обязаны принять и внедрить принудительное законодательство для защиты прав человека от воздействия ИИ в будущем. Должно существовать обязательство по возмещению вреда или ущерба, причиненного нарушением, как отдельному лицу, так и обществу. Для некоторых видов применения ИИ простого обеспечения эффективного средства правовой защиты может быть недостаточно для решения проблемы воздействия такого применения на права человека. Могут потребоваться более далеко идущие меры, такие как запрет или ограничение использования (см. главу IV).

Защита демократии, демократических структур и верховенства закона

Для предотвращения системных сбоев или нарушений в результате централизации процессов принятия решений, управляемых ИИ, в жизненно важных структурах, следует внедрять распределенные процессы принятия решений, а не централизованные, чтобы предотвратить риск катастрофического сбоя. В эти процессы должны быть встроены надлежащие структуры человеческого надзора.⁴¹

Надзор со стороны человека помогает убедиться, что система ИИ не подрывает автономию человека и не вызывает других негативных последствий. Надзор может осуществляться с помощью таких механизмов управления, как «человек в контуре» (HITL), «человек в контуре» (HOTL) или «человек в команде» (HIC). HITL относится к возможности вмешательства человека в каждый цикл принятия решений в системе, что во многих случаях невозможно и нежелательно. HOTL подразумевает возможность вмешательства человека во время цикла проектирования системы и мониторинга ее работы. HIC относится к возможности контролировать общую деятельность системы ИИ (включая ее

41 Руководящие принципы этики для надежного ИИ, Группа экспертов высокого уровня по ИИ, 2019 г.

более широкое экономическое, социальное, правовое и этическое воздействие) и способности принимать решения о том, когда и как использовать систему в любой конкретной ситуации. Это может включать решение не использовать систему ИИ в конкретной ситуации, устанавливать уровни свободы действий человека при использовании системы или обеспечивать возможность отмены решения, принятого системой. Чтобы устранить риск неравенства, правительствам необходимо активно пресекать использование приложений ИИ, которые увеличивают неравенство.

Предотвращение влияния на выборы или манипулирования общественностью с помощью персонализированной информации, управляемой ИИ, – задача не из легких. Можно рассмотреть вопрос о регулировании агитации в Интернете, либо для платформ (социальных) СМИ, либо для политических партий. Очевидно, что при этом возникают вопросы, касающиеся свободы слова. Привлечение людей к управлению и командованию (см. выше) может помочь обнаружить и устранить нежелательное влияние на избирателей.

Важнейшим рычагом в обеспечении ответственного использования ИИ в государственных услугах являются государственные закупки. Если обновить юридически обязательные требования к государственным закупкам, включив в них такие критерии, как справедливость, подотчетность и прозрачность ИИ, это может служить двум целям. С одной стороны, это гарантирует, что правительства будут использовать только те системы, которые совместимы с верховенством закона, а также создаст экономические стимулы для частного сектора разрабатывать и использовать системы, соответствующие принципам верховенства закона. Кроме того, использование ИИ в правительстве должно быть подчинено механизмам надзора, включая судебные постановления и омбудсменов для рассмотрения жалоб.

Что если нынешние права человека, демократия и верховенство закона не смогут защитить нас должным образом?

Нулевой вопрос

Из-за инвазивности некоторых приложений или видов использования ИИ могут возникнуть ситуации, в которых наша нынешняя система прав человека, демократии и верховенства закона не сможет адекватно или своевременно защитить нас, и когда нам, возможно, потребуется сделать паузу для размышлений и найти подходящий ответ на то, что можно считать «нулевым вопросом»: Хотим ли мы разрешить эту конкретную систему ИИ или ее использование, и если да, то при каких условиях? Ответ на этот вопрос должен заставить нас взглянуть на ИИ-систему или использование со всех точек зрения, что может привести к нескольким «решениям»:

На определенную ИИ-систему или ее использование накладывается мораторий, (временно или бессрочно) запрет или ограничения («красные линии») Новые права человека вводятся в качестве гарантий против «нового» негативного воздействия ИИ.

Существующие права человека адаптированы для ответственного развития и использования ИИ

Конкретная ИИ-система или ее использование подчиняется конкретному механизму демократического надзора

Частные владельцы мощных ИИ-систем обязаны согласовывать свои структуры разработки и управления ИИ с интересами тех, на кого влияет система, и общества в целом, что может включать меры по привлечению соответствующих сторон (таких как работники, потребители, клиенты, граждане, политические деятели).

Прежде всего, «воздействие ИИ» следует рассматривать как на индивидуальном, так и на общественном/коллективном уровне, поскольку ИИ может воздействовать как на отдельного человека, так и на более крупные части нашего коллективного общества. Во-вторых, контекст, цель, серьезность,

масштаб и вероятность воздействия важны для определения соответствующих и пропорциональных действий. Для приложений ИИ, создающих неприемлемые риски или угрозы причинения вреда или системного сбоя, которые являются существенными, следует применять осторожный и основанный на принципах подход к регулированию. Для других видов применения ИИ более подходящим может быть подход, основанный на оценке риска.

Красные линии

Красные линии могут быть проведены для определенных ИИ-систем или видов использования, которые считаются слишком влиятельными, чтобы оставлять их бесконтрольными или нерегулируемыми или даже разрешать. Такие применения ИИ могут привести к необходимости запрета, моратория и/или жестких ограничений или условий для исключительного и/или контролируемого использования:

неизбирательное использование распознавания лиц и других форм биометрического распознавания как государственными, так и частными субъектами

массовое наблюдение с помощью ИИ (с использованием распознавания лиц/биометрии, а также других форм отслеживания и/или идентификации с помощью ИИ, например, через службы определения местоположения, онлайн-поведения и т.д.)

Персональное, физическое или психическое отслеживание, оценка, профилирование, подсчет баллов и

подталкивание с помощью биометрических данных и распознавания поведения

Социальный скоринг с использованием ИИ

Скрытые системы ИИ и deepfakes

Интерфейсы «человек – ИИ»

Исключительное использование таких технологий, например, в целях национальной безопасности или медицинского лечения или диагностики, должно быть доказательным, необходимым и соразмерным и допускаться только в контролируемой среде и (если применимо) в течение ограниченного периода времени.

В дополнение к «красным линиям» – мерам, можно рассмотреть следующие адаптированные или новые права человека (неисчерпывающие):

Право человека на автономию, управление и надзор над ИИ

Право на прозрачность/объяснимость результатов ИИ, включая право на объяснение того, как функционирует ИИ, какой логике он следует и как его использование затрагивает интересы соответствующего лица, даже если система ИИ не обрабатывает личные данные, в этом случае право на такую информацию уже существует в соответствии с GDPR.⁴²

Отдельное право на физическую, психологическую и моральную целостность в свете профилирования ИИ, распознавания аффектов

Усиление права на неприкосновенность частной жизни для защиты от массового наблюдения, управляемого ИИ.

Адаптация права на неприкосновенность данных для защиты от неизбирательной, охватывающей все общество онлайн-слежки за людьми с использованием личных и неличных данных (которые часто служат косвенным средством идентификации личности).

Отступление от этих прав в исключительных обстоятельствах, например, в целях безопасности, должно допускаться только при строгих условиях и соразмерным образом.

Сценарии будущего

Экстраполируя в будущее с более длительным временным горизонтом, можно выдвинуть гипотезу о некоторых критических долгосрочных проблемах, которые исследуются, что требует подхода, основанного на риске, ввиду

42 Немитц, П. (2018). Конституционная демократия и технологии в эпоху искусственного интеллекта.

возможных неизвестных и «черных лебедей». В то время как некоторые считают, что искусственный интеллект, искусственное сознание, искусственные моральные агенты, сверхинтеллект могут быть примерами таких долгосрочных проблем (в настоящее время не существующих), многие другие считают их нереальными. Тем не менее, необходимо внимательно следить за развитием событий, чтобы определить, нужна ли постоянная адаптация наших систем прав человека, демократии и верховенства закона.

ALLAI относится к Stichting ALLAI Nederland, фонду в соответствии с голландским законодательством

Данный проект документа содержит только общую и ориентировочную информацию, и ни ALLAI, ни его члены правления, члены консультативного совета, сотрудники, должностные лица, ассоциированные организации или лица, или связанные с ними организации (совместно именуемые «сеть ALLAI») не предоставляют посредством данного документа профессиональные консультации или услуги. Прежде чем принять какое-либо решение или предпринять какие-либо действия, которые могут повлиять на ваши финансы или вашу организацию, вам следует проконсультироваться с квалифицированным профессиональным консультантом. Ни одна организация или лицо из сети ALLAI не несет ответственности за любые прямые или косвенные убытки или ущерб, понесенные любой организацией или лицом, полагающимся на данный документ.

Руководящие принципы этики ИИ: Европейские и глобальные перспективы⁴³

I. Исполнительное резюме

В последние годы частные компании, исследовательские институты и организации государственного сектора выпустили принципы, руководства и другие инструменты «мягкого права» для этичного использования искусственного интеллекта (ИИ). Однако, несмотря на очевидное согласие с тем, что ИИ должен быть «этичным», ведутся споры о том, что представляет собой «этичный ИИ» и какие этические требования, технические стандарты и лучшие практики необходимы для его реализации. Целью данного отчета является составление карты соответствующего корпуса документов «мягкого права» и других этико-правовых рамок, разработанных правительственными и неправительственными организациями по всему миру с двойной целью.

Во-первых, мы хотим отслеживать этот постоянно развивающийся спектр необязательных инструментов управления.

Во-вторых, мы хотим в перспективе оценить влияние ИИ на этические принципы, права человека, верховенство закона и демократию. В отчете используется адаптированный и предварительно проверенный протокол обзорного исследования, чтобы предоставить всеобъемлющий и актуальный обзор текущих усилий в области мягкого права. Мы проанализировали 116 документов, опубликованных, в частности, правительственными учреждениями, неправительственными организациями, академическими институтами и частными компаниями.

Наш анализ выявляет пять выдающихся кластеров этических принципов и оценивает их роль в текущем дискурсе управления, выявляет существующие «слепые пятна» и интерпретационные пробелы в современном ландшафте мягкого права.

Кроме того, мы устанавливаем связь между этическими принципами и правами человека, уделяя особое внимание правам и свободам, закрепленным в

⁴³ Отчет, подготовленный для САНАИ. Страсбург, 15 июня 2020 года, под названием САНАИ(2020)07-fin

Европейской конвенции по правам человека (ЕКПЧ), чтобы оценить, в какой степени защита прав человека является неотъемлемой частью современных необязательных систем управления.

Наконец, мы предоставляем эмпирически обоснованные политические последствия для ученых, исследовательских институтов, финансирующих агентств, правительственных и межправительственных организаций и других заинтересованных сторон, участвующих в продвижении этически ответственных инноваций в ИИ.

II. Основные выводы

Все больше правительственных и неправительственных организаций (включая частные компании и академические организации) разрабатывают руководства по этике или другие документы «мягкого права» по ИИ.

Эти документы «мягкого права» разрабатываются в основном в Европе, Северной Америке и Азии. Глобальный Юг в настоящее время недостаточно представлен в списке организаций, предлагающих руководящие принципы этики ИИ.

Существующие руководства по этике ИИ, как правило, соглашаются с некоторыми общими принципами, но резко расходятся во мнениях относительно деталей того, что следует делать на практике. Более того, ни один этический принцип не является общим для всех 116 документов по этике ИИ, которые мы рассмотрели.

Мы обнаружили растущее согласие по следующим этическим принципам: прозрачность, справедливость, непричинение вреда, ответственность и конфиденциальность. Этические соображения, касающиеся устойчивости, достоинства и солидарности, представляются значительно менее представленными.

Большинство руководств согласны с тем, что ИИ должен быть прозрачным, чтобы избежать потенциальных проблем. Однако неясно, должна

ли прозрачность достигаться путем публикации исходного кода, базовых баз данных или каким-либо другим способом.

Чуть более половины рассмотренных документов по мягкому праву прямо рекомендуют поощрять права человека – или предостерегают от их нарушения – при проектировании, разработке и внедрении систем ИИ.

Регулярные выражения, построенные на основе кодов, выявили значительные различия в охвате тем в документах, подготовленных в странах-членах Совета Европы (СЕ), по сравнению с документами, подготовленными в других странах. По сравнению с остальным миром, документы по мягкому праву, подготовленные в странах-членах Совета Европы, как представляется, подчеркивают этические принципы солидарности, доверия и надежности. В отличие от них, принципы благодеяния и достоинства упоминаются в них более эпизодически.

Принципы конфиденциальности, справедливости и честности показали наименьшие различия между странами-членами СЕ, странами-наблюдателями СЕ и остальным миром, а значит, и самую высокую степень межгеографической и межкультурной стабильности.

III. Основные политические последствия

Документы «мягкого права», издаваемые правительственными и неправительственными организациями (включая частные компании и академические организации), являются полезными инструментами для оказания практического влияния на принятие государственных решений в области ИИ и управления развитием систем ИИ в интересах общества и в соответствии с этическими ценностями и правовыми нормами. Однако подходы «мягкого права» не следует считать заменой обязательного управления. Из-за конфликта интересов усилия по саморегулированию, предпринимаемые частными субъектами ИИ, подвержены особому риску продвижения в обход или вместо обязательного управления со стороны правительственных и межправительственных органов.

Для обеспечения инклюзивности, культурного плюрализма и справедливого участия в принятии коллективных решений по ИИ следует поощрять разработку документов «мягкого права» организациями, расположенными в недопредставленных в настоящее время глобальных регионах, особенно в Африке и Южной Америке.

Сближение существующих инструментов «мягкого права» вокруг пяти общих этических принципов, таких как прозрачность, справедливость, непричинение вреда, ответственность и конфиденциальность, выявляет пять приоритетных областей для надзора и возможного вмешательства со стороны обязательных органов управления как на правительственном, так и на межправительственном уровне.

Для того чтобы эти этические принципы воплотились в эффективное управление, их необходимо концептуально прояснить. Разработчики политики обязаны устранить семантические двусмысленности и противоречивые характеристики этих принципов.

Резкие разногласия в существующих документах по «мягкому праву» относительно интерпретации и практической реализации этих принципов указывают на то, что решения в области обязательного управления, скорее всего, будут вызывать разногласия в обществе, а значит, требуют прозрачного процесса демократического обсуждения.

Недостаточно представленные этические соображения, такие как устойчивость, достоинство и солидарность, нуждаются в дальнейшем тщательном изучении, чтобы избежать импорта в обязательное управление тех же концептуальных пробелов и нормативных слепых пятен, что и в «мягком праве».

Поскольку почти половина рассмотренных документов «мягкого права» не содержит четких рекомендаций по поощрению или предупреждению нарушения прав человека при проектировании, разработке и внедрении систем ИИ, необходимо срочно уделить больше внимания последствиям ИИ для прав человека.

Страны-члены Совета Европы имеют все возможности для того, чтобы направить международное управление ИИ в сторону поощрения прав человека.

Введение

Искусственный интеллект (ИИ) – это изучение и разработка компьютерных систем, способных выполнять задачи, для которых обычно требуется человеческий интеллект. Обычно компьютерные системы считаются интеллектуальными (поэтому их называют «интеллектуальными агентами» или «интеллектуальными машинами»), когда они обладают способностью воспринимать окружающую среду и предпринимать автономные действия, направленные на успешное достижение цели. С исторической точки зрения, хотя общие размышления о механическом мышлении присутствовали в научной и философской литературе с древних времен, область ИИ в узком смысле возникла в 1940-х годах как следствие одновременного прогресса в математической логике (например, тезис Черча-Тьюринга), теории информации, нейробиологии и кибернетике. Область искусственного интеллекта включает в себя множество сложных вычислительных подходов, которые воспроизводят или имитируют когнитивные функции, такие как обучение, память, рассуждения, зрение и обработка естественного языка. Наиболее распространенный из этих подходов называется машинным обучением (МОО) и подразумевает разработку алгоритмов, которые выполняют задачи в отсутствие явных инструкций от человека. В отличие от обычных компьютерных программ, алгоритмы ML строят математические модели на основе обучающих данных и полагаются исключительно на умозаключения и выявление закономерностей для принятия автономных прогнозов и решений. Сегодня ИИ является одним из основных катализаторов технологических преобразований. На заре 2020-х годов системы ИИ встроены в бесчисленное количество систем и устройств, регулярно используемых человеком, таких как мобильные телефоны, социальные сети, автомобили, самолеты, аналитическое программное обеспечение, системы электронной связи, бытовая техника и т.д. ИИ является неотъемлемой частью широкого

спектра человеческой деятельности, включая (но не ограничиваясь) телекоммуникации, транспорт, производство, здравоохранение, банковское дело, страхование, правоохранительные органы и вооруженные силы.

Благодаря своей технологической новизне, способности к автономным действиям и общецелевому характеру, ИИ обладает потенциалом для преобразования человеческого общества более быстрыми темпами и в больших масштабах по сравнению с любой другой технологией. Трансформационный потенциал ИИ был признан экспертами «революционным», а авторы называют развитие ИИ «продолжающейся революцией», которая «изменит почти все сферы деятельности». По этой причине крайне важно и срочно необходимо оценить последствия ИИ для основных принципов и ценностей человеческой жизни, будущего человеческих обществ и систем правил, которые управляют этими обществами, в первую очередь демократии и верховенства закона

В последние годы несколько правительственных и межправительственных организаций, а также негосударственных субъектов выпустили принципы, руководства, рекомендации, рамки управления или другие инструменты мягкого права для ИИ. Инструменты мягкого права – это нормативные документы, которые не являются юридически обязательными или подлежащими исполнению, но имеют убеждающий характер и могут оказывать практическое влияние на принятие решений, сравнимое с обязательными нормативными актами (жесткими законами). Цель этих документов – направить развитие ИИ на социальное благо и в соответствии с этическими ценностями и правовыми нормами. Однако, несмотря на очевидное согласие с тем, что ИИ должен быть «этичным», ведутся споры о том, что представляет собой «этичный ИИ» и какие этические требования, технические стандарты и лучшие практики необходимы для его реализации. Кроме того, из-за быстрого распространения документов «мягкого права», связанных с ИИ, и большого разнообразия их эмитентов, трудно отследить и осмыслить этот постоянно развивающийся свод необязательных норм управления всеобъемлющим и строгим образом.

Данный отчет представляет собой комплексное описание и мета-анализ существующего свода принципов и руководств по этическому ИИ. Этот анализ будет полезен ученым, исследовательским институтам, финансирующим агентствам, правительственным и межправительственным организациям и другим заинтересованным сторонам, участвующим в продвижении этически ответственных инноваций в области ИИ. Кроме того, в нем будет обсуждаться, как эти этические принципы, моральные обычаи и рекомендуемые социальные практики могут быть воплощены в обязательные правила управления, особенно в юридические документы, имеющие обязательную международную силу.

Особое внимание в этом докладе уделяется изучению связи между управлением ИИ и правами человека и перспективной оценке влияния технологий ИИ на права и свободы человека. Права человека – это права, присущие всем людям, независимо от расы, пола, национальности, этнической принадлежности, языка, религии или любого другого статуса. Эти права описывают как моральные принципы, так и правовые нормы в муниципальном и международном праве, на которые человек имеет неотъемлемое право как человеческое существо. Поэтому они неотъемлемы, неприкосновенны и универсальны. Они неотъемлемы, поскольку не могут быть никем отняты; неприкосновенны, поскольку не должны нарушаться ни при каких обстоятельствах; универсальны, поскольку применимы везде и во все времена. Права и свободы человека защищены международными конвенциями. На европейском пространстве основополагающим документом является Европейская конвенция о правах человека (ЕКПЧ), которая была разработана в 1950 году Советом Европы и вступила в силу 3 сентября 1953 года. ЕКПЧ закрепляет набор основных прав и свобод, которые должны быть защищены, давая юридическое обязательство придерживаться стандартов поведения, уважающих эти права и свободы.

Методология

В феврале 2020 года мы провели обзор существующего корпуса инструментов мягкого права, связанных с ИИ. Методы обзорного анализа позволяют синтезировать и сопоставить литературу в определенной области исследовательским способом, поэтому они особенно подходят для отбора и оценки сложных или неоднородных областей исследований. Учитывая отсутствие единой базы данных для инструментов мягкого права, мы разработали протокол для обнаружения и фильтрации, адаптированный из рамок предпочтительнее элементы отчетности для систематических обзоров и мета-анализа (Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)). Протокол, который прошел пилотное тестирование и был выверен до сбора данных, состоял из трех последовательных и итеративно связанных этапов: скрининг, оценка приемлемости и анализ содержания. Эта методология призвана обеспечить формальную и основанную на фактических данных процедуру для картирования, мониторинга и итеративной оценки усилий по управлению мягким правом в области ИИ.

I. Скрининг

На этапе отбора мы объединили ретроспективный анализ существующих репозиторий с целенаправленным и неструктурированным поиском в Интернете. Во-первых, мы проверили следующие четыре хранилища данных и текстовые источники для извлечения релевантных записей, связанных с документами по мягкому праву в области ИИ:

Агентство Европейского союза по фундаментальным правам (11 декабря 2019 года), Список политических инициатив в области искусственного интеллекта.

Fjeld & Nagy (январь 2020 года), Принципиальный искусственный интеллект: Картирование консенсуса в этических и основанных на правах подходах к принципам для ИИ, Исследовательская публикация Гарвардского центра Беркмана-Кляйна.

Jobin, Ienca & Vayena (сентябрь 2019 года). Глобальный ландшафт руководящих принципов этики ИИ.

Природный машинный интеллект. Вонг и CASBS (июнь 2019 года). Ландшафт Флюксуса: Экспансивный взгляд на этику и управление ИИ.

В соответствии со структурой Аркси и О'Мэлли для проведения обзорных исследований (12), структурированный поиск в базе данных был дополнен неструктурированным поиском серой литературы для выявления инструментов мягкого права, которые могли ускользнуть от внимания. Записи оценивались на предмет соответствия требованиям (см. 1.2) и, если соответствие требованиям подтверждалось, включались в окончательный синтез вручную и допускались ко второму этапу.

Наконец, мы изучили список «топ-45 компаний ИИ», составленный в мае 2019 года американским журналом Datamation, специализирующимся на анализе технологий. Каждый из 45 субъектов ИИ, включенных в этот список, был проверен независимо путем доступа к их веб-сайтам и поиска этических или политических заявлений ИИ вручную и с помощью поиска по ключевым словам. Наконец, был проведен неструктурированный веб-поиск для получения информации, которая могла остаться незамеченной в рамках нашей стратегии поиска. Соответствующие требованиям записи были вручную включены в окончательный синтез и допущены ко второму этапу.

II. Оценка соответствия квалификационным требованиям

На этапе оценки соответствия требованиям мы проверили все найденные записи, чтобы оценить их соответствие критериям включения в окончательный синтез. Решения о соответствии критериям принимались на основании критериев включения и исключения, перечисленных в Таблице 1.

Таблица 1 – Критерии приемлемости

Скрининг

	Виды: веб-сайты, письменные статьи и другие документы, опубликованные в Интернете, или их части, такие как специальные веб-страницы, сообщения в блогах, институциональные отчеты и декларации, а также ссылки, содержащиеся в них;
Источники включали	Эмитенты: частные коммерческие организации (компании, корпорации, холдинги и т.д., включая объединения частного сектора); академические и исследовательские учреждения (университеты, профессиональные общества, научные фонды и т.д.); национальные правительственные агентства (министерства, органы по защите данных, антимонопольные органы и т.д.); неправительственные организации, включая некоммерческие организации и благотворительные организации.
	Язык: Английский, немецкий, французский, испанский, голландский, итальянский и греческий (языки, на которых говорят исследователи).
Источники исключены:	Типы: видео, изображения и аудио/подкасты (кроме письменных описаний), книги, журналистские статьи, академические статьи, учебные планы, законодательство, официальные стандарты, резюме конференций;
	Эмитенты: межправительственные и наднациональные организации.
	Язык: другие, кроме вышеперечисленных.

Приемлемость	
Источники	которые ссылаются на «искусственный интеллект» и/или «ИИ», либо прямо в своем названии, либо в своем описании (пример: Google: «Принципы ИИ»); или
включены:	которые не содержат вышеуказанной ссылки в своем названии, но вместо этого упоминают «робот», «робототехника», «большие данные», «машинное обучение» и прямо ссылаются на ИИ или искусственный интеллект как часть роботов и/или робототехники; или
	которые не содержат вышеуказанной ссылки в своем названии, но тематически эквивалентны (ссылаясь вместо этого на «алгоритмы», «предиктивную аналитику», «когнитивные вычисления», «машинное обучение», «большие данные», «глубокое обучение», «автономный» или «автоматизированный».
	который описывает принцип, руководящее положение, стандарт (включая «этика/этика», «принципы», «постулаты», «декларация», «политика», «руководящие принципы», «ценности» и т.д.), внутреннюю стратегию (например, создание консультативного совета) или другой вид инициативы.
	которая выражена нормативным или предписывающим языком (т.е. с модальными глаголами или императивами, такими как «ответственный», «справедливый»,

	«доверительный/надежный» и т.д.); или
	которые основаны на принципах или ценностях (т.е. указывают на предпочтение и/или приверженность определенному этическому видению или курсу действий).
	которые ссылаются на действия/решения/обязательства/курсы действий, применимые к субъекту, который их озвучивает, или другим субъектам частного сектора.
	- веб-сайты и документы о робототехнике, в которых искусственный интеллект не упоминается как часть роботов/робототехники; и
Источники исключены:	веб-сайты и документы о данных или этике данных, в которых искусственный интеллект не упоминается как часть данных;
	веб-сайты и документы об этике ИИ, непосредственно предназначенные для субъектов, не относящихся к частному сектору (например, консалтинг для государственного сектора).
	веб-сайты и документы об этике, основной темой которых не является ИИ (например, деловая этика).

III. Анализ содержания

На втором этапе статьи, включенные в окончательный синтез, оценивались с помощью расширенной версии ранее утвержденного протокола контент-анализа, разработанного авторами (13,14). Этот протокол включает как

количественный, так и качественный анализ. На количественном уровне записи классифицировались по типу инструмента, эмитенту и географическому происхождению эмитента. Кроме того, измерялась и визуально отображалась относительная частота соответствующих количественных данных. Наконец, мы провели полнотекстовую проверку с помощью плагина для поиска ключевых слов, чтобы выявить документы, в которых содержались прямые ссылки на права человека. Документы, включенные в эту категорию, содержали прямые ссылки на сохранение и продвижение прав человека или предотвращение их нарушения при проектировании, разработке или развертывании приложений ИИ. Эти документы отличались от тех, в которых права человека не упоминались или упоминались, но без четкого нормативного заявления об их поощрении или недопущении нарушения в контексте ИИ.

На качественном уровне был проведен тематический контент-анализ для выявления повторяющихся тематических моделей, относящихся к следующим областям: (I) этические принципы и ценности и (II) права человека. Тематический контент-анализ проводился вручную исследователями с помощью программного обеспечения для качественного анализа (NVivo/MAXQDA for Mac). Возникающие тематические модели анализировались углубленно, кодировались и группировались в заранее определенные этические категории на основе этической матрицы, разработанной Jobin, Ienca и Vayena (2019).

Учитывая размер окончательной базы данных синтеза, ручной тематический анализ был дополнен автоматизированным анализом с помощью обработки естественного языка (NLP). Мы извлекли документы и веб-содержимое автоматически, где это было возможно, с помощью пакета Python wget, а остальные добавили вручную. Затем мы составили регулярные выражения (15) из кодов, полученных в результате использования протокола качественного анализа, разработанного Jobin, Ienca & Vayena. Регулярные выражения кодов, относящихся к одной теме, были объединены в одно регулярное. Для обеспечения полноты и включения оригинальные английские

коды были переведены на следующие языки: немецкий, французский, испанский, итальянский и голландский. Чтобы определить охват тем, мы проверили наличие регулярных выражений тем в документах, т.е. мы определили, что тема присутствует в руководстве, если регулярное выражение темы имело хотя бы одно совпадение. Наконец, мы сгруппировали результаты по типам членов и нормализовали по общему количеству руководств в каждой группе. Различия между этическими принципами и ценностями, выдвинутыми в 47 государствах-членах Совета Европы, сравнивались, соответственно, с принципами и ценностями, выдвинутыми в государствах-наблюдателях, а также в остальном мире.

IV. Нормативный этический и политический анализ

На четвертом и последнем этапе был проведен эмпирически обоснованный нормативный этический и политический анализ. Целью этого заключительного компонента исследования является перевод предварительных выводов предыдущих этапов исследования с описательного на нормативно-предписывающий уровень. На этом этапе мы выполнили три последовательных теоретических шага. Во-первых, мы оценили результаты контент-анализа, чтобы определить, какие этические принципы и ценности являются наиболее общими и повторяющимися в анализируемом корпусе документов. Поскольку предыдущие исследования показали значительную вариативность интерпретаций в рамках повторяющихся тематических кластеров, мы дополнили оценку относительных тематических частот подробным анализом их интерпретации. Эта оценка сыграла важную роль в определении того, какие интерпретации принципов являются наиболее эффективными и, следовательно, должны быть приняты и реализованы глобальными субъектами. Во-вторых, мы проанализировали данные нашего обзора, чтобы определить, какие принципы и ценности встречаются реже или отсутствуют в текущем ландшафте руководящих принципов этики ИИ. Этот второй шаг сыграл важную роль в выявлении возможных «слепых пятен» в международных инициативах в области мягкого права и, следовательно, в разработке нормативных

рекомендаций по преодолению этих этических пробелов. Наконец, в-третьих, мы выработали нормативные рекомендации по основным этическим принципам и ценностям, которые требуют приоритетности в международном управлении ИИ. Эта заключительная часть сыграла важную роль в формировании будущих нормативных этических рамок и составлении дорожной карты для международной политики в области ИИ, этики и прав человека. С этой целью мы представили удобное для читателя визуальное резюме результатов исследования и инструментарий для будущего мониторинга и оценки (например, показатели) на стыке ИИ, этики и прав человека.

VI. Выводы

Наш поиск выявил 116 документов, содержащих документы «мягкого права» по ИИ, выпущенные неправительственными организациями до февраля 2020 года. Данные показывают значительный рост числа публикаций с течением времени: 93,9% были выпущены с 2016 года. Пик количества опубликованных на международном уровне документов «мягкого права» был достигнут в 2018 году, а в последующем году произошло незначительное снижение (см. Рисунок 1).

Изменения во времени в публикации документов «мягкого права» по ИИ



45 40 35 30 25 20

Количество публикаций

Рисунок 1 – Изменения во времени в публикации документов «мягкого права» по ИИ

Разбивка данных по типу издающей организации показывает, что большинство документов было выпущено правительственными агентствами ($n=39$), за ними следуют частные компании и объединения частного сектора ($n=36$), академические и исследовательские учреждения, включая научные фонды, профессиональные общества и исследовательские объединения ($n=28$), а также неправительственные организации (НПО), включая некоммерческие организации (НКО) и благотворительные организации ($n=13$). Подробное распределение организаций-эмитентов по типам представлено на Рисунке 2.

Разбивка данных по географическому распределению организаций, выпускающих документы, показывает, что 46% ($n=53,5$) документов «мягкого права» выпущены организациями, расположенными в странах-членах Совета Европы. 32% ($n=37,5$) – организациями, расположенными в странах-наблюдателях Совета Европы. 21% ($n=25$) – организациями, расположенными в странах, которые не являются ни членами, ни наблюдателями Совета Европы. В целом, данные свидетельствуют о значительном представительстве организаций, занимающихся выдачей сертификатов, в экономически развитых странах, причем на США ($n = 29,5$; 25,2%) и Великобританию ($n = 17,5$; 16%) в совокупности приходится более трети всех этических принципов ИИ. Другие страны включают, в порядке убывания, Германию ($n = 8$), Японию ($n = 6$), Финляндию ($n = 4$), Бельгию, Китай, Францию и Нидерланды ($n = 3$), Индию, Италию, Сингапур и Испанию ($n = 2$), Австралию, Австрию, Чешскую Республику, Исландию, Литву, Мальту, Мексику, Новую Зеландию, Норвегию, Россию, Южную Корею, Швецию, Швейцарию, ОАЭ и Ватикан ($n = 1$). Тринадцать документов были выпущены международными организациями или организациями, которые не могут быть приписаны к какой-либо конкретной стране. Африканские и южно-американские страны не представлены независимо от международных организаций. Визуальный обзор географического распределения организаций, выдавших документы, представлен на рисунке 3:

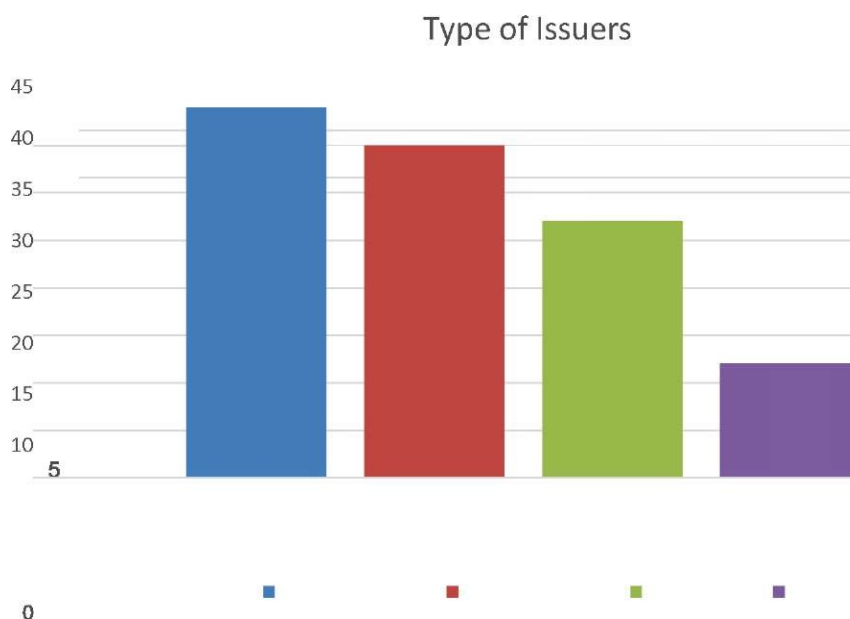


Рисунок 2 – Типы организаций-эмитентов

Более половины документов ($n=62$) содержат прямую ссылку на продвижение, уважение или предотвращение нарушения прав человека. Из этих 31 документ издан организациями, расположенными в странах-членах Совета Европы, 14 – организациями, расположенными в странах-наблюдателях, и 17 – в странах-наблюдателях, не являющихся членами Совета Европы. Документы, изданные организациями, расположенными в странах-членах Совета Европы, содержат ссылки на права человека в 57,9% случаев. Документы из стран, не являющихся членами Совета Европы, содержат ссылки на права человека в 49,6% случаев. Это показывает, что последствия искусственного интеллекта для прав человека чаще затрагиваются организациями, расположенными в странах-членах Совета Европы, по сравнению с остальным миром.

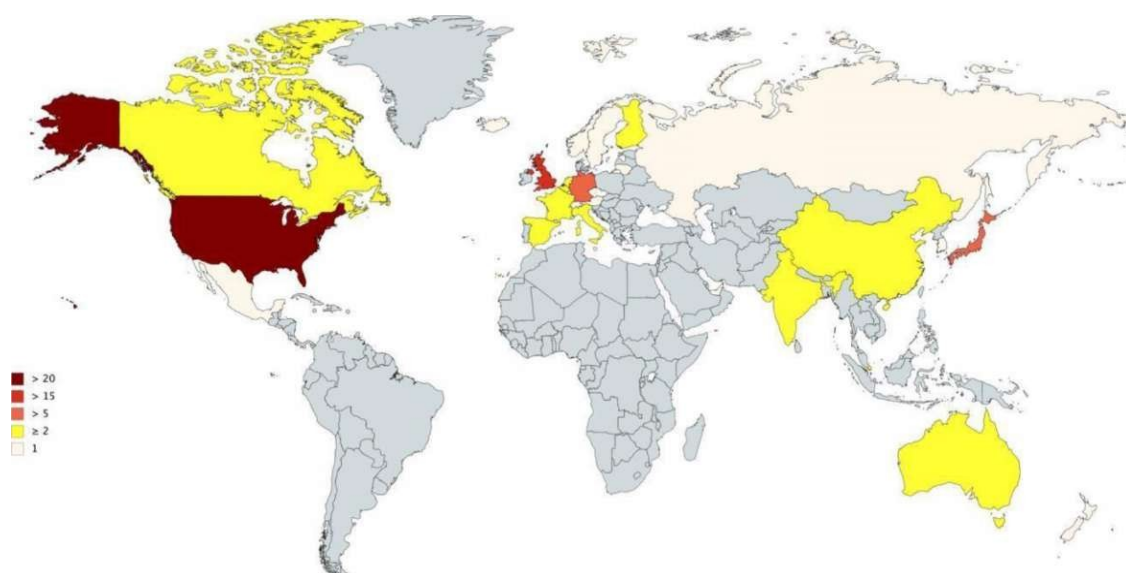


Рисунок 3 – Географическое распределение документов «мягкого права» по стране организации, выдавшей документ

В результате тематического контент-анализа было получено множество этически значимых кодов, которые можно последовательно отнести к одиннадцати общим этическим кластерам, определенным Jobin, Ienca & Vayena (2019)¹³. К ним относятся, в порядке убывания частоты встречаемости источников, в которых они фигурировали: прозрачность, справедливость и честность, нематериальность, ответственность, конфиденциальность, благотворность, свобода и автономия, доверие, достоинство, устойчивость и солидарность. Подробное частотное представление этических принципов и соответствующих кодексов представлено в Таблице 2.

Этический принцип	Количество документов	Включенные коды
Прозрачность	101/116	Прозрачность, объяснимость,
		объяснимость, понятность,
		интерпретируемость,
		коммуникация,
		раскрытие, демонстрация
		Справедливость, честность,

		последовательность,
		вовлеченность, равенство, справедливость, (не)предвзятость,
		дискриминация, разнообразие,
Правосудие и справедливость	97/116	множественность, доступность, обратимость,
		средство правовой защиты, возмещение, вызов, доступ
		и распространение, беспристрастность
		Нематериальность, надежность, безопасность,
		вред, защита, предосторожность,
Нематериальная ответственность	84/116	предотвращение, неприкосновенность (телесная или психическая), отказ от подрывной деятельности
		Ответственность, подотчетность,
Ответственность	79/116	ответственность, действовать добросовестно
		Конфиденциальность, личная или частная
Конфиденциальность	74/116	информация, конфиденциальность

		Выгоды, благотворительность, благополучие, мир, социальное благо, общее благо
Благотворительность	58/116	Благо, благодеяние, благополучие, мир, социальное благо, общее благо
		Свобода, автономия, согласие, выбор, самоопределение, свобода, расширение возможностей
Свобода и автономия	48/116	
Доверие	41/116	Доверие, благонадежность
		Устойчивость, окружающая среда (природа), энергия, ресурсы (энергия)
Устойчивое развитие	20/116	
Достоинства	20/116	Достоинства
Солидарность	10/116	Солидарность, социальное обеспечение, сплоченность

Ни один этический принцип не является общим для всего корпуса документов, хотя наметилось сближение вокруг следующих принципов: прозрачность, справедливость и честность, незлорадство, ответственность и конфиденциальность. Эти принципы упоминаются почти в двух третях всех источников. Тем не менее, дальнейший тематический анализ показывает, что сохраняются значительные семантические и концептуальные расхождения как

в трактовке 11 этических принципов, так и в конкретных рекомендациях или проблемных областях, вытекающих из каждого из них.

Таблица 2 – Частота этических тем и соответствующих кодексов

Регулярные выражения, построенные на основе кодов, показывают значительные различия в охвате тем в документах, подготовленных в странах-членах Совета Европы (СЕ), по сравнению с документами, подготовленными в других странах. По сравнению с документами, подготовленными в странах-наблюдателях СЕ, документы по мягкому праву, подготовленные в странах-членах Совета Европы, подчеркивают следующие этические принципы: прозрачность, устойчивость, свобода и автономия, доверие/надежность и солидарность (см. Рисунок 4). Напротив, принципы справедливости, благодеяния и достоинства упоминаются в них более спорадически. По сравнению с документами, подготовленными в остальном мире (страны-наблюдатели, не являющиеся членами Совета Европы), документы по мягкому праву, подготовленные в странах-членах Совета Европы, подчеркивают принципы доверия/надежности и солидарности, в то время как все остальные принципы затрагиваются реже. Принципы конфиденциальности, справедливости и честности продемонстрировали наименьшую вариативность, а значит, и самую высокую степень межгеографической и межкультурной стабильности.

Подробная тематическая оценка перечисленного представлена ниже.

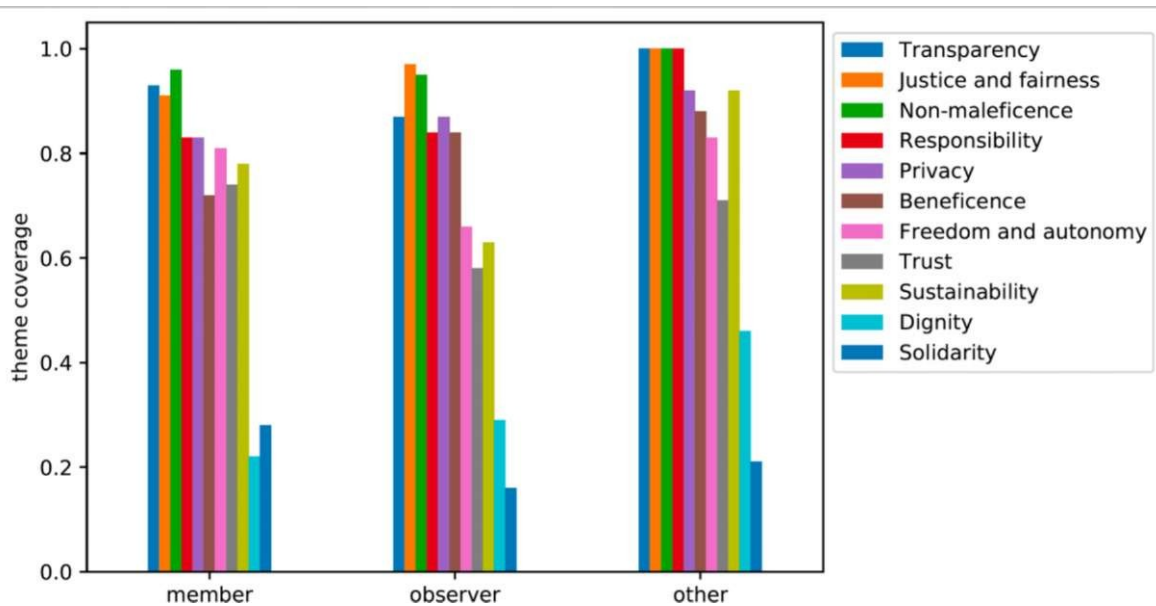


Рисунок 4 – Различия в охвате тем в документах, произведенных в странах-членах Совета Европы (СЕ), по сравнению с документами, произведенными в остальном мире.

Прозрачность: Прозрачность, представленная в 101 из 116 источников, является наиболее распространенным этическим принципом в современном спектре «мягкого права». Тематический анализ выявил значительные различия в интерпретации и обосновании призывов к прозрачности. Эти различия приводят к очевидным расхождениям в стратегиях реализации, предлагаемых для достижения прозрачности в отношении ИИ. Ссылки на прозрачность можно сгруппировать в два основных тематических семейства: (1) прозрачность алгоритмов и методов обработки данных, (2) прозрачность человеческих практик, связанных с проектированием, разработкой и внедрением систем ИИ. Призывы к прозрачности первого типа обычно связаны с продвижением методологических подходов к «объяснимому ИИ», то есть к системам ИИ, результаты и решения которых могут быть понятны людям-экспертам. Эти методы и технологии контрастируют с подходами «черного ящика» к машинному обучению, когда шаги, с помощью которых система ИИ пришла к определенному решению, непонятны человеческим экспертам, включая разработчиков системы. В то время как частные компании, особенно частные субъекты ИИ, стремятся свести прозрачность к интерпретируемости и

объяснимости с помощью технических решений, таких как, в частности, послойное распространение релевантности (LRP) и локальная интерпретируемость, государственные органы, такие как национальные органы по защите данных, подчеркивают важность таких методов надзора, как аудит. Призывы к прозрачности второго типа сосредоточены не на интерпретируемых алгоритмах, а на прозрачности человеческих практик, связанных с данными и ИИ, таких как раскрытие соответствующей информации субъектам данных, избежание секретности при внедрении стратегий ИИ и запрет конфликта интересов между субъектами ИИ и надзорными органами. Призывы к прозрачности такого рода более распространены среди правительственных субъектов и НПО.

Справедливость, честность и равноправие: Справедливость в основном выражается в терминах честности и предотвращения (или смягчения) алгоритмических предубеждений, которые могут привести к дискриминации. Опасения, что ИИ может усилить неравенство и вызвать дискриминацию, встречаются реже в документах «мягкого права», изданных в частном секторе, по сравнению с правительственными органами и научными кругами. Документы расходятся во мнениях относительно того, как достичь справедливости и честности в ИИ. Некоторые источники делают акцент на уважении разнообразия и предпочтении инклюзивности и равенства как при разработке систем ИИ (особенно при составлении обучающих наборов данных), так и при их внедрении в общество. Другие источники призывают к возможности обжалования или опротестования решений, основывая это на праве на возмещение ущерба и средства правовой защиты. Справедливый доступ к преимуществам ИИ также является часто повторяющейся темой. В документах, изданных правительственными структурами, особое внимание уделяется влиянию ИИ на рынок труда, а также необходимости решения демократических или общественных проблем. Мы определили пять основных, не исключających друг друга стратегий реализации для сохранения и продвижения справедливости и честности в ИИ:

- через технические решения, такие как стандарты и лучшие практики;
- путем повышения осведомленности общественности о существующих права и регулировании;
- путем улучшения тестирования, мониторинга и аудита систем ИИ;
- путем развития или укрепления верховенства закона и права на апелляцию, обращение, возмещение или средство правовой защиты;
- посредством системных изменений и процессов, таких как действия и надзор со стороны правительства, более междисциплинарный персонал, а также более активное вовлечение гражданского общества или других соответствующих заинтересованных сторон в интерактивном режиме.

В то время как решения II-V оказались предпочтительными среди государственных учреждений (особенно сотрудников по защите данных), решения типа I оказались более распространенными среди частных субъектов ИИ.

Нематериальность: Упоминания о небезразличии встречаются значительно чаще, чем упоминания о благодеянии, и включают в себя общие призывы к обеспечению безопасности и защиты или утверждения о том, что ИИ никогда не должен причинять предвидимый или непреднамеренный вред. Некоторые документы фокусируются на конкретных рисках или потенциальном вреде, особенно на риске преднамеренного злоупотребления посредством кибервойн и злонамеренного взлома. Наиболее распространенными источниками вреда, упоминаемыми в документах, являются социальная дискриминация, нарушение неприкосновенности частной жизни, телесные или психологические повреждения. Документы «мягкого права», посвященные снижению вреда, часто призывают как к техническим решениям, так и к обязательному вмешательству в управление на уровне исследований, проектирования, разработки и внедрения технологий ИИ. Технические решения включают в себя встроенные оценки качества данных или основы безопасности и конфиденциальности при проектировании, хотя другие выступают за установление отраслевых стандартов. Предлагаемые

стратегии управления включают активное сотрудничество между различными дисциплинами и заинтересованными сторонами, соблюдение существующего или нового законодательства, а также необходимость создания надзорных процессов и практик, в частности тестов, мониторинга, аудита и оценки со стороны внутренних подразделений, клиентов, пользователей, независимых третьих сторон или государственных структур. В некоторых источниках прямо упоминается совместное использование в военных целях – так называемая проблема двойного применения – как основная область внедрения ИИ, требующая вмешательства руководства.

Ответственность и подотчетность: Ссылки на разработку «ответственного ИИ» широко распространены. Тем не менее, понятия ответственности и подотчетности определяются редко. Различные субъекты называются ответственными и подотчетными за действия и решения ИИ. К ним относятся разработчики ИИ, дизайнеры и весь промышленный сектор. В дальнейшем возникли разногласия по поводу того, должен ли ИИ нести ответственность подобно человеку или же люди всегда должны быть единственными субъектами, несущими конечную ответственность за технологические артефакты.

Конфиденциальность: Конфиденциальность широко рассматривается как ценность, которую необходимо отстаивать, и как право, которое необходимо защищать. Хотя вопросы конфиденциальности часто затрагиваются в действующих руководствах по ИИ, нет единого мнения о том, какие уникальные проблемы, если таковые имеются, возникают в связи с развитием ИИ по сравнению с другими технологиями, использующими большие объемы данных. Тематический анализ показывает, что в большинстве документов вопросы конфиденциальности рассматриваются в общих чертах, без установления четкой связи между возможностями ИИ и новыми проблемами конфиденциальности. Несмотря на то, что проблема конфиденциальности ИИ слабо охарактеризована, она часто представляется в связи с вопросами защиты данных и безопасности данных. Предлагаемые стратегии сохранения

конфиденциальности в ИИ можно разделить на три категории: (А) технические решения, такие как дифференциальная конфиденциальность, безопасные многосторонние вычисления и гомоморфное шифрование; (В) решения по привлечению общественности, такие как повышение осведомленности пользователей и субъектов данных, и (С) решения по нормативно-правовым подходам, такие как более четкое определение требований к соблюдению законодательства (особенно в области защиты данных) или даже создание новых законов и правил для учета уникальности ИИ.

Благодеяние: хотя содействие благу (*beneficence* в этических терминах) часто упоминается, оно редко определяется, хотя заметные исключения упоминают содействие благополучию и процветанию человека, миру и счастью, создание социально-экономических возможностей и содействие экономическому процветанию. Аналогичная неопределенность касается субъектов, которые должны получить выгоду от ИИ: частные эмитенты склонны подчеркивать выгоду от ИИ для клиентов, в то время как академические и правительственные источники обычно утверждают, что ИИ должен принести пользу «всем», «человечеству» и «обществу в целом». Стратегии продвижения блага включают согласование ИИ с человеческими ценностями, минимизацию концентрации власти и использование возможностей ИИ для продвижения прав человека.

Свобода и автономия: Документы «мягкого права» связывают ИИ с сохранением или поощрением ряда свобод и свободолюбия. К ним, в частности, относятся свобода выражения мнений, информационное самоопределение, право на частную жизнь и личную автономию. Последнее понятие обычно называют позитивной свободой, в частности, свободой процветать, принимать самостоятельные решения и самоопределять свой собственный курс действий. Однако в меньшинстве документов автономия рассматривается как негативная свобода, например, свобода от технологических экспериментов, манипуляций или наблюдения. Предлагаемые решения для сохранения свободы и автономии в ИИ включают обеспечение прозрачности и объяснимости ИИ, повышение

грамотности в области ИИ, обеспечение информированного согласия или, наоборот, активное воздержание от сбора и распространения данных в отсутствие информированного согласия.

Доверие и надежность: чуть более одного из трех документов «мягкого права» призывают к доверительным исследованиям и технологиям ИИ или к продвижению культуры доверия среди ученых и инженеров. Однако некоторые документы прямо предостерегают от чрезмерного доверия к ИИ, утверждая, что доверие может возникнуть только среди равных и не должно быть делегировано ИИ. Предложения по созданию или поддержанию доверия включают образование, надежность, подотчетность, процессы мониторинга и оценки целостности систем ИИ с течением времени, а также инструменты и методы, обеспечивающие соблюдение норм и стандартов.

Устойчивость: Устойчивость упоминается спорадически, обычно в связи с защитой окружающей среды или даже улучшением экосистемы и биоразнообразия планеты. Некоторые документы требуют, чтобы системы ИИ устойчиво обрабатывали данные и повышали свою энергоэффективность для минимизации экологического воздействия (4,7). Меньшая часть документов посвящена социальной устойчивости, то есть обеспечению ответственности в связи с потенциальной потерей рабочих мест и расширению возможностей для инноваций.

Достоинство: хотя понятие достоинства остается неопределенным в существующих руководствах, документы «мягкого права» указывают, что оно является прерогативой людей, но не роботов. Ссылки на достоинство тесно переплетаются с защитой и продвижением прав человека. Утверждается, что ИИ должен не уменьшать или уничтожать, а уважать, сохранять или даже увеличивать человеческое достоинство. Считается, что достоинство будет сохранено, если оно будет уважаться разработчиками ИИ в первую очередь и поощряться через новое законодательство, через инициативы по управлению или через технические и методологические руководства, изданные правительством.

Солидарность: Солидарность — это наименее повторяющаяся этическая тема, которая в основном упоминается в связи с последствиями ИИ для рынка труда. Источники призывают к усилению системы социальной защиты, чтобы справиться с долгосрочными последствиями ИИ для человеческого труда. Они подчеркивают необходимость перераспределения выгод от ИИ, чтобы не угрожать социальной сплоченности (6,5) и уважать потенциально уязвимые лица и группы. И наконец, содержится предупреждение о том, что практика сбора и обработки данных, ориентированная на отдельных людей, может подорвать солидарность в пользу «радикального индивидуализма».

I. Ограничения

Данное исследование имеет несколько ограничений. Во-первых, с библиографической точки зрения, руководства и мягкие политические документы являются примером серой литературы, поэтому они не индексируются в обычных научных базах данных. Поэтому их поиск неизбежно менее воспроизводим и объективен по сравнению с систематическим поиском в базах данных рецензируемой литературы. В соответствии с передовой практикой обзора серой литературы, это ограничение было смягчено путем разработки протокола обнаружения и отбора, который прошел пилотное тестирование перед сбором данных. Хотя результаты поиска в поисковых системах персонализированы, риск влияния персонализации на обнаружение был снижен благодаря широте поиска по ключевым словам и включению результатов. Языковая предвзятость могла перекосить наш корпус в сторону англоязычных результатов. Мы минимизировали это ограничение, включив записи, написанные на следующих языках (помимо английского): немецком, французском, итальянском, испанском и голландском. Ключевые слова и коды на перечисленных языках были переведены на английский и включены в анализ. В нашем контент-анализе присутствуют типичные ограничения качественных методов анализа. Следуя лучшим практикам контент-анализа, это ограничение было смягчено путем разработки

индуктивной стратегии кодирования, которая проводилась независимо двумя рецензентами для минимизации субъективной предвзятости. Наконец, учитывая быстрый темп публикации руководящих документов по ИИ, существует вероятность того, что после завершения нашего поиска были опубликованы новые программные документы. Чтобы минимизировать этот риск, параллельно с анализом данных и до 1 марта 2020 года проводился постоянный мониторинг литературы.

II. Обсуждение и нормативно-этический анализ

Мы обнаружили стремительный рост числа и разнообразия документов «мягкого права» по ИИ, что свидетельствует о все более активном участии международного сообщества в обязательном управлении в этой технологической области. Организации, выпускающие руководства, принципы и другие документы «мягкого права» по ИИ, представляют широкий спектр секторов. В частности, почти эквивалентная доля документов, выпущенных государством (т.е. правительственными организациями) и частным сектором (компаниями и объединениями частного сектора), указывает на то, что этические проблемы ИИ касаются как государственных структур, так и частных предприятий. Однако решения, предлагаемые для решения этических проблем ИИ, значительно отличаются друг от друга, причем государственные субъекты отдают предпочтение техническим решениям, таким как объяснимый и интерпретируемый ИИ, а не обязательному регулированию и углубленному этическому анализу. Кроме того, относительная недопредставленность таких географических регионов, как Африка и Южная Америка, указывает на то, что международные дебаты по этическим аспектам ИИ, возможно, ведутся неравномерно. Более экономически развитые страны (MEDC) формируют эти дебаты в большей степени, чем другие, что вызывает опасения по поводу пренебрежения местными знаниями, культурным плюрализмом и глобальной справедливостью. Эти результаты подтверждают неравномерное географическое представительство и распределение субъектов этики ИИ,

отмеченное в предыдущих исследованиях. Однако по сравнению с предыдущими исследованиями наш обзор показывает, что в международном необязательном управлении теперь участвуют новые субъекты из ранее непредставленных стран. К ним относятся субъекты из сверхдержав ИИ, то есть стран-лидеров в области ИИ, таких как Китай, а также стран со средним уровнем дохода из ранее непредставленных регионов мира, таких как Россия и Мексика.

Распространение усилий в области «мягкого права» можно интерпретировать как реакцию руководства на передовые исследования в области ИИ, объем исследований и объем рынка которого резко возросли в последние годы (16). Наш анализ показывает появление очевидного сближения заинтересованных сторон в продвижении этических принципов прозрачности, справедливости, непричинения вреда, ответственности и конфиденциальности. Тем не менее, наш тематический анализ выявляет существенные расхождения в отношении четырех основных факторов: (I) как интерпретируются этические принципы, (II) почему они считаются важными, (III) к какому вопросу, области или субъектам они относятся, и (IV) как они должны быть реализованы. Кроме того, остается неясным, какие этические принципы должны быть приоритетными, как должны разрешаться конфликты между этическими принципами, кто должен осуществлять этический надзор за ИИ и как исследователи и учреждения могут соблюдать разработанные рекомендации. Эти выводы свидетельствуют о существовании пробела на стыке формулирования принципов и их реализации на практике, который вряд ли может быть решен с помощью технических знаний или подходов «сверху вниз».

Несмотря на то, что ни один этический принцип прямо не поддерживается всеми существующими руководствами, прозрачность, справедливость и честность, небезвредность, ответственность и конфиденциальность упоминаются более чем в половине всех руководств. Такое внимание может свидетельствовать о развивающейся конвергенции

этических принципов ИИ вокруг этих принципов в глобальном политическом ландшафте. В частности, преобладание призывов к прозрачности, справедливости и честности указывает на зарождающийся моральный приоритет, требующий прозрачных процессов на всем континууме ИИ (от прозрачности разработки и проектирования алгоритмов до прозрачной практики использования ИИ), а также предостерегающий мировое сообщество от риска того, что ИИ может усилить неравенство, если соображения справедливости и честности не будут должным образом учтены. Обе эти темы тесно переплетаются с темой ответственности, поскольку продвижение как прозрачности, так и справедливости, по-видимому, предполагает повышение ответственности и подотчетности со стороны создателей и внедрителей ИИ.

Утверждается, что прозрачность не является этическим принципом как таковым, а скорее «проэтическим условием для обеспечения или ослабления других этических практик или принципов». Такая характеристика прозрачности как проэтического условия для других принципов прослеживается в Декларации соответствия поставщика (SDoC) компании IBM, которая помогает предоставить информацию о четырех ключевых столпах надежного ИИ. Предположительно проэтический характер прозрачности может отчасти объяснить ее большую распространенность по сравнению с другими этическими принципами. Примечательно, что в существующих руководствах большое значение придается поощрению ответственности и подотчетности, однако лишь немногие из них подчеркивают обязанность всех заинтересованных сторон, участвующих в разработке и внедрении ИИ, действовать добросовестно. Это несоответствие, вероятно, связано с тем, что в существующих руководствах не удается установить полное соответствие между принципами и реализуемыми требованиями, при этом некоторые принципы остаются неохарактеризованными или оторванными от требований, необходимых для их реализации.

Поскольку кодексы, относящиеся к не-малефициции, превышают кодексы, относящиеся к бенефициции, похоже, что для нынешнего

сообщества ИИ моральные обязательства по предотвращению вреда превалируют над поощрением добра. Этот факт можно отчасти интерпретировать как проявление так называемого негативистского предубеждения, т.е. общего когнитивного предубеждения придавать больший вес негативным сущностям, гипотезы, которую подчеркнул когнитивный психолог Стивен Пинкер в недавнем углубленном анализе Группы научного прогнозирования (STOA) Европейского парламента. Эта негативная характеристика этических ценностей еще больше подчеркивается тем фактом, что существующие руководства в основном сосредоточены на том, как сохранить неприкосновенность частной жизни, достоинство, автономию и свободу личности, несмотря на достижения в области ИИ, при этом в значительной степени игнорируя вопрос о том, можно ли активно продвигать эти принципы посредством ответственных инноваций в области ИИ.

Вопрос доверия к ИИ, хотя и затрагивается менее чем в трети всех источников, решает важнейшую этическую дилемму в управлении ИИ: определить, является ли морально желательным укрепление общественного доверия к ИИ. Хотя некоторые источники, особенно те, которые были подготовлены в частном секторе, подчеркивают важность укрепления доверия к ИИ посредством образовательной и просветительской деятельности, меньшее число источников утверждает, что доверие к ИИ может фактически ослабить контроль и подорвать некоторые общественные обязательства производителей ИИ. Такая возможность ставит под сомнение доминирующее в этике ИИ мнение о том, что укрепление общественного доверия к ИИ является фундаментальным требованием для этического управления. В связи с доверием мы отметили дополнительные концептуальные проблемы. Во-первых, концептуальная ясность в отношении значения и динамики доверия, по-видимому, отсутствует во всех существующих документах. Большинство источников не указали доверителя и доверенное лицо в описываемых ими доверительных отношениях, следовательно, пренебрегли тем, что «доверие» — это реляционное и очень сложное отношение, в котором участвуют, по крайней

мере два субъекта, доверяющие друг другу делать или не делать определенную деятельность. На эти отношения влияет широкий спектр факторов, например, культура, системы убеждений, контекст, а также черты характера участников доверительных отношений. Эти контекстуальные факторы, как представляется, игнорируются в текущей литературе. Самое главное, что черта «благонадежность» и реляционная конструкция «доверие» часто смешивались или использовались как взаимозаменяемые субъектами ИИ, которых мы рассмотрели. Такое смешение не только приводит к концептуальной путанице, но и может породить ложные надежды среди пользователей ИИ и политиков. Доверие и благонадежность – это разные понятия, и благонадежность сама по себе не ведет к доверительным отношениям. Дальнейшая работа по управлению в этой области должна прояснить это важнейшее концептуальное различие и потребовать большего уточнения требований к доверительным отношениям.

Относительная недостаточная представленность тем устойчивости и солидарности говорит о том, что эти темы, возможно, в настоящее время остаются вне поля зрения основного этического дискурса по ИИ. Недопредставленность принципов, связанных с устойчивостью, особенно проблематична в свете того, что развертывание ИИ требует огромных вычислительных ресурсов, которые, в свою очередь, требуют большого потребления энергии. Однако воздействие ИИ на окружающую среду связано не только с негативными последствиями цифровых инфраструктур, оказывающих большое влияние на окружающую среду, но и с возможностью использования ИИ на благо экосистем и всей биосферы. Этот последний момент, отмеченный в докладе Всемирного экономического форума, хотя и не отраженный в руководящих принципах ИИ, разработанных той же организацией, требует более широкого одобрения, чтобы закрепиться в этическом повествовании об ИИ. Этический принцип солидарности упоминается редко, обычно в связи с разработкой инклюзивных стратегий по предотвращению потери рабочих мест и несправедливого распределения

бремени. Мало внимания уделяется продвижению солидарности через появляющуюся возможность использования опыта ИИ для решения гуманитарных проблем – задача, которой в настоящее время занимаются, в частности, межправительственные организации, такие как Управление ООН по обслуживанию проектов (ЮНОПС) или Всемирная организация здравоохранения (ВОЗ), а также частные компании, такие как Microsoft. Поскольку гуманитарные издержки антропогенного изменения климата быстро растут, принципы устойчивости и солидарности представляются строго взаимосвязанными, хотя и слабо представленными по сравнению с другими принципами.

Хотя числовые данные свидетельствуют о наметившемся сближении позиций по продвижению некоторых этических принципов, углубленный тематический анализ рисует более сложную картину, поскольку существуют критические различия в том, как интерпретируются эти принципы, а также какие требования считаются необходимыми для их реализации. Результаты показывают, что для практического достижения этического ИИ предлагаются различные и зачастую противоречивые меры. Например, потребность во все более крупных и разнообразных массивах данных для «беспристрастного» ИИ трудно совместить с требованием предоставить людям больший контроль над их данными и их использованием, чтобы уважать их частную жизнь и автономию. Аналогичные контрасты возникают между требованием избегать вреда любой ценой и требованием сбалансировать риски и выгоды. Кроме того, следует отметить, что оценка рисков и выгод приведет к различным результатам в зависимости от того, в интересах чьего благополучия она будет оптимизирована теми или иными субъектами. Если не решить эту проблему, такие расхождения и противоречия могут подорвать попытки разработать глобальную повестку дня для этического ИИ.

Несмотря на общее согласие с тем, что ИИ должен быть этическим, внутри и между руководствами по этическому ИИ возникают значительные расхождения. Кроме того, сохраняется неопределенность в отношении того, как

этические принципы и руководства должны быть реализованы. Эти проблемы имеют последствия для научной политики, управления технологиями и этики исследований. На уровне политики они призывают к активизации совместных усилий правительственных организаций по согласованию и определению приоритетов их программ в области ИИ, причем эти усилия могут быть предприняты при посредничестве и содействии межправительственных организаций. Хотя гармонизация желательна, однако она не должна происходить ценой уничтожения культурного и морального плюрализма в отношении ИИ. Поэтому основной задачей при разработке глобальной повестки дня в области ИИ является достижение баланса между необходимостью межнациональной гармонизации и уважением к культурному разнообразию и моральному плюрализму. Эта задача потребует разработки совещательных механизмов для урегулирования разногласий относительно ценностей и последствий достижений ИИ между различными заинтересованными сторонами из разных регионов мира. На уровне управления технологиями гармонизация, как правило, осуществляется в виде стандартизации. Усилия в этом направлении были предприняты, в частности, Институтом инженеров по электротехнике и электронике (IEEE) в рамках инициативы «Этически согласованная разработка». Наконец, для оценки этической обоснованности применения ИИ в научных исследованиях, особенно в академической сфере, все чаще будут требоваться «мягкие» механизмы управления, такие как независимые наблюдательные советы (IRB). Однако применение ИИ правительствами или частными корпорациями вряд ли попадет под их надзор, если только не будут существенно расширены полномочия IRB.

В целом, наши выводы показывают, что международное сообщество не пришло к единому мнению относительно того, что представляет собой этический ИИ и какие требования необходимы для его достижения. Тем не менее, заметны признаки сближения позиций вокруг понятий прозрачности, незлонамеренности, ответственности и конфиденциальности. Обогащение нынешнего этического дискурса ИИ за счет более глубокой оценки важнейших,

но недостаточно представленных этических принципов, таких как человеческое достоинство, солидарность и устойчивость, вероятно, приведет к созданию более четко сформулированного этического ландшафта для ИИ. Кроме того, желательно сместить акцент с формулирования принципов на их практическое воплощение. Глобальная повестка дня для этического ИИ должна балансировать между необходимостью межнациональной и междоменной гармонизации и уважением к культурному разнообразию и моральному плюрализму. В целом, наш обзор обеспечивает полезную отправную точку для понимания присущего разнообразия существующих принципов и руководств по этическому ИИ и очерчивает задачи, стоящие перед мировым сообществом.

III. Последствия для политики

Множество международных усилий по разработке документов «мягкого права» по ИИ предоставляет ценную косвенную информацию о том, как человечество будет реагировать на многочисленные проблемы управления, создаваемые ИИ. Международное сообщество, похоже, сходится во мнении о важности прозрачности, отсутствия злоупотреблений, ответственности и конфиденциальности для разработки и внедрения этичного ИИ. Однако обогащение нынешнего этического дискурса ИИ за счет более глубокой оценки важнейших, но недостаточно представленных этических принципов, таких как человеческое достоинство, солидарность и устойчивость, вероятно, приведет к созданию более четко сформулированного этического ландшафта искусственного интеллекта. Кроме того, следующим шагом должно стать смещение акцента с формулирования принципов на их практическое воплощение. Глобальная повестка дня для этического ИИ должна уравнивать необходимость межнациональной и междоменной гармонизации с уважением к культурному разнообразию и моральному плюрализму.

Эти выводы имеют последствия для государственной политики, управления технологиями и этики исследований. На уровне политики

необходимо более тесное сотрудничество между заинтересованными сторонами для взаимного согласования различных этических программ в области ИИ и достижения процедурного сближения не только этических принципов, но и их реализации. Хотя глобальный консенсус может быть желателен, он не должен достигаться ценой уничтожения культурного и морального плюрализма и может потребовать разработки совещательных механизмов для урегулирования разногласий между заинтересованными сторонами из разных регионов мира. Такие усилия могут быть предприняты при посредничестве и содействии межправительственных организаций, таких как Совет Европы. Кроме того, они могут быть дополнены подходами «снизу вверх» с участием всех заинтересованных сторон на равных условиях.

Вмешательство в политику в этой области должно прояснить, как этические рекомендации по ИИ соотносятся с существующим национальным и международным регулированием. Несмотря на предполагаемую социотехническую уникальность ИИ, документы «мягкого права» по ИИ не действуют в этико-правовом вакууме. Напротив, этические руководства и другие инструменты «мягкого права» в конечном итоге должны будут действовать в контексте, уже насыщенном правилами, включая жесткое право (обязательное управление). Если не учитывать контекст этих правил, это может подорвать внедрение принципов в действенное и эффективное международное управление. Примером тому может служить прозрачность – наиболее часто повторяющийся этический принцип. Несмотря на то, что он часто встречается, принцип прозрачности, как правило, упоминается без прямой связи с основополагающим обязательным нормативным актом. Сегодня учреждения, использующие технологии ИИ, уже подчиняются многочисленным правилам прозрачности в рамках существующих правовых систем, таких как Закон о справедливом кредитном информировании в США и конкретные практические требования к контролерам и обработчикам данных, изложенные в статьях (12-14). Общего регламента ЕС по защите данных (GDPR). Аналогичным образом,

уточнение различий между «доверием» и «надежностью» является важной задачей для политиков.

Помимо интеграции жесткого и мягкого права, дополнительной проблемой является воплощение этических принципов на практике и поиск гармонизации между различными этическими кодексами ИИ. На уровне управления технологиями многообещающие попытки гармонизации были предприняты в рамках инициатив по стандартизации, таких как инициативы Института инженеров по электротехнике и электронике (IEEE), т.е. крупнейшей в мире профессиональной технической организации, занимающейся продвижением технологических инноваций. IEEE прилагает усилия по этике ИИ как для автономных и интеллектуальных систем общего назначения в рамках инициативы «Ethically Aligned Design» (26), так и для конкретных областей, таких как «Дорожная карта стандартов нейротехнологий для интерфейса мозг-машина», разработанная Ассоциацией стандартов IEEE.

Еще одно политическое последствие касается надзора за исследованиями. Механизмы исследовательской этики, такие как независимые наблюдательные советы (IRBs), будут все чаще требоваться для оценки этической обоснованности применения ИИ в научных исследованиях, особенно в академической сфере. Однако применение ИИ правительствами или частными корпорациями вряд ли попадет под их надзор, если только не будут существенно расширены полномочия IRB.

В целом, тематическое разнообразие и информационная насыщенность проанализированных нами документов свидетельствует о том, что инструменты «мягкого права», изданные правительственными и неправительственными организациями (включая частные компании и академические организации), являются полезными инструментами для оказания практического влияния на принятие государственных решений в области ИИ. При адекватной концептуализации, разработке и составлении инициатив «мягкого права» они обладают потенциалом для направления развития систем ИИ на благо общества и в соответствии с этическими ценностями и правовыми нормами. Однако

подходы «мягкого права» не следует считать заменой обязательного управления. Усилия по саморегулированию, предпринимаемые частными субъектами ИИ, подвержены особому риску продвижения в обход или вместо обязательного управления со стороны правительственных и межправительственных органов. Этот риск подчеркнул немецкий философ Томас Метцингер, член экспертной группы высокого уровня ЕС по ИИ, который заметил, что значительная часть дискурса по этике ИИ формируется частным сектором.

Неравномерное географическое представительство организаций, выпускающих руководства по этике ИИ, требует тщательного мониторинга и анализа со стороны международных, особенно межправительственных, организаций. Для обеспечения инклюзивности, культурного плюрализма и справедливого участия в принятии коллективных решений по ИИ следует поощрять разработку документов «мягкого права» организациями, расположенными в недопредставленных в настоящее время глобальных регионах, особенно в Африке и Южной Америке. Межправительственные организации, такие как Совет Европы, могут сыграть решающую роль в создании международных платформ для взаимного обмена мнениями и дебатов по вопросам этики и управления ИИ.

Сближение существующих инструментов «мягкого права» вокруг пяти общих этических принципов, таких как прозрачность, справедливость, небезразличие, ответственность и конфиденциальность, выявляет пять приоритетных областей для надзора и возможного вмешательства со стороны обязательных органов управления как на правительственном, так и на межправительственном уровне. Приоритетность реализации этих принципов может способствовать созданию основного набора норм, основанных на широко согласованных этических заповедях. Более того, их широкое признание как частными, так и государственными субъектами, вероятно, обеспечит более высокую степень соблюдения. Тем не менее для того, чтобы эти этические принципы воплотились в эффективное управление, их необходимо

концептуально прояснить. Разработчики политики обязаны устранить семантическую двусмысленность и противоречивые характеристики этих принципов. Резкие разногласия в существующих документах «мягкого права» относительно интерпретации и практической реализации этих принципов указывают на то, что обязательные решения в области управления, скорее всего, будут вызывать разногласия в обществе, а значит, требуют прозрачного процесса демократического обсуждения.

Параллельно с этим необходимо более тщательно изучить относительную недопредставленность этических соображений, таких как соображения устойчивости, достоинства и солидарности, чтобы избежать переноса в обязательное управление тех же концептуальных пробелов и нормативных «слепых пятен», что и в «мягком праве». Обязательное управление должно дополнять и заполнять пробелы не обязательных подходов, а не отражать те же «слепые пятна» «мягкого права». Для адекватного решения проблем устойчивости и солидарности ИИ может потребоваться более тесное сотрудничество между агентствами по охране окружающей среды, министерствами труда и занятости, а также министерствами технологий и инноваций.

Поскольку почти половина рассмотренных документов «мягкого права» не содержит четких рекомендаций по поощрению или предупреждению нарушения прав человека при проектировании, разработке и внедрении систем ИИ, необходимо срочно уделить больше внимания последствиям ИИ для прав человека. Страны-члены Совета Европы имеют все возможности для того, чтобы направить международное управление ИИ в сторону поощрения прав человека. Последствия ИИ для прав человека должны быть тщательно изучены на различных уровнях: во-первых, они должны быть изучены на уровне прав и обязанностей в философском смысле, поскольку они действуют независимо от принятия закона как обоснованные моральные нормы. Во-вторых, его следует оценивать на уровне международного права прав человека. В этом отношении Европейская конвенция по правам человека (ЕКПЧ) может сыграть ключевую

роль в международных доктринальных исследованиях и дискуссиях по вопросам ИИ. Соблюдение конвенции является важнейшим требованием для обеспечения социально ответственного развития и внедрения новой технологии. Поэтому крайне важно оценить влияние социотехнических преобразований, вызванных ИИ, на основные права и свободы, постулированные в ЕКПЧ. Такая оценка воздействия должна преследовать двоякую цель:

(I) оценка того, повлияет ли ИИ на права и свободы человека или создаст новые риски для них;

(II) оценка того, может ли ответственное развитие ИИ и общественное обсуждение в связи с ним способствовать поощрению этих прав и свобод.

Следует подчеркнуть, что поскольку технологии разрабатываются не в вакууме, а в социально-историческом контексте человеческой практики, обычаев и норм, эффективные стратегии оценки воздействия должны рассматривать ИИ не абстрактно, а в контексте текущей практики и норм.

Наконец, важно исследовать взаимосвязь между ИИ и правами человека не только с точки зрения высокого уровня, но также и в первую очередь путем изучения правозащитной значимости конкретных областей применения ИИ, таких как, в частности, робототехника, большие данные, автономное оружие и интерфейс мозг-компьютер.

Благодарности

Авторы выражают благодарность д-ру Анне Жобин, Каролине Игнатиадис и Мануэлю Шнайдеру, чья работа способствовала подготовке данного доклада.

Ссылки

Мичи Д., Шпигельхальтер Д. Ж., Тейлор К. Машинное обучение. Нейронная и статистическая классификация. 1994;13.

Аппенцеллер Т. Революция ИИ в науке. Science. 2017;357:16-17.

- Харари Ю.Н. Перезагрузка для революции ИИ. *Nature News*. 2017;550(7676):324.
- Helbing D, Frey BS, Gigerenzer G, et al. Переживет ли демократия большие данные и искусственный интеллект? In: *Towards Digital Enlightenment*. Springer; 2019:73-98.
- Ливингстон С., Риссе М. Будущее влияние искусственного интеллекта на человека и права человека. *Ethics & International Affairs*. 2019;33(2):141-158.
- Немитц П. Конституционная демократия и технологии в эпоху искусственного интеллекта. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2018;376(2133):20180089.
- Ашрафян Х. Интеллектуальные роботы должны отстаивать права человека. *Nature*. 2015;519(7544):391-391.
- Ван Эст Р, Герритсен Дж, Кул Л. Права человека в эпоху роботов: Вызовы, возникающие в связи с использованием робототехники, искусственного интеллекта, виртуальной и дополненной реальности – экспертный доклад, подготовленный для Комитета по культуре. Science, Education and Media of the Parliamentary Assembly of the Council of Europe (PACE) (Rathenau Institute): [https://www/rathenau.nl/sites/default/files/2018-02/Human%20Rights%20in%20the%20Robot%20Age-Rathenau%20Instituut-2017.pdf](https://www.rathenau.nl/sites/default/files/2018-02/Human%20Rights%20in%20the%20Robot%20Age-Rathenau%20Instituut-2017.pdf) (January 5, 2019). 2017.
- Raso FA, Hilligoss H, Krishnamurthy V, Bavitz C, Kim L. Artificial Intelligence & Human Rights: Возможности и риски. Исследовательская публикация Центра Беркмана Кляйна. 2018(2018-6).
- Ассамблея УГ. Всеобщая декларация прав человека. Генеральная Ассамблея ООН. 1948;302(2).
- Моубрей А. Европейская конвенция по правам человека. In: *International Human Rights Law*. Routledge; 2016:287-304.
- Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Международный журнал методологии социальных исследований*. 2005;8(1):19-32.

- Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*. 2019;1(9):389-399.
- Ienca M, Ferretti A, Hurst S, Puhon M, Lovis C, Vayena E. Considerations for ethics review of big data health research: A scoping review. *PloS one*. 2018;13(10):e0204937.
- Li Y, Krishnamurthy R, Raghavan S, Vaithyanathan S, Jagadish H. Regular expression learning for information extraction. Доклад представлен на: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing 2008.
- Shoham Y, Perrault R, Brynjolfsson E, et al. The AI Index 2018 annual report. Руководящий комитет AI Index, Human-Centered AI Initiative, Стэнфордский университет, Стэнфорд, Калифорния. 2018.
- Турилли М, Флориди Л. Этика информационной прозрачности. Этика и информационные технологии. 2009;11(2):105-112.
- Rozin P, Royzman EB. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*. 2001; 5(4):296-320.
- Ito TA, Larsen JT, Smith NK, Cacioppo JT. Негативная информация сильнее давит на мозг: смещение негативности в оценочных категоризациях. *Journal of personality and social psychology*. 1998; 75(4):887.
- Peter J. Bentley, Miles Brundage, Olle Haggstrom, Thomas Metzinger. Должны ли мы бояться искусственного интеллекта? Европейская парламентская исследовательская служба. 2018:Scientific Foresight Unit (STOA).
- Брайсон Дж. Никто не должен доверять искусственному интеллекту. Наука и технологии: Инновации, управление, технологии. 2018; 11:14.
- Winfield AF, Jirotko M. Этическое управление необходимо для создания доверия в робототехнике и системах искусственного интеллекта. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2018;376(2133):20180085.
- Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243. 2019.
- Форум WE. Harnessing Artificial Intelligence for the Earth.

Scheffran J, Brzoska M, Kominek J, Link PM, Schilling J. Climate change and violent conflict. *Наука*. 2012;336(6083):869-871.

IEEE. Этически выверенный дизайн. IEEE Standards v1. 2016 (Global Initiative).

Расмуссен Т. Социальная теория и коммуникационные технологии. Routledge; 2019.

Лю Х-Й, Завиеска К. От ответственной робототехники к режиму прав человека, ориентированному на вызовы робототехники и искусственного интеллекта. *Этика и информационные технологии*. 2017:1-13.

Mantelero A. AI and Big Data: План оценки воздействия на права человека, социальные и этические аспекты. *Computer Law & Security Review*. 2018;34(4):754-772.

Вайена Е, Тасиулас Дж. Динамика больших данных и прав человека: Случай научных исследований. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2016;374(2083):20160129.

Хейнс К. Права человека и использование автономных систем вооружения (AWS) при охране правопорядка в стране. *Hum Rts Q*. 2016;38:350.

Асаро П. О запрете автономных систем оружия: права человека, автоматизация и дегуманизация принятия смертоносных решений. *International Review of the Red Cross*. 2012;94(886):687-709.

Иенка М, Андорно Р. К новым правам человека в эпоху нейронауки и нейротехнологий. *Life Sci Soc Policy*. 2017;13(1):5.

Анализ международных юридически обязательных документов. Заключительный отчет⁴⁴

Исполнительное резюме

Последняя волна развития искусственного интеллекта (ИИ) оказывает все большее преобразующее воздействие на общество и поднимает новые вопросы в различных областях, от предиктивной медицины до модерации медиаконтента, от квантифицированного «я» до судебных систем, не упуская из виду вопросы воздействия на окружающую среду.

Таким образом, анализ международных юридически обязательных документов является обязательной отправной точкой для определения существующей правовой базы, выявления ее руководящих ценностей и проверки того, правильно ли эта база и ее принципы решают все вопросы, поднятые ИИ.

С целью сохранения гармонизации существующей правовой базы в области прав человека, демократии и верховенства закона, данное исследование призвано внести вклад в разработку будущего регулирования ИИ, опираясь на существующие обязательные инструменты, контекстуализируя их принципы и предоставляя ключевые руководящие принципы регулирования для будущей правовой базы.

Теоретическая основа этого подхода базируется на предположении, что общие принципы, предусмотренные международными документами по правам человека, должны лежать в основе всех видов человеческой деятельности, включая инновации на базе ИИ. Более того, только система прав человека может служить универсальной основой для регулирования ИИ, в то время как другие сферы (например, этика) не имеют такого глобального измерения, в большей степени зависят от контекста и характеризуются разнообразием теоретических подходов.

44 Отчет, подготовленный для САНАИ, Страсбург, 15 июня 2020 года, под названием САНАИ(2020)08-fin.

Анализ существующих обязательных правовых инструментов, содержащийся в данном документе, не ограничивается гармонизирующим исследованием, извлечением общих ценностей и принципов из определенного набора правил по ИИ. Более артикулированное исследование проводится на разных этапах.

После первоначального отраслевого анализа, направленного на определение ключевых руководящих принципов в четырех основных областях (защита данных, здравоохранение, демократия и правосудие), эти принципы рассматриваются в контексте изменений в обществе, вызванных ИИ. При этом мы извлекаем пользу из существующих необязательных документов, которые обеспечивают более детальное применение принципов, закрепленных в международных правовых документах, а в некоторых случаях также содержат конкретные рекомендации по ИИ.

Такая контекстуализация руководящих принципов и правовых ценностей обеспечивает их более тонкую и детальную формулировку с учетом специфики продуктов и услуг ИИ и помогает лучше решать проблемы, возникающие в связи с ИИ. Это позволяет сформулировать первоначальный набор положений для будущего регулирования ИИ с упором на наиболее сложные вопросы в каждом из рассмотренных секторов.

Учитывая большое количество документов, принятых несколькими международными и межправительственными органами, а также параллельно проводимое САНАИ исследование этических инструментов, данный документ посвящен юридически обязательным документам, а также необязательным документам, принятым для их реализации.

Исследование состоит из двух частей. В первой определяются сфера охвата и методология данного анализа, а во второй представлены результаты секторального анализа руководящих принципов.

В рамках секторального анализа первыми рассматриваются две ключевые области – здравоохранение и защита данных. Пересечение этих двух сфер интересно с точки зрения направленности данного исследования, учитывая

большое количество приложений ИИ, касающихся данных здравоохранения, и общность этих двух областей. Это отражено в нескольких положениях.

Конвенции Овьедо и Конвенции 108+, а также в не имеющих обязательной силы документах. Более того, индивидуальное самоопределение играет центральную роль как в области защиты данных, так и в области биомедицины, поэтому проблемы ИИ – с точки зрения сложности и непрозрачности процедур и операций по обработке – особенно актуальны и имеют общие проблемы.

Четвертый и пятый разделы посвящены демократии и правосудию. Здесь поле исследования шире, и нет комплексных правовых инструментов, которые могли бы обеспечить конкретные секторальные принципы, такие как Конвенция 108+ или Конвенция Овьедо. Поэтому анализ в большей степени сосредоточен на принципах высокого уровня и их контекстуализации с более ограниченной проработкой ключевых руководящих положений по сравнению с предыдущими разделами.

В последнем разделе представлен обзор выявленных руководящих принципов и предложена структура гармонизации, указывающая на существующие взаимосвязи и точки соприкосновения между этими принципами и, в то же время, подчеркивающая уникальный вклад каждого сектора в будущее регулирование ИИ.

Основная цель данного исследования заключается не в том, чтобы добавить новый список руководящих принципов к уже представленным различными органами и структурами, а в том, чтобы достичь иного результата в методологическом и содержательном плане.

Во-первых, проведенный анализ и предложенное решение имеют свои корни и опираются на права и свободы человека, применяя конкретный подход, основанный на существующих международных правовых инструментах. Другие исследования часто являются отраслевыми и имеют другой набор нормативных ссылок (национальных или региональных) или используют теоретический подход, провозглашая принципы или ссылаясь на права

человека в общем и абстрактном виде. Хотя эти работы способствуют расширению правовых и этических дебатов по ИИ, их влияние с точки зрения вклада в нормативно-правовую базу ограничено и не имеет конкретного контекста в рамках стандартов Совета Европы по правам человека, демократии и верховенству закона.

Во-вторых, результатом анализа юридически обязательных документов, включая необязательные документы, принятые для их реализации, является не просто список принципов, каким бы точным он ни был. Выявление общих руководящих принципов важно, но недостаточно для создания дорожной карты для будущего регулирования ИИ. Прозрачность, подотчетность, человеческий надзор и многие другие принципы, уже перечисленные в нескольких хартиях по ИИ, являются абстрактными понятиями без надлежащей контекстуализации.

Основной вклад данного исследования заключается в том, чтобы предоставить именно такую контекстуализацию в отношении правовой базы и проблем ИИ. Если в данном документе удастся предложить конкретные и эффективные способы формулирования и кодификации этих руководящих принципов в отношении ИИ и конкретного включения стандартов Совета Европы в области прав человека, демократии и верховенства закона в контуры будущего регулирования ИИ, то он достигнет своей цели – поможет сформулировать отношения между людьми и ИИ с правовой точки зрения.

II. Сфера применения и методология

Как и Интернет, электричество и паровая энергия, искусственный интеллект (ИИ) включает в себя целый ряд различных технологий, оказывающих широкое влияние на различные сферы деятельности человека и общества.

В этом контексте множество различных правовых инструментов могут иметь важное значение для регулирования применения ИИ. В то же время, эти правовые инструменты были приняты в эпоху, предшествующую ИИ, что

может снизить их эффективность в обеспечении адекватного и конкретного ответа на новые вызовы ИИ.

Таким образом, анализ международных юридически обязательных документов является обязательной отправной точкой для определения существующей правовой базы, выявления ее руководящих ценностей и проверки того, насколько эта база и ее принципы должным образом учитывают все вопросы, поднимаемые ИИ, с целью сохранения гармонизации существующей правовой базы в области прав человека, демократии и верховенства закона.

Данный подход не ставит своей целью создание совершенно новой и всеобъемлющей справочной системы, поскольку регулирование должно быть сосредоточено на том, какие изменения ИИ принесет обществу, а не на перестройке всех областей, в которых может применяться ИИ. Такой целевой подход становится возможным благодаря использованию существующих обязательных документов, контекстуализации их руководящих принципов и разработке ключевых руководящих принципов регулирования для будущей правовой базы для ИИ, которая может охватывать области, которые в настоящее время не регулируются существующими обязательными документами.

В этой связи важно подчеркнуть разницу между существующими юридически обязывающими документами и другими документами, такими как документы «мягкого права» или этические хартии по ИИ. Юридически обязывающие документы существовали до появления нынешней весны ИИ. Они не были разработаны с учетом ИИ и не содержат конкретного набора правил для этой области, в то время как документы по мягкому праву и этике ИИ дают конкретную направленность, хотя и с разных точек зрения.

Поэтому анализ существующих обязательных правовых инструментов не ограничивается исследованием гармонизации (т.е. извлечением общих ценностей и принципов из определенного набора правил по ИИ), а требует более четко сформулированного процесса, в котором гармонизация является

лишь одним из нескольких этапов. Этот процесс можно разделить на три отдельных этапа: (I) картирование и определение ключевых принципов, (II) контекстуализация и (III) гармонизация.

Сценарий

Последняя волна развития ИИ оказывает все большее преобразующее воздействие на общество и ставит новые вопросы в различных областях, от предиктивной медицины до модерации медиаконтента, от квантифицированного «я» до судебных систем, не упуская из виду вопросы воздействия на окружающую среду.

Быстрая эволюция прикладного ИИ за последние несколько лет была несовместима с конкретным правовым ответом в виде международных юридически обязывающих документов, посвященных ИИ. Именно поэтому мы наблюдаем развитие двух различных операционных стратегий для решения этих вопросов: (I) значительные усилия по интерпретации существующей правовой базы в свете вопросов, связанных с ИИ (см., например, продолжающиеся дебаты по положениям GDPR о прозрачности и автоматизированном принятии решений); (II) использование необязательных правил для контекстуализации принципов, предусмотренных существующими обязательными документами (например, T-PD(2019)01 Guidelines on Artificial Intelligence and Data Protection; CEPEJ. 2019. Европейская этическая хартия по использованию искусственного интеллекта (ИИ) в судебных системах и их окружении⁴⁵).

Поэтому будущее регулирование ИИ должно опираться на эти усилия, уделяя внимание как руководящим принципам и ценностям, вытекающим из существующих обязательных инструментов, так и их соответствующим необязательным реализациям, которые в некоторых случаях уже учитывают новый сценарий ИИ.

⁴⁵ Европейская комиссия по эффективности правосудия (CEPEJ). 2018. Европейская этическая хартия по использованию искусственного интеллекта в судебных системах и их окружении.

Направленность и методология исследования

Основной целью данного исследования является определение ключевых принципов будущего регулирования ИИ путем анализа существующей правовой базы. Таким образом, методология исследования является дедуктивной и предполагает извлечение этих принципов из разнообразных нормативных актов, касающихся областей, в которых потенциально могут быть приняты решения в области ИИ.

Теоретическая основа данного подхода базируется на предположении, что общие принципы, предусмотренные международными документами по правам человека, должны лежать в основе всех видов человеческой деятельности, включая инновации на базе ИИ.⁴⁶ Более того, только система прав человека может служить универсальной основой для регулирования ИИ, в то время как другие сферы (например, этика) не имеют такого глобального измерения, в большей степени зависят от контекста и характеризуются разнообразием теоретических подходов.

На этом фоне возникает множество вопросов, таких как: когда система ИИ должна принимать решение? Какие критерии должна применять система? Кто несет ответственность за решения, которые могут негативно повлиять на человека и общество? В связи с этими и многими другими возникающими вопросами необходимо пересмотреть существующие нормативные акты.

Чтобы создать гармонизированную нормативно-правовую базу для решения проблем ИИ, общие и высокоуровневые руководящие принципы и ценности, которые должны быть закреплены, должны быть взяты из международных хартий прав человека (например, Всеобщей декларации прав человека, Международного пакта о гражданских и политических правах, Конвенции о защите прав человека и основных свобод, Хартии основных прав Европейского союза).

⁴⁶ См. также Комитет министров. 2020. Рекомендация CM/Rec(2020) о воздействии алгоритмических систем на права человека.

Эти руководящие принципы должны рассматриваться в рамках трансформационного сценария, управляемого ИИ, который во многих случаях потребует их адаптации. Эти принципы остаются в силе, но их действие должно быть пересмотрено в свете социальных и технических изменений, вызванных ИИ (например, свобода выбора в случае так называемых «черных ящиков»). Это позволит более контекстуализированно и детально применять принципы, чтобы они могли внести конкретный вклад в формирование будущего регулирования ИИ.

Для проведения данного исследования нам необходимо начать с определения основных областей исследования, учитывая как потенциальное воздействие ИИ, так и сферы деятельности Совета Европы. В связи с этим были выбраны четыре ключевые области: данные, здравоохранение, демократия и правосудие.

Анализ и ожидаемые результаты

В исследовании используется подход «сверху вниз» с целью внесения вклада в будущую нормативную базу ИИ, которая будет реализована с помощью дополнительных обязательных и необязательных инструментов, как это произошло в области биомедицины. Ожидаемым результатом является набор положений, касающихся исследуемых областей, и ключевых общих руководящих принципов, основанных на всестороннем анализе всего корпуса обязательных инструментов, включая принятые необязательные инструменты.

Первый этап: Составление карты и определение ключевых принципов. В различных исследуемых областях будут определены руководящие принципы. Первый этап анализа основан на различных субъектах, так как обязательные к исполнению инструменты являются отраслевыми, а не основанными на правах. Следующие две таблицы представляют собой первый пример такого сопоставления, основанного на предварительном обзоре сфер защиты данных и правосудия, для определения руководящих принципов для будущего регулирования ИИ.

Рисунок 1: Защита данных

Обязательные инструменты	Конвенция 108 Конвенция о киберпреступности
Затронутые области	Системы принятия решений Групповая конфиденциальность и коллективное измерение Профилирование
Связанные необязательные инструменты	СЕ. 2019. Руководство по последствиям искусственного интеллекта для защиты данных СЕ. 2017. Руководящие принципы по защите физических лиц в отношении обработки персональных данных в мире Больших Данных СЕ. 2010. Рекомендация о защите физических лиц в отношении автоматической обработки персональных данных в контексте профилирования [пересматривается]. ЮНЕСКО. 2019. Предварительное исследование возможного нормативного документа по этике искусственного интеллекта ОЭСР. 2019. Рекомендация Совета по искусственному интеллекту 40-я Международная конференция комиссаров по защите данных и конфиденциальности. 2018

Руководящие принципы и правовые ценности	Подотчетность Риск-ориентированный подход Принцип предосторожности Качество и безопасность данных Прозрачность Справедливость Контекстуальный подход Роль экспертов Участие/инклюзивность Свобода выбора/автономия Человек контроль/надзор Осведомленность Грамотность Ответственные инновации Сотрудничество между надзорными органами
Обязательные инструменты	Всеобщая декларация прав человека Международный пакт о гражданских и политических правах Международная конвенция о ликвидации всех форм расовой дискриминации Конвенция о ликвидации всех форм дискриминации в отношении женщин Конвенция о защите прав человека и основных свобод Хартия основных прав Европейского Союза
Затронутые районы	Обработка судебных решений и данных Предиктивная полицейская деятельность
Соответствующие необязательные документы	СЕРЕJ. 2019. Европейская этическая хартия по использованию искусственного интеллекта (ИИ) в судебных системах и их окружении

Основные принципы и правовые ценности	Недискриминация Качество и безопасность данных Прозрачность Беспристрастность Справедливость Контекстуальный подход Свобода выбора/ Независимость судей (процесс принятия решений) Человеческий контроль/надзор Гарантии права на справедливое судебное разбирательство

Второй этап: Контекстуализация.

Руководящие ценности, выявленные в процессе картирования, должны быть контекстуализированы в свете изменений в обществе, вызванных ИИ. На этом этапе будут полезны существующие необязательные документы, которые обеспечивают более детальное применение принципов, закрепленных в обязательных документах, а в некоторых случаях также содержат конкретные указания по ИИ.

Такая контекстуализация руководящих принципов и правовых ценностей обеспечит их более тонкую и детальную формулировку с учетом специфики продуктов и услуг ИИ. Таким образом, на данном этапе можно будет сформулировать первоначальный набор положений для будущего регулирования ИИ, сосредоточив внимание на наиболее сложных вопросах в каждом секторе.

Третий этап: гармонизация (межсекторальный анализ). На основе отраслевого анализа, проведенного в данном исследовании, в последнем разделе будет составлен список ключевых руководящих принципов, общих для различных сфер. Эти общие принципы станут краеугольным камнем общего ядра будущих положений об ИИ.

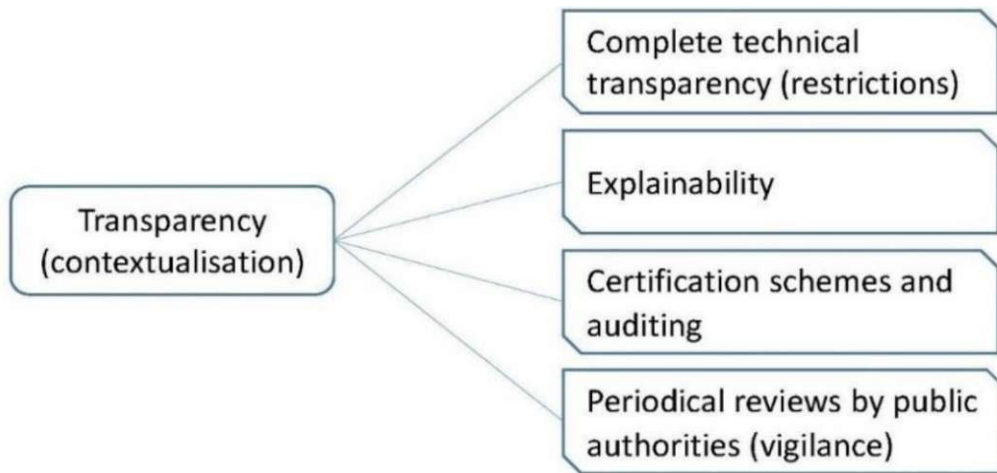


Рисунок: Реализация принципа прозрачности в зависимости от контекста

III. Анализ

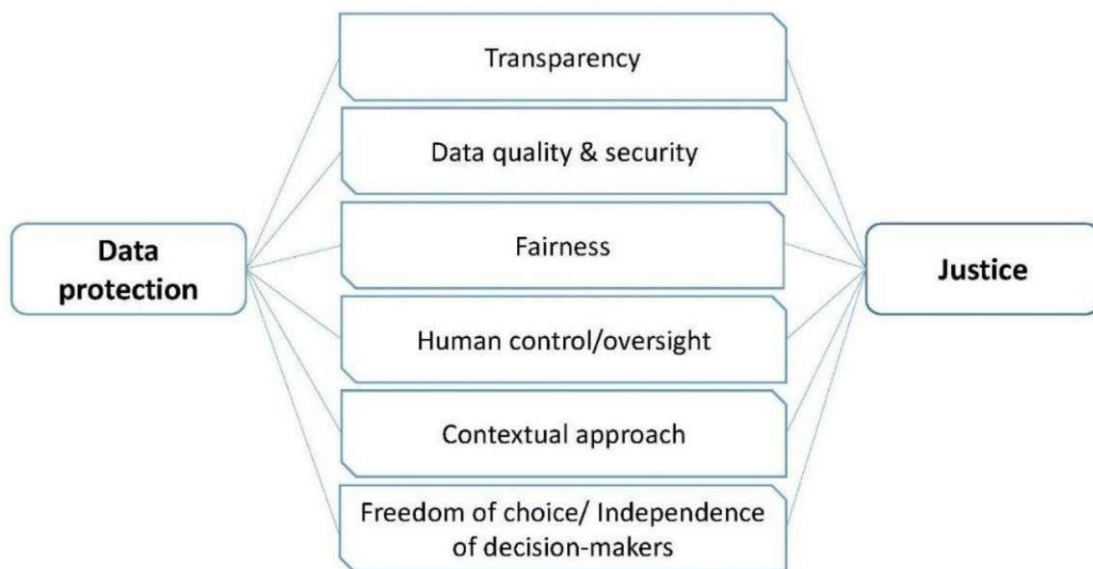


Рисунок: Общие руководящие ценности в области защиты данных и правосудия

Учитывая большое количество документов, принятых несколькими международными и межправительственными органами, а также параллельно проводимое САНАИ исследование этических инструментов, данная часть посвящена юридически обязательным документам, включая необязательные документы, принятые для их реализации. Поэтому этические руководства на данном этапе не рассматриваются, а документы, касающиеся будущих

стратегий регулирования (например, белые книги), принимаются во внимание только в качестве справочной информации.

Данная часть состоит из шести разделов, за которыми следуют некоторые заключительные соображения. В первом разделе представлен общий обзор существующих инструментов, принятых Советом Европы, и основных основополагающих принципов/ценностей. Это помогает определить потенциальные основные принципы будущего регулирования ИИ и его согласованность с существующей системой.

Второй и третий разделы посвящены двум ключевым и смежным областям: здравоохранению и защите данных. Пересечение этих двух сфер интересно с точки зрения направленности данного исследования, учитывая большое количество приложений ИИ, касающихся данных здравоохранения, и общность этих двух областей. Это отражено в ряде положений Конвенции Овьедо и Конвенции 108+, а также в необязательных документах.⁴⁷ Более того, индивидуальное самоопределение играет центральную роль как в области защиты данных, так и в биомедицине, поэтому проблемы ИИ – с точки зрения сложности и непрозрачности процедур и операций по обработке данных – особенно актуальны и имеют общие проблемы.

Четвертый и пятый разделы посвящены демократии и правосудию. Здесь область исследования шире, и нет общих правовых инструментов, которые могли бы обеспечить специфические для сектора принципы, такие как Конвенция 108+ или Конвенция Овьедо. Поэтому анализ сосредоточен на принципах высокого уровня и их контекстуализации, в результате чего ключевые руководящие положения проработаны более ограниченно, чем в предыдущих разделах.

В разделе 6 представлен общий обзор выявленных руководящих принципов и предложена схема гармонизации, которая подчеркивает как существующие взаимосвязи между этими принципами, так и уникальный вклад каждого сектора в будущее регулирование ИИ.

Как следует из комментариев, полученных в ходе мониторинга, описанного в следующем разделе, технологии ИИ оказывают влияние на различные сектора и поднимают вопросы, касающиеся большого количества нормативных документов. Поэтому данное первоначальное исследование является отправной точкой, сфокусированной на четырех основных упомянутых областях. Однако, несмотря на ограниченный охват, результаты подтверждают предложенную методологию и дают ряд указаний на будущие положения в регулировании ИИ.

I. Общий обзор

*Как ИИ влияет на различные ситуации*⁴⁸

Поскольку ИИ влияет на целый ряд ситуаций, рассматриваемых различными обязательными документами, охватывающими несколько областей, нам необходимо провести анализ, основанный на фактах, чтобы определить ключевые принципы и общие ценности, которые следует рассмотреть для будущего регулирования.

В связи с этим в период с 12 по 28 февраля 2020 года был проведен первоначальный мониторинг, в котором приняли участие различные отделения Совета Европы, чтобы воспользоваться отраслевым опытом различных подразделений, которые на протяжении многих лет работали в различных областях, связанных с правами человека, демократией и верховенством закона.

С помощью опроса, основанного на открытых вопросах, различным опрошенным подразделениям было предложено предоставить информацию по следующим направлениям: (I) обязательные документы, (II) области воздействия (применения), (III) соответствующие необязательные документы, (IV) руководящие принципы и правовые ценности, и (V) отсутствующие принципы/вопросы. Благодаря позитивной приверженности различных областей, удалось собрать множество различных видов информации.

С методологической точки зрения, структура этого предварительного исследования, основанного на открытых вопросах, неизбежно влияет на результаты количественного анализа. Основные ограничения связаны с использованием различных и частично пересекающихся общих категорий, а также с различными уровнями детализации и конкретизации ответов.

Тем не менее, благодаря агрегированию по макро-областям и фокусированию на сходстве (т.е. частоте) в выявленных принципах и ценностях, мы смогли добиться определенной перспективы в результатах, и в ходе исследования была составлена более подробная карта имеющихся необязательных инструментов, принятых Советом Европы, которые могут помочь в создании правовой базы для будущего регулирования (см. Приложение 1).

Что касается затронутых областей (см. Приложение 2), упражнение предлагает сфокусировать будущее регулирование ИИ по двум основным направлениям: использование ИИ и развитие ИИ. В обоих случаях различные права человека и основные свободы потенциально затрагиваются или могут играть важную роль в формировании будущих сценариев ИИ.⁴⁹

Что касается использования ИИ, существуют четыре основные области применения и соответствующего регулирования: прогнозный анализ и системы поддержки принятия решений, автоматизированные системы принятия решений, сбор доказательств/компьютерная криминалистика и создание контента.

Первые две области хорошо известны и обсуждаемы, поскольку они охватывают чрезвычайно широкий спектр применения (см. Приложение 2). Тем не менее, различие между системами поддержки принятия решений и автономными системами принятия решений имеет решающее значение с точки зрения ценностно-ориентированного проектирования и роли человека в процессе принятия решений: различная природа этих двух типов систем

обязательно потребует различных процедурных и материальных гарантий в регулировании ИИ.

Последние две области являются отраслевыми, но их следует рассматривать отдельно, поскольку они не касаются непосредственно процесса принятия решений, но обеспечивают доказательства, лежащие в его основе (сбор доказательств и компьютерная криминалистика), или влияют на процессы создания (создание контента). В этих случаях основные вопросы представляются иными и больше сосредоточены на процедурных аспектах и их согласованности с традиционными (т.е. не основанными на ИИ) подходами.

Хотя большинство существующей литературы и руководств посвящено системам ИИ и их потенциальным последствиям, важное влияние ИИ на права человека и основные свободы также связано с развитием ИИ и предоставлением услуг ИИ. В связи с этим будущее регулирование ИИ должно тщательно рассмотреть вопросы, связанные с условиями труда людей, вовлеченных во всю цепочку поставок продуктов и услуг ИИ.⁵⁰

Второй блок информации, полученной в ходе мониторинга, касается руководящих принципов и правовых ценностей, которые должны лежать в основе будущего развития и использования ИИ (см. Приложение 3). Здесь разнообразие понятий, используемых опрошенными подразделениями, позволяет говорить об агрегировании принципов и ценностей.

Результат этого анализа позволил сгруппировать руководящие принципы и ценности вокруг ряда ключевых элементов, которые выделились по распределению (частоте):

Борьба с дискриминацией (15)

Многообразие, инклюзия и плюрализм (13) Конфиденциальность и защита данных (11) Прозрачность (9) Равенство (8)

Доступ к правосудию, справедливый суд (7) Контроль над людьми (7)

Беспристрастность (6)

Доступ к информации (5) Безопасность (5)

⁵⁰ См. также Crawford and Joler, 2018.

Справедливость (5)

Участие (5) Свобода выбора (5)

Свобода выражения и творчества (5)

Подотчетность (3) Компетентность и способность (2) Независимость (3)

Индивидуальная автономия (3)

Культурное сотрудничество (2)

Устойчивость и долгосрочная ориентация (2)

Несмотря на упомянутые ограничения анализа, очевидно, что первые три принципа рассматриваются как ключевые элементы будущего регулирования ИИ и, следовательно, будут находиться в центре его внимания. Это подтверждает и второй набор принципов/ценностей, который тесно связан с первым: прозрачность и человеческий контроль являются важными факторами недискриминации и защиты данных, а доступ к правосудию – общим условием для решения любых потенциальных нарушений этих ценностей. Аналогичным образом, хотя и более предметно, равенство различным образом связано с первыми тремя основными ценностями/принципами. Другие ценности/принципы, затрагивающие различные конкретные проблемы реализации ИИ, различаются более широко.

Эта работа позволила определить первый список руководящих принципов регулирования ИИ, которые уже кодифицированы в обязательных и необязательных правовых документах, но нуждаются в контекстуализации в области ИИ. В отраслевом анализе эта контекстуализация, основанная на углубленном анализе международных юридически обязательных инструментов, будет достигнута путем оценки любых потенциальных пробелов в существующей нормативно-правовой базе по секторам.

Поскольку ИИ является межотраслевой технологией, ожидается, что результаты этого анализа могут предложить аналогичные регулятивные вмешательства в других областях, как указано в части, посвященной методологии. После завершения отраслевого анализа все эти потенциальные

меры вмешательства будут систематизированы, чтобы избежать дублирования и объединить их в целостную систему, основанную на ключевых ценностях.

II. Защита данных

За последнее десятилетие международная нормативно-правовая база в области защиты данных претерпела значительное обновление. Правовые инструменты, сформированные на основе принципов, определенных в 1970-х и 1980-х годах, уже не отвечали изменившемуся социально-техническому ландшафту, созданному растущей доступностью пропускной способности для передачи данных, хранения данных и вычислительных ресурсов (облачные вычисления), прогрессирующей информатизацией значительной части нашей жизни и окружающей среды (IoT), а также масштабным и прогностическим анализом данных на основе Big Data и Machine Learning.

В Европе основными ответами на эти изменения стали модернизированная версия Конвенции 108 (Конвенция 108+) и GDPR. Аналогичный пересмотр нормативно-правовой базы был осуществлен или осуществляется в других международных контекстах – например, в ОЭСР.⁵¹ – или в отдельных странах.

Однако, учитывая стремительное развитие последней волны развития ИИ, эти новые обязательные документы не могут напрямую решить некоторые специфические для ИИ проблемы и для преодоления этого пробела было принято несколько необязательных документов, а также обсуждаемые будущие стратегии регулирования.⁵²

Поэтому для целей данного исследования были проанализированы следующие необязательные правовые документы:⁵³ T-PD(2019)01, Руководство по искусственному интеллекту и защите данных [GAI]; T-PD(2017)1, Руководство по защите физических лиц в отношении обработки персональных

51 См. ОЭСР. 2013. Рекомендация Совета относительно руководящих принципов, регулирующих защиту частной жизни и трансграничные потоки персональных данных, C(80)58/FINAL, с поправками от 11 июля 2013 года C(2013)79.

52 См. Европейская комиссия. 2020. Доклад о последствиях искусственного интеллекта, Интернета вещей и робототехники для безопасности и ответственности, COM(2020) 64 final; Европейская комиссия. 2020. Белая книга по искусственному интеллекту – европейский подход к совершенству и доверию, COM(2020) 65 final; Европейская комиссия. 2020. Европейская стратегия в отношении данных, COM(2020) 66 final.

53 См. Приложение 4.

данных в мире Больших Данных; Рекомендация CM/Rec(2019)2 Комитета министров Совета Европы для государств-членов по защите данных, связанных со здоровьем;⁵⁴ Рекомендация CM/Rec(2010)13 Комитета министров Совета Европы государствам-членам о защите физических лиц в отношении автоматической обработки персональных данных в контексте профилирования; ЮНЕСКО. 2019. Предварительное исследование возможного нормативного документа по этике искусственного интеллекта [ЮНЕСКО]; OECD. 2019. Рекомендация Совета по искусственному интеллекту [ОЭСР]; 40-я Международная конференция комиссаров по защите данных и конфиденциальности. 2018 [ICDPPC]. Декларация об этике и защите данных в искусственном интеллекте.

Эти документы различаются по своему характеру: в то время как документы, принятые Советом Европы, определяют различные конкретные требования и положения, другие документы в основном основаны на принципах, устанавливая несколько принципов, но не предоставляя или предоставляя лишь частично более подробное руководство с точки зрения конкретных требований. Следующие параграфы иллюстрируют ключевые принципы, вытекающие из этих различных инструментов, и то, как они могут быть контекстуализированы в рамках сценария ИИ.

Некоторые из этих принципов, относящиеся к области защиты персональных данных (например, качество данных), могут быть распространены на неперсональные данные, в основном в отношении влияния использования неперсональных данных (например, агрегированных данных) на отдельных лиц и группы лиц в контексте процессов принятия решений (например, данные о мобильности или данные о потреблении энергии).

⁵⁴ Данная Рекомендация заменила Рекомендацию № R(97)5 о защите медицинских данных. См. также Rec(2016)8 об обработке персональных данных о здоровье в целях страхования, включая данные, полученные в результате генетических тестов, и пояснительный меморандум к ней.

Главенство человека

Системы ИИ должны служить человечеству, и любое создание, развитие и использование систем ИИ должно полностью уважать права человека, демократию и верховенство закона.⁵⁵

Человеческий контроль

Приложения ИИ должны обеспечивать значимый контроль со стороны человека за их воздействием на человека и общество.⁵⁶

Прозрачность и расширяемость

Каждый человек имеет право на соответствующее информирование, когда он непосредственно взаимодействует с системой ИИ, предоставляя адекватную и простую для понимания информацию о цели и последствиях работы этой системы, включая наличие автоматизированных решений, с тем чтобы проверить постоянное соответствие ожиданиям людей, обеспечить общий человеческий контроль над такими системами и дать возможность тем, кто пострадал от системы ИИ, оспорить ее результаты.⁵⁷

Каждое лицо также имеет право по запросу получить информацию о причинах, лежащих в основе процесса принятия решений на основе ИИ, когда результаты такого процесса применяются к нему или к ней.⁵⁸ Кроме того, государства содействуют научным исследованиям в области объяснимого искусственного интеллекта и передовой практики прозрачности и проверяемости систем ИИ.⁵⁹

55 См. CM/Rec(2019)2; ICDPPC; GAI, пп. I.1 и II.1; ЮНЕСКО. См. также GDPR, Recital no. 4.

56 См. GAI, п. I.6.

57 См. ICDPPC, CM/Rec(2019)2, ОЭСР, ЮНЕСКО. См. также Рекомендацию CM/Rec(2020)1 о воздействии алгоритмических систем на права человека.

58 См. Конвенцию 108+; GAI II.11.

59 См. ICDPPC.

Предупредительный подход

Когда потенциальные риски применения ИИ неизвестны или неопределенны, развитие ИИ должно основываться на принципе предосторожности.⁶⁰

Управление рисками

Разработчики, производители и поставщики услуг ИИ должны оценивать и документировать возможные негативные последствия применения ИИ для прав человека и основных свобод, а также принимать соответствующие меры по предотвращению и снижению рисков, начиная с этапа проектирования (подход «права человека в процессе проектирования») и на протяжении всего жизненного цикла.⁶¹ Неблагоприятные последствия включают в себя последствия, вызванные использованием деконтекстуализированных данных и деконтекстуализированных алгоритмических моделей.⁶²

Разработчики, производители и поставщики услуг ИИ должны консультироваться с компетентными надзорными органами, если приложения ИИ потенциально могут существенно повлиять на права человека и основные свободы отдельных лиц.⁶³

Риск повторной идентификации

Необходимо ввести соответствующие меры для защиты от любой возможности того, что анонимные и агрегированные данные могут привести к повторной идентификации субъектов данных.⁶⁴

60 См. GAI, п. II.2.

61 22 См. GAI, пункты II.2 и II.3; ICDPPC; ОЭСР; ЮНЕСКО. См. также Рекомендацию CM/Rec(2020)1 о воздействии алгоритмических систем на права человека.

62 См. GAI, п. II.5.

63 См. GAI, п. III.5.

64 См. CM/Rec(2010)13.

Качество и минимизация данных

Разработчики ИИ должны критически оценивать качество, характер, происхождение и объем используемых персональных данных, сокращая ненужные, избыточные или маргинальные данные на этапах разработки и обучения ИИ, а также отслеживая точность модели по мере поступления в нее новых данных. Использование синтетических данных может рассматриваться как одно из возможных решений для минимизации объема персональных данных, обрабатываемых приложениями ИИ.⁶⁵

Роль экспертов

Разработчикам, производителям и поставщикам услуг ИИ рекомендуется создавать независимые комитеты экспертов из различных областей, а также взаимодействовать с независимыми академическими институтами, которые могут способствовать разработке приложений ИИ, основанных на правах человека, этически и социально ориентированных, а также выявлению потенциальной предвзятости. Такие комитеты могут играть особенно важную роль в областях, где прозрачность и взаимодействие с заинтересованными сторонами могут быть более сложными из-за конкурирующих интересов и прав, например, в области предиктивного правосудия, предотвращения и выявления преступлений.⁶⁶

Необходимо создать соответствующие механизмы для обеспечения независимости этих комитетов экспертов.⁶⁷

Участие и демократический надзор за разработкой ИИ

Должны быть разработаны партисипативные формы оценки рисков, основанные на активном вовлечении отдельных лиц и групп, на которые потенциально может повлиять применение ИИ. Отдельные лица, группы и

⁶⁵ См. GAI п. II.4; OECD. и CM/Rec(2020)1.

⁶⁶ См. также ниже раздел II.3.

⁶⁷ См. GAI, пункты II.6 и II.7; ICDPPC. См. также Статью11, ЮНЕСКО. Декларация о геноме человека и правах человека (11 ноября 1997 г.).

другие заинтересованные стороны должны быть информированы и активно вовлечены в обсуждение того, какую роль ИИ должен играть в формировании социальной динамики и в процессах принятия решений, затрагивающих их интересы.⁶⁸ Отступления могут быть введены в интересах общества, если это соразмерно в демократическом обществе и с адекватными гарантиями.

Человеческий надзор

Продукты и услуги ИИ должны быть разработаны таким образом, чтобы обеспечить право человека не быть объектом решения, существенно влияющего на него, основанного исключительно на автоматизированной обработке данных, без учета его мнения. Продукты и услуги ИИ должны обеспечивать возможность общего контроля над ними со стороны человека.⁶⁹

Кроме того, следует сохранить роль человеческого вмешательства в процессы принятия решений на основе ИИ и свободу людей, принимающих решения, не полагаться на результат рекомендаций, предоставленных с помощью ИИ.⁷⁰

Бдительность алгоритмов

Разработчики, производители и поставщики услуг ИИ должны принять формы бдительности в отношении алгоритмов, которые способствуют подотчетности всех соответствующих заинтересованных сторон путем оценки и документирования ожидаемого воздействия на человека и общество на каждом этапе жизненного цикла системы ИИ на постоянной основе, чтобы обеспечить соблюдение прав человека, верховенства закона и демократии.⁷¹ Правительства должны регулярно предоставлять отчеты об использовании ИИ в полицейской деятельности, разведке и безопасности.⁷²

68 См. GAI, пп. II.7 и III.8; ICDPPC. См. также CM/Rec(2020)1.

69 См. Конвенцию 108+; GAI п. II.8; ICDPPC; ЮНЕСКО.

70 См. GAI п. III. 4.

71 См. GAI п. II.10; OECD; ICDPPC. See also CM/Rec(2020)1.

72 См. UNESCO.

Свобода выбора

В целях повышения доверия пользователей разработчикам, производителям и поставщикам услуг ИИ рекомендуется разрабатывать свои продукты и услуги таким образом, чтобы гарантировать свободу выбора пользователей в отношении использования ИИ, предоставляя реальные альтернативы приложениям ИИ.⁷³

Право на возражение

Право на возражение должно быть обеспечено в отношении систем искусственного интеллекта, основанных на технологиях, которые влияют на мнения и развитие личности человека.⁷⁴

Операционная совместимость

Оперативная совместимость между системами ИИ должна осуществляться в полном соответствии с принципами законности, необходимости и пропорциональности, обеспечивая соответствующие гарантии прав человека, демократии и верховенства закона.⁷⁵

Сотрудничество

Сотрудничество должно поощряться между надзорными органами, обладающими компетенцией, связанной с искусственным интеллектом.⁷⁶

Цифровая грамотность, образование и профессиональная подготовка

Разработчики политики должны инвестировать ресурсы в цифровую грамотность и образование, чтобы повысить осведомленность и понимание субъектами данных приложений ИИ и их последствий. Они также должны поощрять профессиональную подготовку разработчиков ИИ, чтобы повысить осведомленность и понимание потенциального влияния ИИ на человека и

⁷³ См. GAI, п. II.9.

⁷⁴ См. GAI, п. II.12. и II.4.

⁷⁵ См. CM/Rec(2019)2.

⁷⁶ См. ICDPPC; GAI, и. III.6.

общество. Они должны поддерживать исследования в области ИИ, ориентированного на права человека.⁷⁷

Добросовестность научных исследований

Если субъект данных выходит из проекта научного исследования, отзыв согласия не влияет на законность обработки, основанной на согласии до его отзыва. Личные данные должны быть уничтожены или обезличены таким образом, чтобы не поставить под угрозу научную обоснованность исследования, а субъект данных должен быть соответствующим образом проинформирован.⁷⁸

III. Здоровье

Европейская нормативно-правовая база в сфере здравоохранения характеризуется наличием нескольких обязательных документов Совета Европы и ряда отраслевых документов, принятых на уровне ЕС, в соответствии с различным характером, сферой действия и полномочиями этих двух организаций в сфере регулирования.

Европейская конвенция по правам человека, а также Конвенция 108+ и Европейская социальная хартия содержат ряд общих положений об охране здоровья и смежных правах. Однако эти положения и принципы, уже изложенные в других общих документах на международном уровне,⁷⁹ в Конвенции Овьедо находят более широкую и более отраслевую контекстуализацию.

Поэтому Конвенция Овьедо – единственный многосторонний обязательный документ, полностью ориентированный на биомедицину, – и дополнительные протоколы к ней являются основным ориентиром для

⁷⁷ См. ICDDPPC; ОЭСР; GAI, п. III.9; ЮНЕСКО. См. также CM/Rec(2020)1.

⁷⁸ См. Конвенцию 108+; CM/Rec(2019)2.

⁷⁹ Некоторые из общих принципов, закрепленных в данной конвенции, были подтверждены в предыдущих международных документах по правам человека, таких как Международный пакт о гражданских и политических правах, Международный пакт об экономических, социальных и культурных правах и Конвенция о правах ребенка от 20 ноября 1989 года.

определения ключевых принципов в этой области,⁸⁰ которые нуждаются в дальнейшей проработке и, при необходимости, усилении для регулирования применения ИИ. Кроме того, Конвенция дополняется двумя необязательными документами: Рекомендацией о данных⁸¹ и Рекомендация об исследованиях биологических материалов человеческого происхождения.⁸² Первая из этих двух рекомендаций иллюстрирует тесную связь между биомедициной (и здравоохранением в целом) и обработкой данных, которая будет рассмотрена ниже.

Большая часть существующего регулирования в сфере здравоохранения сосредоточена на медицинском лечении, исследованиях (включая медицинские испытания) и медицинских устройствах/продуктах. ИИ потенциально может повлиять на все эти сферы, учитывая его применение в точной медицине,⁸³ диагностике, а также в медицинских устройствах и услугах.

Хотя Конвенция Овьедо и связанные с ней необязательные документы были приняты в эпоху до появления ИИ, они предусматривают конкретные гарантии в отношении самоопределения, обработки генома человека и исследований с участием человека, которые не затрагиваются применением ИИ в этой области и не требуют изменений.

Тем не менее, самоопределение в области биомедицины сталкивается с теми же проблемами, которые уже обсуждались при обработке данных. Несмотря на различную природу согласия на медицинское лечение и согласия на обработку данных, высокий уровень сложности и, зачастую, определенная степень неясности применения ИИ может подрвать эффективное осуществление автономии личности в обоих случаях.

На этом фоне основной вклад Конвенции Овьедо в будущее регулирование ИИ касается не тех гарантий, которые она предоставляет в конкретных секторах, а заключается в важном наборе общих принципов и

80 См. Andorno, 2005; Seatzu, 2015.

81 См. Рекомендацию CM/Rec(2019)2 о защите данных, связанных со здоровьем.

82 См. Рекомендацию CM/Rec(2016)6 об исследованиях биологических материалов человеческого происхождения.

83 См. Azencott, 2018; Ferryman and Pitcan, 2018.

ценностей, которые могут быть экстраполированы из нее, чтобы сформировать строительный блок будущего регулирования ИИ.

Основной вклад Совета Европы в области медицины касается следующих восьми областей: человеческое достоинство, права человека, профессиональные стандарты, общее правило информированного согласия, частная жизнь и право на информацию, недискриминация, защита лиц, проходящих исследования, и общественные дебаты. Вклад данной Конвенции в обсуждение будущего регулирования ИИ выходит за рамки биомедицины, так как некоторые положения, сосредоточенные на правильном балансе между технологиями и правами человека, могут быть распространены в целом за пределы области ИИ, как описано в следующих пунктах.⁸⁴

Главенство человека

В геополитическом и экономическом контексте, характеризующемся конкуренцией в развитии ИИ, примат человека в целом должен быть подтвержден в качестве ключевого элемента европейского подхода: лучшие показатели систем на базе ИИ и их эффективность не должны превалировать над интересами и благополучием людей. Применение этого принципа должно охватывать как разработку (например, системы, созданные с нарушением прав и свобод человека), так и использование систем ИИ.⁸⁵

Равный доступ к медицинскому обслуживанию

Принцип справедливого доступа может быть распространен на доступ к преимуществам ИИ. Это подразумевает принятие соответствующих мер для устранения рисков, связанных с цифровым разрывом, дискриминацией,

⁸⁴ Человеческое достоинство и информированное согласие не включены в таблицу, поскольку первое является ценностью, общей для документов, принятых Советом Европы в области прав человека, демократии и верховенства закона, а информированное согласие – это принцип, который также актуален в контексте обработки данных.

⁸⁵ См. также Конвенцию Овьедо, статья 2..

маргинализацией уязвимых лиц или культурных меньшинств и ограничениями в доступе к информации.⁸⁶

Профессиональные стандарты

Поэтому разработка ИИ охватывает несколько областей знаний, и в тех случаях, когда разработка систем ИИ может оказать воздействие на человека и общество, она должна осуществляться в соответствии с соответствующими профессиональными обязательствами и стандартами каждой области знаний. Необходимые профессиональные стандарты и навыки должны быть основаны на современном уровне техники.⁸⁷

Государства поощряют профессиональную подготовку для повышения осведомленности и понимания ИИ и его потенциальных последствий для отдельных лиц и общества. Они должны поддерживать исследования в области ИИ, ориентированные на права человека. Государства также должны сотрудничать в определении общих образовательных программ и общих стандартов для специалистов, занимающихся вопросами ИИ и общества.

При использовании ИИ в сфере здравоохранения особое внимание должно уделяться доверию пациента к своему врачу и взаимному доверию, которые не должны быть скомпрометированы использованием ИИ.

Защита лиц, не способных дать согласие, и лиц, не способных дать согласие на проведение исследований

Соблюдение принципа благодеяния должно рассматриваться как требование в тех случаях, когда, учитывая сложность или непрозрачность методов лечения на основе ИИ, индивидуальное согласие страдает рядом ограничений и не может быть исключительной основой для лечения.⁸⁸

86 См. также Конвенцию Овьедо, статья 3.

87 См. также Конвенцию Овьедо, статья 4.

88 См. также Конвенцию Овьедо, статьи 6 и 17.

Частная жизнь и право на информацию

Согласно Статье 10 Конвенции Овьедо, медицинские приложения ИИ должны гарантировать право на информацию и уважать желание человека не быть информированным, за исключением случаев, когда соблюдение желания человека не быть информированным представляет серьезный риск для здоровья других людей.⁸⁹

Принцип борьбы с дискриминацией в области здравоохранения должен быть дополнен запретом на любую форму дискриминации в отношении человека или группы людей на основании прогнозов будущего состояния здоровья.⁹⁰

Роль экспертов

Следует учитывать опыт комитетов по этике в области биомедицины, внедряя междисциплинарные комитеты экспертов по оценке приложений ИИ.⁹¹

Общественные дебаты

Фундаментальные вопросы, возникающие в связи с развитием ИИ, должны стать предметом соответствующего общественного обсуждения в свете, в частности, соответствующих социальных, экономических, этических и правовых последствий, и чтобы их возможное применение стало предметом соответствующих консультаций.⁹²

Эти соображения показывают, что существующая правовая база в области биомедицины содержит важные принципы и элементы, которые могут быть распространены на будущее регулирование ИИ, даже за пределами сектора здравоохранения.

С другой стороны, ряд недостатков, вызванных влиянием ИИ, остается нерешенным в следующих областях.

89 См. также Конвенцию Овьедо, статья 10.

90 См. также Конвенцию Овьедо, статья 11.

91 См. также Конвенцию Овьедо, ст. 16.

92 См. также Конвенцию Овьедо, ст. 28.

Системы принятия решений [Контекстуальный подход, Справедливость, Качество данных, Человеческий контроль/надзор].

В последние годы все больше приложений ИИ разрабатывается и используется в медицинском секторе для диагностики с применением аналитических и мл-решений. Создаются масштабные массивы данных, и предиктивная аналитика используется для того, чтобы попытаться найти решения для клинических случаев на основе существующих знаний и практики. Аналогичным образом, мл-приложения в распознавании изображений, похоже, могут повысить возможности обнаружения рака. Кроме того, в области точной медицины широкомасштабный сбор и анализ многочисленных источников данных (медицинских данных, а также немедицинских данных, таких как качество воздуха и жилья) используются для разработки индивидуальных представлений о здоровье и болезни.

Использование клинических данных, медицинских знаний и практики, а также немедицинских данных само по себе не является чем-то новым в медицине и исследованиях общественного здравоохранения. Однако масштабы сбора данных, детализация собранной информации, сложность (а в некоторых случаях и непрозрачность) обработки данных и прогностический характер результатов анализа вызывают опасения относительно потенциальной слабости систем принятия решений.

Большинство из этих вопросов не ограничивается сектором здравоохранения, поскольку потенциальная необъективность (включая отсутствие разнообразия и исключение аутсайдеров и небольших групп населения), качество данных, деконтекстуализация, контекстуальный характер маркировки данных⁹³ и повторное использование данных являются общими.⁹⁴ В соответствии с принятой методологией,⁹⁵ руководство в области защиты

93 Ferryman and Pitcan, 2018, 19-20 («Поскольку обозначения заболеваний, таких как сепсис, не являются четкими, отдельные обозначения могут использоваться для описания совершенно разных клинических реалий» и «эти записи были созданы не для исследований, а для выставления счетов, что может быть источником систематических ошибок и предвзятости»).

94 См. выше раздел II.2.

95 См., Общие руководящие ценности в области защиты данных и правосудия.

данных может быть применено и в этом случае, а аспекты качества данных распространены на неперсональные данные.

Самоопределение [свобода выбора/автономия, осознанность]

Непрозрачность приложений ИИ и трансформационное использование данных в крупномасштабном анализе данных подрывают традиционное понятие согласия как при обработке данных⁹⁶ так и при медицинском лечении, что предполагает принятие новых схем – таких как широкое⁹⁷ или динамическое согласие – которые, однако, могут лишь частично способствовать решению этой проблемы.

Отношения между врачом и пациентом

Несколько факторов, касающихся диагностики на основе ИИ, таких как потеря знаний, которые не могут быть закодированы в данных,⁹⁸ чрезмерная зависимость от ИИ при принятии медицинских решений, влияние местной практики на наборы данных для обучения и потенциальное увольнение в медицинском секторе⁹⁹ могут повлиять на отношения между врачом и пациентом¹⁰⁰ и должны быть оценены при внедрении ИИ в этой области.

96 См. также Рекомендацию CM/Rec(2019)2 о защите данных, связанных со здоровьем.

97 См. также Конвенцию 108+. Пояснительный отчет, 43 («В контексте научных исследований часто невозможно полностью определить цель обработки персональных данных для целей научных исследований в момент сбора данных. Поэтому субъектам данных должно быть разрешено давать свое согласие на определенные области научных исследований в соответствии с признанными этическими стандартами научных исследований. Субъекты данных должны иметь возможность дать свое согласие только на определенные области исследований или части исследовательских проектов в той степени, в которой это допускается предполагаемой целью») и Рекомендация CM/Rec(2019)2 о защите данных, связанных со здоровьем, 15.6 («Поскольку не всегда возможно заранее определить цели различных исследовательских проектов в момент сбора данных, субъекты данных должны иметь возможность выразить согласие на определенные области исследований или определенные части исследовательских проектов в той степени, в которой это допускается предполагаемой целью, с должным учетом признанных этических стандартов»).

98 См. Каруана и др., 2015.

99 См. Cabitza, Rasoini, and Gensini, 2017.

100 См. также Хельсинкскую декларацию ВМА – Этические принципы проведения медицинских исследований с участием человека, 9 июля 2018 года, <https://www.wma.net>.

Управление рисками [Риск-ориентированный подход, Подотчетность]

Область медицинских изделий¹⁰¹ представляет собой интересный пример управления рисками, учитывая значительные последствия, которые использование этих изделий может иметь для людей. Европейский Союз уже принял модель классификации¹⁰² на основе риска, основанную на прогрессивных гарантиях в зависимости от класса риска каждого устройства (от процедур оценки соответствия под исключительной ответственностью производителя или вмешательства нотифицированного органа до инспекции нотифицированным органом и, в случаях наивысшего риска, требования предварительного разрешения перед выпуском на рынок).

Модель, основанная на таких прогрессивных гарантиях, может быть обобщена для будущего регулирования ИИ, а также принята вне сферы медицинских устройств, с акцентом на воздействие на права человека и основные свободы. Однако классификация продуктов/услуг ИИ является более сложной задачей, учитывая их разнообразие и различные области применения: следует ввести несколько отраслевых классификаций или принять общие критерии, основанные на процедурах оценки рисков.

Кроме того, конкретные положения о бдительности в отношении ИИ и принятие принципа предосторожности при разработке ИИ, как обсуждалось выше,¹⁰³ могут помочь в решении этих проблем.

IV. Демократия

Демократия охватывает чрезвычайно широкий спектр социальных и правовых вопросов,¹⁰⁴ большинство из которых, вероятно, будут реализованы при поддержке ИКТ¹⁰⁵. В этом сценарии ИИ может сыграть важную роль в

101 См. также Европейская комиссия, 2014.

102 См. директиву 93/42/ЕЕС.

103 См. выше раздел II.2.

104 См. например, Совет Европы. Генеральный директорат по демократии – Европейский комитет по демократии и управлению. 2016. Сборник наиболее актуальных текстов Совета Европы в области демократии.

105 См. например, Генеральный директорат по демократии и политическим вопросам – Директорат демократических институтов. 2009. Проект «Надлежащее управление в информационном обществе», СМ(2009)9 Addendum 3. Руководства по показателям и глоссарий, относящиеся к Рекомендации Rec(2009) 1 Комитета министров для государств-членов по электронной демократии (e-democracy), подготовленной

настоящем и будущем развитии цифровой демократии в информационном обществе.

По сравнению с другими рассмотренными областями (защита данных и здравоохранение), широкое измерение этой темы затрудняет определение единого обязательного отраслевого правового инструмента для ссылки. Демократия и ее различные аспекты рассматриваются в нескольких международных документах, начиная с Декларации прав человека ООН и Международного пакта о гражданских и политических правах. Аналогичным образом, в европейском контексте ключевые принципы демократии присутствуют в нескольких международных источниках.

Основываясь на статье 25 МПГПП, мы можем выделить две основные области вмешательства: (I) участие¹⁰⁶ и надлежащее управление, и (II) выборы. Несомненно, трудно или невозможно провести красную линию между этими областями, поскольку они взаимосвязаны различными способами. ИИ может оказать влияние на все из них: участие (например, вовлечение граждан, платформы участия), надлежащее управление (например, электронное правительство, процессы принятия решений, умные города), предвыборный этап (например, финансирование, таргетинг и профилирование, пропаганда), выборы (например, предсказание результатов выборов, электронное голосование) и послевыборный период (например, разрешение избирательных споров).

Как и в любой классификации, это различие характеризуется пределом направленности. Стоит отметить, что это функциональная классификация, основанная на различных воздействиях ИИ, без намерения дать юридическое или политическое представление о демократии и ее различных ключевых элементах. Поэтому взаимосвязь между участием, надлежащим управлением и

Специальным комитетом Совета Европы по электронной демократии (CANDE); Дополнительный протокол к Европейской хартии местного самоуправления о праве на участие в делах местного органа власти, 2009, Статья 2.2.III.

106 Более подробный анализ см. в Faye Jacobsen, 2013. См. также Maisley, 2017.

выборами можно рассматривать с разных сторон и формировать по-разному, объединяя определенные области или подразделяя их еще больше.

Участие выражается как через участие в демократических дебатах, так и через избирательный процесс, но то, как инструменты ИИ взаимодействуют с участием в этих двух случаях, отличается, и существуют отдельные международные правовые инструменты, специфичные для избирательного процесса.

Участие и эффективное управление

Право на участие в общественных делах (статья 25 Пакта) основано на широком понятии «общественные дела»,¹⁰⁷ которое включает в себя публичные дебаты и диалог между гражданами и их представителями, с тесной связью со свободой выражения мнений, собраний и ассоциаций.¹⁰⁸ В этом отношении ИИ актуален с двух разных точек зрения: как средство для участия и как предмет решений, принимаемых на основе участия.

Если рассматривать ИИ как средство, то технические и образовательные барьеры могут подрывать осуществление права на участие. Поэтому инструменты участия, основанные на ИИ, должны учитывать риски недопредставленности и непрозрачности процессов участия (например, платформы для разработки законопроектов). В то же время, ИИ также является предметом решений, основанных на участии, поскольку они включают решения о развитии ИИ в целом и его использовании в государственных делах.

Основанные на ИИ партисипативные платформы (например, Consul,¹⁰⁹ Citizenlab,¹¹⁰ Decidim¹¹¹) могут внести значительный вклад в демократический

107 См. Управление Верховного комиссара ООН по правам человека. 1996. Замечание общего порядка № 25: Право на участие в общественных делах, право голоса и право равного доступа к государственной службе (ст. 25). CCPR/C/21/Rev.1/Add.7.

108 См. также Управление Верховного комиссара ООН по правам человека. 1981. Замечание общего порядка № 1 КЭСКП: Отчетность государств-участников, п. 5 («содействовать общественному контролю за политикой правительства в отношении экономических, социальных и культурных прав и поощрять участие различных экономических, социальных и культурных секторов общества в разработке, осуществлении и пересмотре соответствующей политики»).

109 См. <<https://consulproject.org/en/>>

110 См <<https://www.citizenlab.co/>>

111 См <<https://decidim.org/>>

процесс, способствуя взаимодействию граждан, определению приоритетности целей и совместным подходам в принятии решений¹¹² по темам, представляющим общий интерес на различных уровнях (район, муниципалитет, мегаполис, регион, страна).¹¹³ Поскольку эти платформы используются в социальной среде и собирают информацию, те же аспекты, которые уже обсуждались в отношении защиты данных, включая безопасность, можно вспомнить и здесь, распространив на эти приложения рекомендации, рассмотренные в предыдущем разделе о данных.

Однако возникают и другие, более конкретные вопросы, связанные с инструментами ИИ для демократического участия (в том числе для предотвращения и борьбы с коррупцией¹¹⁴), которые ассоциируются со следующими четырьмя основными областями: прозрачность, подотчетность, инклюзивность и открытость. В этом отношении общие принципы, изложенные в международных обязательных документах, имеют важное воплощение в Рекомендации CM/Rec(2009)1 Комитета министров для государств-членов по электронной демократии (e-democracy), которая обеспечивает основу для дальнейшей разработки руководящих принципов в области ИИ в отношении демократии.

Прозрачность является требованием для использования технологических приложений в демократических целях.¹¹⁵ Этот принцип является общим для областей, проанализированных выше, данных и здравоохранения. Однако прозрачность — это понятие, зависящее от контекста. Если в этих областях прозрачность тесно связана с самоопределением, то здесь она приобретает более широкое значение. В демократическом процессе прозрачность является не только требованием для самоопределения граждан в отношении технического инструмента, но и компонентом демократического процесса

112 См. также Совет Европы. Руководящие принципы гражданского участия в принятии политических решений. CM(2017)83-final. Принято Комитетом министров 27 сентября 2017 года на 1295-м заседании заместителей министров.

113 См. также Рекомендацию CM/Rec(2009)2 по оценке, аудиту и мониторингу участия и политики участия на местном и региональном уровне.

114 См. Конвенцию ООН против коррупции, 2003, Статья 13.

115 См. Рекомендацию CM/Rec(2009)1 об электронной демократии (e-democracy), параграф 6.

участия.¹¹⁶ Прозрачность больше не имеет индивидуального измерения, а приобретает коллективное измерение как гарантия демократического процесса.

В этом контексте использование решений на основе ИИ для электронной демократии должно быть прозрачным в отношении их логики и функционирования (например, отбор контента на платформах с широким участием), предоставляя четкую, легкодоступную, понятную и обновляемую информацию об используемых инструментах ИИ.¹¹⁷

Более того, при реализации этого понятия прозрачности следует также учитывать круг различных пользователей этих инструментов,¹¹⁸ применяя доступный подход с ранних этапов разработки инструментов ИИ. Это необходимо для обеспечения эффективной прозрачности в отношении уязвимых и ослабленных групп населения, придавая дополнительную ценность доступности в этом контексте.

Прозрачность и доступность тесно связаны с характером архитектуры, используемой для создания систем ИИ. Поэтому открытый исходный код и открытые стандарты¹¹⁹ могут способствовать демократическому надзору за наиболее важными приложениями ИИ.¹²⁰ Бывают случаи, когда открытость страдает от ограничений, обусловленных характером конкретного применения ИИ (например, предотвращение преступлений). В этих случаях проверяемость, а также схемы сертификации играют более важную роль, чем они уже играют в отношении систем ИИ в целом.¹²¹

116 См. также Руководящие принципы гражданского участия в принятии политических решений. CM(2017)83-final, IV.

117 См. Рекомендацию CM/Rec(2009)1 об электронной демократии (е-демократии), п. 6 («облегчает и расширяет доступ, доступность [...] путем использования, где это возможно, прозрачных [...] средств») и Приложение к Рекомендации CM/Rec(2009)1, п. P.57. См. также Рекомендацию CM/Rec(2016)5 о свободе Интернета. Приложение, пункты 2.1.3 и 3.2.

118 См. также Рекомендацию CM/Rec(2018)4 об участии граждан в местной общественной жизни, Приложение, п. B.IV..

119 См. также Рекомендацию CM/Rec(2009)1 об электронной демократии (e-democracy), п. 6 и Приложение, параграф P.54.

120 См. также Рекомендацию CM/Rec(2009)1 об электронной демократии (e-democracy), Приложение, п. G.58.

121 84 Стоит подчеркнуть, что схемы аудита и сертификации играют важную роль и в случаях архитектуры ИИ с открытым исходным кодом, поскольку такой характер не подразумевает само по себе отсутствие предвзятости или каких-либо других недостатков. См. также Рекомендацию CM/Rec(2009)1 об электронной демократии (e-democracy), Приложение, пункты P.55 и G.57 («Программное обеспечение для электронной

В контексте применения ИИ для содействия демократическому участию важную роль может также играть интероперабельность¹²² поскольку она облегчает интеграцию между различными сервисами/платформами для электронной демократии и на различных географических уровнях. Этот аспект уже актуален для электронной демократии в целом,¹²³ и поэтому должен быть распространен на разработку систем на основе ИИ.

Еще одним ключевым принципом электронной демократии, как и в секторах данных и здравоохранения, является подотчетность. В отличие от предыдущих рассмотренных принципов, подотчетность не приобретает здесь иного значения, и поэтому не кажется, что в контексте ЭИ требуется реализация в конкретном секторе, кроме ее общего применения.

Наконец, учитывая роль СМИ в контексте демократического участия и в соответствии с Рекомендацией CM/Rec(2016) Комитета министров Совета Европы,¹²⁴ приложения ИИ не должны ставить под угрозу конфиденциальность и безопасность коммуникаций и защиту журналистских источников и информаторов.¹²⁵

При рассмотрении различных аспектов разработки решений ИИ для демократического участия, прежде всего, необходимо учитывать, что демократический подход несовместим с техно-детерминистским подходом. Поэтому решения ИИ для решения проблем общества должны быть результатом инклюзивного процесса. Следовательно, такие ценности, как защита меньшинств, плюрализм и разнообразие, должны быть необходимым условием при разработке этих решений.

демократии должно быть либо с открытым исходным кодом, который можно проверить, либо, в качестве альтернативы, сертифицировано независимым органом»).

122 См. также Рекомендацию CM/Rec(2009)1 об электронной демократии (e-democracy), Приложение, пункты P. 56, G.56, 59 и 60.

123 См. также Рекомендацию CM/Rec(2009)1 об электронной демократии (e-democracy), п. 6.

124 См. Рекомендацию CM/Rec(2016)4 о защите журналистики и безопасности журналистов и других работников СМИ, Приложение, п. 2; Совет Европы, Парламентская ассамблея. 2019. Резолюция 2254 (2019)1. Свобода СМИ как условие демократических выборов.

125 См. также Парламентская ассамблея, Резолюция 2300 (2019)1, Улучшение защиты информаторов по всей Европе; Рекомендация CM/Rec(2014)7 о защите информаторов.

С точки зрения демократии, первый вопрос, который мы должны задать: действительно ли нам нужно решение данной проблемы на основе ИИ,¹²⁶ в отличие от других вариантов, учитывая потенциальное влияние ИИ на права и свободы? Если ответ на этот вопрос положительный, то следующим шагом будет изучение вопроса о внедрении ценностей в разработку ИИ.¹²⁷

Предлагаемые решения в области ИИ должны разрабатываться с точки зрения прав человека, обеспечивая полное соблюдение прав человека и основных свобод, включая принятие для этой цели инструментов и процедур оценки.¹²⁸ В случае применения ИИ, оказывающего большое влияние на права и свободы человека, например, в избирательных процессах, необходимо предварительно оценить соответствие законодательству. Кроме того, системы ИИ для решения государственных задач должны быть проверяемы, и, если это не исключается конкурирующими преобладающими интересами, результаты проверок должны быть общедоступны.

Еще один важный аспект, который необходимо рассмотреть, — это партнерство государственного и частного секторов, которое часто характеризует услуги ИИ для граждан, взвешивание оптимального выбора между собственными и сторонними решениями, включая множество различных комбинаций этих двух крайностей. В этом отношении, когда решения ИИ полностью или частично разрабатываются частными компаниями, прозрачность контрактов и четкие правила доступа и использования данных граждан имеют решающее значение с точки зрения демократического надзора.

Ограничения на доступ и использование данных граждан актуальны не только с точки зрения защиты данных (принципы минимизации данных и ограничения целей), но и в более широком смысле в отношении большого объема данных, генерируемых сообществом, которые также включают

126 См. также Рекомендацию CM/Rec(2020)1, Приложение, п. 5.7.

127 См. также Декларацию Комитета министров о манипулятивных возможностях алгоритмических процессов (Принята Комитетом министров 13 февраля 2019 года на 1337-м заседании заместителей министров, п. 7.

128 См. Рекомендацию CM/Rec(2009)1, пункты 5 и 6, и Приложение к Рекомендации CM/Rec(2009)1, п. G.67. См. также выше раздел II.2 о данных и роли комитетов экспертов и A Mantelero, 'AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment' (2018) 34 Computer Law & Security Review 754.

неперсональные данные и агрегированные данные. Этот вопрос следует рассматривать как компонент демократии в цифровой среде, где коллективное измерение цифровых ресурсов, создаваемых сообществом, должно подразумевать формы гражданского контроля и надзора, как это происходит в отношении других ресурсов территории/сообщества (например, окружающей среды). Здесь следует напомнить уже высказанные выше соображения об открытости как ключевом элементе инструментов демократического участия, учитывая их влияние на разработку систем ИИ. Кроме того, при проектировании, разработке и внедрении этих систем следует также учитывать принятие экологически безопасной и устойчивой стратегии.¹²⁹

Наконец, стоит отметить, что хотя ИИ-дизайн является ключевым компонентом этих систем, дизайн не является нейтральным. Ценности могут быть заложены в технологические артефакты,¹³⁰ включая системы ИИ. Эти ценности могут быть выбраны намеренно, и в контексте электронной демократии это должно быть основано на демократическом процессе. Но ценности также могут быть непреднамеренно заложены в решения ИИ, в силу культурного, социального и гендерного состава команд разработчиков ИИ. По этой причине инклюзивность имеет здесь дополнительную ценность с точки зрения вовлечения и разнообразия¹³¹ в разработку ИИ.

Что касается надлежащего управления,¹³² то здесь можно повторить принципы, рассмотренные для электронной демократии.¹³³ Это касается «умных» городов и сенсорного управления окружающей средой, где центральную роль играют открытые, прозрачные и инклюзивные процессы принятия решений. Аналогичным образом, использование ИИ для надзора за

129 См. также Рекомендацию CM/Rec(2009)1, Приложение, п. Р. 58.

130 См. также P-P Verbeek, 2011, 41 -65.

131 См. также P-P Verbeek, 2011, 41 -65.

132 См. Рекомендацию CM/Rec(2009)1 об электронной демократии (е-демократии), Приложение, п.п. Р.4 («[...] надлежащее управление, которое представляет собой эффективное, действенное, партисипативное, прозрачное и подотчетное демократическое осуществление власти в электронной форме и включает неформальную политику и неправительственных игроков»).

133 См. также Рекомендацию Rec(2004)15 об электронном управлении («электронное управление»); Совет Европы. 2008. 12 принципов надлежащего управления, закрепленных в Стратегии по инновациям и надлежащему управлению на местном уровне, одобренной решением Комитета министров Совета Европы в 2008 году.

деятельностью местных органов власти в целях аудита,¹³⁴ и борьбы с коррупцией должно быть основано на открытости,¹³⁵ (программное обеспечение с открытым исходным кодом), прозрачности и проверяемости.

В более широком смысле ИИ может использоваться во взаимодействии правительства и граждан для автоматизации запросов граждан и запросов информации.¹³⁶ Однако в этих случаях важно гарантировать право знать, что мы взаимодействуем с машиной¹³⁷ и иметь контакт с человеком. Более того, доступ к государственным услугам не должен зависеть от предоставления данных, которые не нужны и не соразмерны цели.

Особое внимание также следует уделить потенциальному использованию ИИ в человеко-машинном взаимодействии для реализации стратегий подталкивания.¹³⁸ Здесь, в силу сложности и неясности принимаемых технических решений, ИИ может усилить пассивную роль граждан и негативно повлиять на демократический процесс принятия решений.

В противном случае, активный подход, основанный на сознательном и активном участии в достижении целей сообщества, должен быть предпочтительнее и лучше управляться инструментами участия ИИ. В случае принятия стратегий подталкивания следует придерживаться подхода, основанного на фактических данных.

Наконец, использование систем ИИ для решения задач управления поднимает сложные вопросы о взаимоотношениях между людьми, принимающими решения, и ролью ИИ в процессе принятия решений.¹³⁹ Эти вопросы более актуальны в отношении функций, оказывающих большое влияние на права и свободы личности, как в случае с юрисдикционными решениями. По этой причине в разделе 5 будут проанализированы вопросы

134 См. также Рекомендацию CM/Rec(2019)3 о надзоре за деятельностью местных органов власти, Приложение, Руководство по совершенствованию систем надзора за деятельностью местных органов власти, пункты 4 и 9.

135 См. также Savaget, Chiarini and Evans, 2019, где обсуждается бразильский случай «Opera Serenata de Amor» (OSA).

136 См. далее. 2017.

137 См. также GAI 2.11.

138 См. exmultis, Sunstein, 2015a; Sunstein, 2015a; Sunstein and Thaler, 2003; Thaler and Sunstein, 2008.

139 См. также Calo and Citron, 2020, Forthcoming.

прозрачности (включая объяснимость) рассуждений ИИ и взаимосвязь между использованием ИИ и свободой лиц, принимающих решения.

Выборы

Как и в других областях, влияние ИИ на избирательные процессы широко и по-разному затрагивает этапы до выборов, во время выборов и после них. Однако анализ, сфокусированный на этапах избирательного процесса, не позволяет адекватно выделить различные способы взаимодействия решений ИИ с ним.

Поэтому влияние ИИ лучше представить в виде следующего разграничения: ИИ для избирательного процесса (электронное голосование, прогнозирование результатов и разрешение избирательных споров) и ИИ для избирательных кампаний (микротаргетинг и профилирование, пропаганда и фальшивые новости). Если в первой области ИИ — это в основном технологическое усовершенствование существующего процесса, то в сфере избирательных кампаний профилирование и пропаганда на основе ИИ вызывают новые проблемы, которые лишь частично учитываются существующей правовой базой. Кроме того, в ряде документов подчеркивается активная роль государств в создании благоприятной среды для свободы выражения мнений.¹⁴⁰

Что касается технологической реализации электронной демократии (электронное голосование, предсказание результатов и разрешение избирательных споров), то здесь также актуальны некоторые из ключевых принципов, упомянутых в отношении демократического участия.

140 См. Рекомендацию CM/Rec(2018)1 о плюрализме СМИ и прозрачности владения СМИ; Совместную декларацию о «фальшивых новостях», дезинформации и пропаганде, Специальный докладчик Организации Объединенных Наций (ООН) по вопросам свободы мнений и их выражения, Представитель Организации по безопасности и сотрудничеству в Европе (ОБСЕ) по вопросам свободы СМИ, Специальный докладчик Организации американских государств (ОАГ) по вопросам свободы выражения мнений и Специальный докладчик Африканской комиссии по правам человека и народов (АКПЧН) по вопросам свободы выражения мнений и доступа к информации (3 марта 2017 года). См. также Рекомендацию CM/Rec(2016)5 о свободе Интернета, Приложение, пункты 1.5, 2.1 и 3; Европейская комиссия за демократию через право (Венецианская комиссия). 2019. Совместный доклад Венецианской комиссии и Директората информационного общества и действий против преступности Генерального директората по правам человека и верховенству права (DGI) о цифровых технологиях и выборах, п. 151 .Е; OSCE, 2020. См. также Bychawska-Siniarska, 2017.

Доступность,¹⁴¹ прозрачность,¹⁴² открытость,¹⁴³ управление рисками и подотчетность (включая принятие процедур сертификации и аудита)¹⁴⁴ являются основополагающими элементами технологических решений, принятых на этих этапах избирательного процесса.

Что касается ИИ для агитации (микротаргетинг и профилирование, пропаганда и фальшивые новости), некоторые из поднятых вопросов касаются обработки персональных данных в целом. Поэтому принципы, изложенные в Конвенции 108+, могут быть применены и надлежащим образом контекстуализированы.¹⁴⁵

В случае пропаганды и дезинформации необходимы более конкретные и новые ответы.¹⁴⁶ Здесь существующие обязательные и необязательные инструменты не устанавливают конкретных положений, учитывая новизну дезинформации, основанной на новых формах коммуникации, таких как социальные сети, которые отличаются от традиционных¹⁴⁷ СМИ и часто обходят профессиональное посредничество журналистов.

Однако общие принципы, такие как принцип невмешательства государственных органов в деятельность СМИ по оказанию влияния на выборы,¹⁴⁸ могут быть распространены на эти новые формы пропаганды и дезинформации. Учитывая использование ИИ для автоматизации пропаганды, будущее регулирование ИИ должно распространить сферу действия общих принципов невмешательства на системы на основе ИИ, используемые для

141 См. Рекомендацию CM/Rec(2017)5 о стандартах электронного голосования, Приложение I, Стандарты электронного голосования, пункты 1 и 2..

142 См. Рекомендацию CM/Rec(2017)5, Приложение I, п. 32. См. также Совет Европы. Генеральный директорат по демократии и политическим вопросам – Директорат демократических институтов. 2011. Руководство по прозрачности выборов с использованием электронных средств.

143 См. Рекомендацию CM/Rec(2017)5, Приложение I, п. 35.

144 См. Рекомендацию CM/Rec(2017)5, Приложение I, пункты 36, 37, 38, 39 и 40.

145 Рекомендация CM/Rec(2010)13 о защите физических лиц в отношении автоматической обработки персональных данных в контексте профилирования и ее текущий обзор, см. Совет Европы, Консультативный комитет Конвенции о защите физических лиц в отношении автоматической обработки персональных данных. 2019. Профилирование и Конвенция 108+: Предложения по обновлению. T-PD(2019)07BISrev.

146 См. Manheim and Kaplan, 2019; Европейская комиссия – Сети, контент и технологии – Генеральный директорат по коммуникациям, «Многомерный подход к дезинформации Доклад независимой группы высокого уровня по фальшивым новостям и онлайн-дезинформации» (2018). См. также Stoll v. Switzerland [GC], no.69698/01, § 104.

147 См. также Рекомендацию CM/Rec(2011)7 о новом понятии СМИ.

148 См. Рекомендацию CM/Rec(2007)15 о мерах, касающихся освещения избирательных кампаний в СМИ, п. I.1.

предоставления ложной, вводящей в заблуждение и вредной информации. Кроме того, для предотвращения такого вмешательства государства¹⁴⁹ и провайдеры социальных сетей должны принять подход, основанный на принципе «by-design», чтобы повысить свою устойчивость к дезинформации и пропаганде.

Аналогичным образом, обязательство освещать избирательные кампании справедливым, сбалансированным и беспристрастным образом¹⁵⁰ должно повлечь за собой обязательства для СМИ и операторов социальных сетей в отношении прозрачности логики алгоритмов,¹⁵¹ используемых для отбора контента, обеспечения плюрализма и разнообразия мнений,¹⁵² включая критические.¹⁵³

Более того, государства и посредники должны продвигать и облегчать доступ к инструментам для обнаружения дезинформации и нечеловеческих агентов, а также поддерживать независимые исследования влияния дезинформации и проекты, предлагающие пользователям услуги по проверке фактов.¹⁵⁴

Учитывая важную роль, которую играет реклама в дезинформации и пропаганде, критерии, используемые решениями на основе ИИ для политической рекламы, должны быть прозрачными,¹⁵⁵ проверяемыми и обеспечивать равные условия для всех политических партий и кандидатов.¹⁵⁶

149 См. также Совместную декларацию о «фальшивых новостях», дезинформации и пропаганде, п. 2.с.

150 См. Рекомендацию CM/Rec(2007)15, п. II.1.

151 См. также Совместную декларацию о «фальшивых новостях», дезинформации и пропаганде; Рекомендацию CM/Rec(2018)2 о роли и обязанностях интернет-посредников, Приложение, пп. 2.1.3 и 2.3.5 («Ввиду нынешней ограниченной способности автоматизированных средств оценивать контекст, посредники должны тщательно оценивать воздействие автоматизированного управления контентом на права человека и должны обеспечивать человеческий обзор там, где это необходимо. Они должны учитывать риск чрезмерно строгого или слишком мягкого подхода, обусловленного неточными алгоритмическими системами, и влияние, которое эти алгоритмы могут оказать на услуги, которые они предоставляют для общественных дискуссий»).

152 См. также Кодекс практики ЕС по дезинформации, 2018.

153 См. также Рекомендацию CM/Rec(2016)4, Приложение, п. 15.

154 См. также Совместную декларацию о «фальшивых новостях», дезинформации и пропаганде, п. 4.е; Европейская комиссия за демократию и закон. 2019, пара. 151.D.

155 См. также Совет Европы. Парламентская ассамблея. Резолюция 2254 (2019)1. Свобода СМИ как условие демократических выборов, пункты 9.2 и 11.1; Европейская комиссия по демократии через право (Венецианская комиссия). 2019. Совместный доклад Венецианской комиссии и Директората информационного общества и действий против преступности Генерального директората по правам человека и верховенству права (DGI) о цифровых технологиях и выборах, пункты 151.A и 151.B.

156 См. также Рекомендацию CM/Rec(2007)15, п. II.5.

Кроме того, посредники должны пересмотреть свои рекламные модели, чтобы убедиться, что они не оказывают негативного влияния на разнообразие мнений и идей.¹⁵⁷

V. Правосудие

Как и в предыдущем разделе, сфера правосудия является обширной областью, и анализ всего спектра последствий использования ИИ для правосудия и связанных с ним последствий для демократии был бы слишком амбициозным. В соответствии с рамками данного исследования, в этом разделе описываются основные проблемы, связанные с использованием ИИ, и принципы, которые, основываясь на международных юридически обязательных документах, могут способствовать его будущему регулированию.

Правосудие отличается от защиты данных и здравоохранения отсутствием конкретных и специальных обязательных документов, таких как Конвенция 108+ и Конвенция Овьедо. Поэтому данный анализ в большей степени сосредоточен на контекстуализации общих руководящих принципов, чем на конкретных правовых инструментах.

Этому способствует Европейская этическая хартия по использованию искусственного интеллекта (ИИ) в судебных системах и их окружении, принятая СЕРЕJ в 2019 году, в которой напрямую рассматриваются отношения между правосудием и ИИ. Хотя этот не имеющий обязательной силы документ классифицируется как этическая хартия, в значительной степени он касается правовых принципов, закрепленных в международных документах.

Руководящие принципы для развития ИИ в области правосудия можно почерпнуть из следующих обязательных документов: Всеобщая декларация прав человека, Международный пакт о гражданских и политических правах, Конвенция о защите прав человека и основных свобод, Международная конвенция о ликвидации всех форм расовой дискриминации, Конвенция о

157 См. также Совместную декларацию о «фальшивых новостях», дезинформации и пропаганде, п. 4.e.

ликвидации всех форм дискриминации в отношении женщин и Конвенция о защите прав человека и основных свобод.

*Свободы.*¹⁵⁸

Учитывая диапазон типов и целей операций в этой области, а также различные профессиональные фигуры и процедуры, в данном разделе проводится функциональное разграничение между двумя областями: (I) судебные решения и альтернативные способы разрешения споров (АСР) и (II) предупреждение/прогнозирование преступлений. Прежде чем анализировать и контекстуализировать ключевые принципы, относящиеся к этим двум областям, мы должны предложить несколько общих замечаний, которые также могут быть применимы к деятельности государственной администрации в целом.¹⁵⁹

Прежде всего, стоит отметить, что по сравнению с человеческими решениями, а точнее судебными решениями, логика систем ИИ не похожа на юридические рассуждения. Вместо этого они просто выполняют коды, основанные на данных, ориентированных на математический/статистический подход.

Кроме того, уровень ошибок ИИ близок или ниже, чем у человеческого мозга в таких областях, как маркировка изображений, но в более сложных задачах принятия решений уровень ошибок выше. Так обстоит дело с юридическими рассуждениями при решении проблем.¹⁶⁰ В то же время, если неправильная классификация изображения кошки может иметь ограниченные

¹⁵⁸ См. также, применительно к территории ЕС, Хартию основных прав Европейского Союза.

¹⁵⁹ См. выше раздел II.4.

¹⁶⁰ См. Dupont et al., 2018, 148 («Deep Learning не имеет естественного способа работы с иерархической структурой, что означает, что все доступные переменные рассматриваются на одном уровне, как «плоские» или неиерархические. Это представляет собой серьезное препятствие, когда решения имеют большой моральный или юридический вес, который должен превалировать над другими характеристиками»). См. также Osoba and Welser, 2017, 18 («Другой угол зрения на проблему заключается в том, что суждения в пространстве социального поведения часто являются нечеткими, а не четко определенными бинарными критериями [...]. Мы способны научиться ориентироваться в сложных нечетких отношениях, таких как правительства и законы, часто полагаясь для этого на субъективные оценки. Системы, опирающиеся на количественные рассуждения (например, большинство искусственных агентов), могут имитировать этот эффект, но часто требуют для этого тщательной разработки. Для улавливания этого нюанса могут потребоваться не только специалисты по компьютерам и данным»). См. также Cummings et al., 2018, 13.

негативные последствия, то ошибки в юридических решениях оказывают большое влияние на права и свободу людей.

Стоит отметить, что разница между ошибками в принятии решений человеком и машиной имеет важное последствие с точки зрения масштаба: в то время как человеческие ошибки влияют только на отдельные случаи, плохое проектирование и предвзятость в ИИ неизбежно влияют на всех людей в тех же или похожих обстоятельствах, при этом инструменты ИИ применяются к целому ряду случаев. Это может привести к групповой дискриминации, негативно влияющей на лиц, принадлежащих к различным категориям.

Учитывая текстовый характер юридических документов, обработка естественного языка (NLP) играет важную роль в приложениях ИИ для сферы правосудия. В связи с этим возникает несколько критических вопросов, касающихся коммерческих решений, разработанных с ориентацией на англоязычный рынок, что делает их менее эффективными в правовой среде, где используются не только английский язык.¹⁶¹ Более того, юридические решения часто характеризуются неявными невыраженными рассуждениями, которые могут быть применимы к экспертным системам, но не к инструментам машинного обучения на основе языка. Наконец, наличие общих положений требует предварительного знания соответствующей юридической интерпретации и постоянного обновления, что невозможно получить из анализа текста.

Все эти ограничения предполагают осторожное и более критическое применение ИИ в сфере правосудия, чем в других областях, а в отношении судебных решений и ADR следует проводить различие между делами, характеризующимися рутинной и основанной на фактах оценкой, и делами, характеризующимися значительным запасом юридического мышления и усмотрения.¹⁶²

161 Совет адвокатов и юридических обществ Европы, 2020, 29.

162 См. следующий раздел о различии между кодифицированным правосудием и справедливым правосудием.

Судебные решения и ADR

Несколько так называемых продуктов Legal Tech AI не оказывают прямого влияния на процессы принятия решений в судах или альтернативные способы разрешения споров (ADR), а скорее способствуют управлению контентом и знаниями, организационному управлению и оценке эффективности.¹⁶³ Эти приложения включают, например, инструменты для категоризации договоров, выявления расходящихся или несовместимых договорных положений, электронного обнаружения, помощи в составлении проектов, поиска правовых положений, проверки соблюдения законодательства. Кроме того, некоторые приложения могут предоставлять базовые функции решения проблем на основе стандартных вопросов и стандартных ситуаций (например, юридические чат-боты). Хотя в таких случаях ИИ оказывает влияние на юридическую практику и юридические знания, что поднимает различные этические вопросы,¹⁶⁴ потенциальные негативные последствия для прав человека, демократии и верховенства закона ограничены. В значительной степени они связаны с неэффективностью или недостатками этих систем.

В случае управления контентом и знаниями, включая исследования и анализ документов, эти недостатки могут порождать неполное или неточное представление фактов или ситуаций, но это влияет на мета-продукты, результаты исследовательского инструмента, которые должны быть интерпретированы и адекватно мотивированы при использовании в суде. Правила ответственности, например, в контексте ответственности за качество продукции, могут решить эти проблемы.

Кроме того, предвзятость (плохой отбор дел, неправильная классификация и т.д.), влияющая на стандартные текстовые компьютерные инструменты поиска для анализа законодательства, прецедентного права и литературы, может быть устранена путем соответствующего обучения и

163 См. Европейская комиссия по эффективности правосудия (CEPEJ). 2018. Европейская этическая хартия по использованию искусственного интеллекта в судебных системах и их окружении, Приложение II.

164 См. также Нуньес, 2017.

подготовки специалистов в области права, а прозрачность систем ИИ (т.е. описание их логики, потенциальной предвзятости и ограничений) может уменьшить негативные последствия.

Прозрачность должна также характеризовать использование судами ИИ для юридических исследований и анализа документов. Судьи должны быть прозрачными в том, какие решения зависят от ИИ и как результаты, предоставленные ИИ, используются для аргументации, в соответствии с принципами справедливого судебного разбирательства и равенства сторон.¹⁶⁵

Наконец, прозрачность может играть важную роль в отношении юридических чат-ботов, основанных на ИИ, позволяя пользователям понять их логику и используемые ресурсы (например, список проанализированных дел). Полная прозрачность также должна включать источники, используемые для обучения этих алгоритмов, и доступ к базе данных, используемой для предоставления ответов. Там, где эти базы данных являются частными, должны быть доступны аудиты третьей стороны для оценки качества наборов данных и того, как были устранены потенциальные предубеждения, включая риск недопредставленности или перепредставленности определенных категорий (недискриминация).

Другие критические вопросы затрагивают приложения ИИ, разработанные для автоматизации альтернативного разрешения споров или поддержки судебных решений. Здесь различие между кодифицированным правосудием и справедливым правосудием¹⁶⁶ предполагает, что для целей принятия решений ИИ следует ограничить делами, характеризующимися рутинными и основанными на фактах оценками. Это предполагает важность проведения дальнейших исследований по классификации различных видов процессов принятия решений, чтобы определить те рутинизированные

165 См. также Европейская комиссия по эффективности правосудия (СЕРЕЈ). 2018. Европейская этическая хартия по использованию искусственного интеллекта в судебных системах и их окружении.

166 См. Re and Solow-Niederman, 2019, 252-254 («Справедливое правосудие подразумевает как размышления о ценностях, установленных правовой системой, так и обоснованное применение этих ценностей в контексте [...] Кодифицированное правосудие относится к рутинному применению стандартных процедур к набору фактов [...] Короче говоря, кодифицированное правосудие видит пороки усмотрения, тогда как справедливое правосудие видит его достоинства»).

приложения юридических рассуждений, которые можно поручить ИИ, сохраняя в любом случае человеческий обзор, который также гарантирует юридическую креативность лиц, принимающих решения.¹⁶⁷

Что касается справедливого правосудия, то, как отмечается в литературе,¹⁶⁸ его логика сложнее, чем простой исход отдельных дел. Выраженные и невыраженные ценности и соображения, как юридические, так и неюридические, характеризуют аргументацию судов и не воспроизводятся логикой ИИ. Системы на основе ОД не способны выполнять юридические рассуждения. Они извлекают выводы, выявляя закономерности в юридических массивах данных, что не то же самое, что разработка юридического обоснования.

Учитывая более широкий контекст социальной роли судов, юриспруденция является развивающейся системой, открытой для новых социальных и политических вопросов. Поэтому инструменты ИИ, зависящие от пути, могут помешать этому эволюционному процессу: дедуктивный и зависящий от пути характер некоторых решений ИИ-мл (Machine Learning) может подорвать важную роль людей, принимающих решения, в эволюции права на практике и юридической аргументации.

Более того, на индивидуальном уровне зависимость от пути может также повлечь за собой риск «детерминистского анализа»,¹⁶⁹ что приведет к возрождению детерминистских доктрин в ущерб доктринам индивидуализации наказания и в ущерб принципу реабилитации и индивидуализации при вынесении приговора.

Кроме того, в ряде случаев, включая ADR, как для согласования требований сторон, так и для анализа психологической составляющей действий

167 См. также Clay, 2019. В связи с этим, например, правовая система, предоставляющая компенсацию за физические травмы на основе эффективного родового ущерба, может быть автоматизирована, но она не сможет пересмотреть основу правовой аргументации и распространить компенсацию на неличный и экзистенциальный ущерб.

168 См. Re и Solow-Niederman, 2019.

169 См. Европейская комиссия по эффективности правосудия (CEPEJ). 2018. Европейская этическая хартия по использованию искусственного интеллекта в судебных системах и их окружении, 9.

человека (вина, преднамеренность) требуется эмоциональный интеллект, которым системы ИИ не обладают.

Эти проблемы отражены в существующей правовой базе, представленной международными правовыми документами. Всеобщая декларация прав человека (статьи 7 и 10), МПГПП (статья 14), Конвенция о защите прав человека и основных свобод (статья 6), а также Хартия основных прав Европейского Союза (статья 47) подчеркивают следующие ключевые требования к осуществлению судебной власти: равенство перед законом, беспристрастность, независимость и компетентность. Инструменты ИИ не обладают этими качествами, что ограничивает их вклад в процесс принятия решений судами.

Как заявила Европейская комиссия по эффективности правосудия, «нейтральность алгоритмов — это миф, поскольку их создатели сознательно или непреднамеренно переносят в них свои собственные системы ценностей». Многочисленные случаи предвзятого отношения к применению ИИ подтверждают, что эти системы слишком часто — хотя во многих случаях и непреднамеренно — дают частичное представление об обществе и отдельных случаях, что несовместимо с принципами равного обращения перед законом и недискриминации.¹⁷⁰ Качество данных и другие формы оценки качества (оценка влияния, аудиты и т.д.) могут снизить этот риск¹⁷¹ но, учитывая степень потенциально затронутых интересов в случае необъективных решений, риски остаются высокими в случае справедливого правосудия и представляются непропорциональными по сравнению с преимуществами, в основном, с точки зрения эффективности для системы правосудия.¹⁷²

Дальнейшие опасения затрагивают принципы справедливого судебного разбирательства и равенства сторон,¹⁷³ когда судебные решения основываются на результатах запатентованных алгоритмов, учебные данные и структура

170 См. также СЕРЕJ, 2018.

171 См. также СЕРЕJ, 2018.

172 См. также Рекомендацию CM/Rec(2020)1, Приложение, п. 11.

173 См. также СЕРЕJ, 2018, Приложение I, п. 138.

которых не являются общедоступными.¹⁷⁴ Широкое понятие прозрачности может решить эти вопросы в отношении использования ИИ в судебных решениях, но прозрачность ИИ – сама по себе сложная цель – не может решить другие структурные и функциональные возражения, приведенные выше.

Кроме того, специалисты по исследованию данных могут по-разному формировать инструменты ИИ на этапах разработки и обучения, поэтому, если бы инструменты ИИ стали обязательной частью процесса принятия решений, правительства, выбирающие инструменты для использования судами, могли бы косвенно вмешиваться в независимость судей.

Этот риск не устраняется тем фактом, что судья остается свободным игнорировать решения ИИ при наличии конкретной мотивации. Хотя человеческий надзор является важным элементом,¹⁷⁵ его эффективное воздействие может быть подорвано психологической или утилитарной (экономически эффективной) склонностью человека, принимающего решения, воспользоваться решением, предоставляемым ИИ.¹⁷⁶

Предупреждение преступности

Сложность выявления и предотвращения преступлений стимулировала исследования в области применения ИИ для облегчения человеческой деятельности. В последние годы было разработано несколько решений¹⁷⁷ и все больше литературы в области предиктивной полиции, которая представляет собой проактивный подход к предотвращению преступлений, основанный на данных. По сути, имеющиеся решения преследуют две разные цели:

174 См. также СЕРЕJ, Приложение I, п. 131 («отсутствие прозрачности в процессах работы алгоритмов, разработанных частными компаниями (которые претендуют на интеллектуальную собственность), стало еще одной причиной для беспокойства. Если принять во внимание тот факт, что они берут исходные данные у самих государственных органов, отсутствие подотчетности перед гражданами создает серьезную демократическую проблему [...] примером тому может служить случай, когда ProPublica раскрыла недостатки алгоритма COMPAS после отказа компании-владельца поделиться ими»).

175 См. также СЕРЕJ, 2018.

176 См. также Mantelero, 2019 («якобы надежный характер решений, основанных на математике ИИ, может побудить тех, кто принимает решения на основе алгоритмов, довериться картине людей и общества, которую предлагают аналитики»).

177 См. Заврник, 2019; Агентство Европейского союза по фундаментальным правам, 2018, 98-100; Особа и Вельзер, 2017.

предсказать, где и когда могут произойти преступления, или предсказать, кто может совершить преступление.¹⁷⁸

Эти две цели имеют различное потенциальное воздействие на права и свободы человека, которое более выражено, когда ИИ используется для индивидуальных прогнозов. Однако в обоих случаях мы можем повторить здесь соображения об общих проблемах, связанных с ИИ (неизвестность, права интеллектуальной собственности, масштабный сбор данных¹⁷⁹, и т.д.) которые обсуждались в предыдущих разделах и частично решались прозрачностью, качеством данных, защитой данных, аудитом и другими мерами. Стоит отметить, что роль прозрачности¹⁸⁰ в судебном контексте может быть ограничена, чтобы не ослаблять сдерживающий эффект этих инструментов. Поэтому полная прозрачность может быть заменена аудитом и надзором со стороны независимых органов.

Если оставить в стороне организационные аспекты, касающиеся ограничения самостоятельности сотрудников полиции при выполнении своих обязанностей, то основные вопросы, связанные с использованием ИИ для прогнозирования преступности на географической и временной основе, касаются влияния этих инструментов на право на недискриминацию.¹⁸¹ Самореализующиеся предубеждения, предубеждения сообщества¹⁸² и исторические предубеждения¹⁸³ могут порождать формы стигматизации определенных групп и районов, где они обычно проживают.

Там, где анализ данных используется для классификации преступлений и вывода доказательств о преступных сетях, запатентованные решения

178 Таксономию методов прогнозирования см. в Perry et al., 2013, где выделены следующие четыре категории: методы прогнозирования преступлений (ориентированные на места и время совершения преступлений), методы прогнозирования преступников (ориентированные на отдельных лиц), методы прогнозирования личности преступников (ориентированные на отдельных лиц) и методы прогнозирования жертв преступлений (ориентированные на группы и, в некоторых случаях, на отдельных лиц).

179 См. также Рекомендацию Rec(2001)10 о Европейском кодексе полицейской этики, Приложение, п. 42.

180 См. также Barrett, 2017, 361-62.

181 См. Агентство Европейского союза по основным правам, 2018, 10.

182 См. также Barrett, 2017, 358-59 («Для некоторых цель коллективной безопасности заслуживает односторонней жертвы некоторой степени индивидуальных прав в данном конкретном контексте. Но этот расчет должен измениться, если жертва не является коллективной, а ограничивается группами меньшинств или становится принципиально произвольной в силу неприемлемой степени погрешности.»).

183 См. Беннет Мозес и Чан, 2018.

поднимают вопросы соблюдения принципов справедливого судебного разбирательства и равенства сторон в отношении сбора и использования доказательств. Более того, если повседневная деятельность политических департаментов направляется прогностическим программным обеспечением, это поднимает проблему подотчетности принятых стратегий, поскольку они частично определяются программным обеспечением и, следовательно, компаниями-разработчиками программного обеспечения, а не полицией.

Более острый конфликт с правами человека возникает в области инструментов предиктивного полицейского контроля, которые используют профилирование для поддержки индивидуального прогнозирования. Помимо вопроса обработки данных и профилирования,¹⁸⁴ эти решения могут также негативно повлиять на принцип презумпции невиновности,¹⁸⁵ процессуальную справедливость и право на недискриминацию.¹⁸⁶

В то время как вопросы недискриминации могут быть частично решены, оставшиеся конфликты представляются более трудноразрешимыми. С точки зрения прав человека и соразмерности (включая право на уважение частной и семейной жизни¹⁸⁷), риск нанесения ущерба этим принципам представляется высоким и неадекватно уравновешенным доказательствами преимуществ для индивидуальных и коллективных прав и свобод.¹⁸⁸ В свете будущего регулирования ИИ это должно побудить к тщательному рассмотрению этих вопросов, принимая во внимание различие между техническими возможностями.

Наконец, с более широкой и всеобъемлющей точки зрения прав человека, сосредоточение внимания на преступности с помощью инструментов ИИ, основанных на данных, приводит к краткосрочному фактическому подходу, который недооценивает социальные проблемы, которые часто связаны с

184 См. выше раздел II.2.

185 См. также Рекомендацию Rec(2001)10 о Европейском кодексе полицейской этики, Приложение, п. 47.

186 См. также Рекомендацию Rec(2001)10 о Европейском кодексе полицейской этики, Приложение, п. 49.

187 См. van Brakel and De Hert, 2011, 183. См. также Сабо и Висси против Венгрии [2016] Европейский суд по правам человека Четвертая секция. Жалоба № 37138/14.

188 См. Meijer and Wessels, 2019.

преступностью и требуют долгосрочных социальных стратегий, включающих эффективное укрепление индивидуальных и социальных прав и свобод.¹⁸⁹

IV. Гармонизация выявленных принципов

В предыдущих разделах было определено несколько руководящих принципов для будущего регулирования ИИ. Эти принципы были контекстуализированы с учетом проблем, связанных с ИИ в различных областях.

ИИ в различных рассмотренных областях, однако стоит обратить внимание на существующий уровень гармонизации между этими принципами.

Результаты данного исследования показывают, что в ограниченном числе случаев существуют общие принципы (примат человека, индивидуальное самоопределение, отсутствие дискриминации, надзор за людьми). Это объясняется несколькими факторами.

Во-первых, некоторые принципы являются отраслевыми. Это касается, например, независимости судей или принципов справедливого судебного разбирательства и равенства сторон, которые касаются только правосудия.¹⁹⁰

Во-вторых, некоторые руководящие принципы одинаковы в разных областях, но с различными нюансами в каждом контексте. Это верно для прозрачности, которая часто рассматривается как ключевая в регулировании ИИ, но приобретает разные значения в разных нормативных контекстах.

В области здравоохранения и персональных данных прозрачность относится к информации, предоставляемой людям о лечении, касающемся их, с особым вниманием к процессу и связанным с ним рискам и с сильным оттенком индивидуального самоопределения. Однако прозрачность также актуальна в сфере защиты данных для контроля за осуществлением власти над данными в руках государственных и частных организаций. Этот иной облик

¹⁸⁹ См. также Розенбаум, 2006, 245-266.

¹⁹⁰ См. также принципы равного доступа и благодеяния в секторе здравоохранения, или принципы невмешательства государственных органов в деятельность СМИ по влиянию на выборы и обязательство предоставлять равные условия всем политическим партиям и кандидатам в предвыборной рекламе.

прозрачности затем рассматривается в связи с применением ИИ для демократического участия и надлежащего управления. В то же время в контексте правосудия прозрачность имеет более сложное значение, будучи жизненно важной для защиты основных прав и свобод (например, использование ИИ в судах), но также требуя ограничений, чтобы не нанести ущерб конкурирующим интересам (например, выявление и предотвращение преступлений при предиктивной работе полиции).

Таким образом, мы можем сделать вывод, что прозрачность является руководящей ценностью, но мы должны выйти за рамки простого утверждения о прозрачности как ключевом принципе регулирования ИИ. Как и в случае с другими ключевыми принципами (такими как участие, инклюзивность, демократический надзор и открытость), необходима надлежащая контекстуализация, принятие положений, учитывающих различные контексты, в которых они действуют.

В-третьих, некоторые принципы отличаются друг от друга, но относятся к одной и той же концептуальной области, принимая различные нюансы в разных контекстах. Так обстоит дело с подотчетностью и руководящими принципами по управлению рисками в целом. Здесь уровень детализации и соответствующие требования могут быть более или менее проработанными. Например, в области защиты данных существует несколько положений, реализующих эти принципы со значительной степенью детализации, в то время как в области демократии и правосудия эти принципы менее проработаны в отношении приложений с интенсивным использованием данных, таких как искусственный интеллект.

Наконец, существуют определенные компоненты стратегии регулирования ИИ, которые являются не принципами, а операционными подходами и решениями, общими для различных областей, хотя и требующими разработки в зависимости от контекста. Это относится к важной роли, которую играют образование и обучение, операционная совместимость и экспертные комитеты.

Такие соображения позволяют предположить, что достижима лишь частичная гармонизация. Поэтому нормативный подход к ИИ должен основываться на юридически обязательном документе, включающем как общие положения – с упором на общие принципы и оперативные решения – так и более конкретные и отраслевые положения, охватывающие те принципы, которые актуальны только в данной области, или случаи, когда один и тот же принцип по-разному контекстуализируется в различных областях.

V. Выводы

Данный анализ подтвердил обоснованность принятого методологического подхода, который сосредоточен на контекстуализации руководящих принципов, извлеченных из юридически обязательных и необязательных намеренных инструментов. В то же время, он также показал сложность систематизации положений широкого спектра инструментов, которые отличаются не только своим обязательным характером, но и специфической направленностью и подходом, а также структурой.

Результаты также подтвердили, что существующие рамки, основанные на правах человека, демократии и верховенстве закона, могут обеспечить надлежащий и общий контекст для разработки более конкретного обязательного документа для регулирования ИИ в соответствии с принципами и ценностями, закрепленными в международных правовых документах, способного более эффективно решать вопросы, поднимаемые ИИ.

Эти международные рамки неизбежно заставляют нас подтвердить центральную роль человеческого достоинства в контексте ИИ, где нельзя допустить, чтобы машинные решения приводили к дегуманизации людей. Это также может предполагать введение конкретных ограничений на ИИ, когда он разрабатывается или используется таким образом, который не соответствует уважению человеческого достоинства,¹⁹¹ правам человека, демократии и верховенству закона.

191 См. также ЮНЕСКО. 1997. Декларация о геноме человека и правах человека, Статья 11.

С точки зрения будущего регулирования ИИ, этот положительный методологический и содержательный результат не исключает существования некоторых пробелов. В основном они касаются широких областей, таких как демократия и правосудие, где возможны различные варианты и интерпретации в зависимости от политического и общественного видения будущих отношений между людьми и машинами.

Дальнейшие исследования в области прав человека и ИИ, а также продолжающиеся дебаты на международном и региональном уровне будут способствовать устранению этих пробелов. Однако, учитывая развивающуюся природу ИИ, желателен подход, основанный на совместном регулировании.

Поэтому обязательный документ, устанавливающий правовые рамки для ИИ, включая как общие принципы, так и конкретные положения, касающиеся конкретных вопросов, может сочетаться с подробными правилами, изложенными в дополнительных необязательных отраслевых документах. Такая модель обеспечит как четкую нормативную базу, так и гибкость, необходимую для решения задач технологического развития.

Ссылки

Андорно, Р. 2005. Конвенция Овьедо: Европейская правовая база на стыке прав человека и здравоохранения. *Journal of International Biotechnology Law*, January 2005. <https://doi.org/10.1515/jibl.2005.2.4.133>, accessed 20.02.2020.

Azencott, C.-A. 2018. Машинное обучение и геномика: Точная медицина против конфиденциальности пациентов. *Phil. Trans. R. Soc. A* 376, no. 2128 (13 September 2018): 20170350. <https://doi.org/10.1098/rsta.2017.0350>, accessed 14.01.2020.

Barrett, L. 2017. Reasoably Suspicious Algorithms: Predictive Policing at the United States Border. 41 *N.Y.U. Rev. Law & Social Change* 327.

Беннет Мозес, Л., Чан, Дж. 2018. Алгоритмическое прогнозирование в полицейской деятельности: Assumptions, Evaluation, and Accountability. 28 *Policing and Society* 806.

- Бычавска-Синярска, Д. 2017. Защита права на свободу выражения мнения в соответствии с Европейской конвенцией по правам человека. Совет Европы.
- Кабица, Ф., Расоини, Р., и Дженсини, Г.Ф. 2017. Непреднамеренные последствия машинного обучения в медицине. JAMA 318, № 6 (8 августа 2017): 517. <https://doi.org/10.1001/jama.2017.7797>, accessed 18.12.2019.
- Calo, R., Citron, D.K. 2020. Автоматизированное административное государство: A Crisis of Legitimacy. Emory Law Journal, Forthcoming. <https://ssrn.com/abstract=3553590>, accessed 20.04.2020.
- Каруана, Р. и др. 2015. Интеллектуальные модели для здравоохранения: прогнозирование риска пневмонии и 30-дневной госпитализации. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Чам, Швейцария: Springer International Publishing AG.
- Клей, Т. (ред.). 2019. L'arbitrage en ligne. Rapport du Club des Juristes. Paris. 58 <https://www.leclubdesjuristes.com/les-commissions/larbitrage-en-ligne/>, accessed 30.05.2020.
- Комитет министров. 1999. Рекомендация № R (99) 5 о защите частной жизни в Интернете. Принята Комитетом министров 23 февраля 1999 года на 660-м заседании заместителей министров.
- Комитет министров. 2001. Рекомендация Rec(2001)10 Комитета министров государствам-членам о Европейском кодексе полицейской этики. Принята Комитетом министров 19 сентября 2001 года на 765-м заседании заместителей министров.
- Комитет министров. 2003. Рекомендация CM/Rec(2003)4 об общих правилах борьбы с коррупцией при финансировании политических партий и избирательных кампаний. Принята Комитетом министров 8 апреля 2003 года на 835-м заседании заместителей министров.
- Комитет министров. 2004. Рекомендация CM/Rec(2004)15 об электронном управлении («E-Governance»). Принята Комитетом министров 15 декабря 2004 года на 909-м заседании заместителей министров.

Комитет министров. 2007. Рекомендация CM/Rec(2007)15 Комитета министров государствам-членам о мерах, касающихся освещения избирательных кампаний в средствах массовой информации. Принята Комитетом министров 7 ноября 2007 года на 1010-м заседании заместителей министров.

Комитет министров. 2009. Рекомендация CM/Rec(2009)1 Комитета министров государствам-членам об электронной демократии (e-democracy). Принята Комитетом министров 18 февраля 2009 года на 1049-м заседании заместителей министров.

Комитет министров. 2009. Рекомендация CM/Rec(2009)2 Комитета министров государствам-членам по оценке, аудиту и мониторингу политики участия и участия на местном и региональном уровне. Принята Комитетом министров 11 марта 2009 года на 1050-м заседании заместителей министров.

Комитет министров. 2010. Рекомендация CM/Rec(2010)13 о защите физических лиц в отношении автоматической обработки персональных данных в контексте профилирования. Принята Комитетом министров 23 ноября 2010 года на 1099-м заседании заместителей министров.

Комитет министров. 2011. Рекомендация CM/Rec(2011)7 Комитета министров государствам-членам о новом понятии СМИ. Принята Комитетом министров 21 сентября 2011 года на 1121-м заседании заместителей министров.

Комитет министров. 2012. Рекомендация CM/Rec(2012)3 о защите прав человека в отношении поисковых систем. Принята Комитетом министров 4 апреля 2012 года на 1139-м заседании заместителей министров.

Комитет министров. 2012. Рекомендация CM/Rec(2012)4 о защите прав человека в отношении служб социальных сетей. Принята Комитетом министров 4 апреля 2012 года на 1139-м заседании заместителей министров.

Комитет министров. 2014. Рекомендация CM/Rec(2014)7 о защите лиц, сообщающих о нарушениях. Принята Комитетом министров 30 апреля 2014 года на 1198-м заседании заместителей министров.

Комитет министров. 2016. Рекомендация CM/Rec(2016)1 о защите и поощрении права на свободу выражения мнений и права на частную жизнь в отношении

сетевого нейтралитета. Принята Комитетом министров 13 января 2016 года на 1244-м заседании заместителей министров.

Комитет министров. 2016. Рекомендация CM/Rec(2016)4 Комитета министров государствам-членам о защите журналистики и безопасности журналистов и других работников СМИ. Принята Комитетом министров 13 апреля 2016 года на 1253-м заседании заместителей министров.

Комитет министров. 2016. Рекомендация CM/Rec(2016)5 Комитета министров государствам-членам о свободе Интернета. Принята Комитетом министров 13 апреля 2016 года на 1253-м заседании заместителей министров.

Комитет министров. 2017. Рекомендация CM/Rec(2017)5 Комитета министров государствам-членам о стандартах для электронного голосования. Принята Комитетом министров 14 июня 2017 года на 1289-м заседании заместителей министров.

Комитет министров. 2018. Рекомендация CM/Rec(2018)1 Комитета министров государствам-членам о плюрализме СМИ и прозрачности владения СМИ. Принята Комитетом министров 7 марта 2018 года на 1309-м заседании заместителей министров.

Комитет министров. 2018. Рекомендация CM/Rec(2018)2 Комитета министров государствам-членам о роли и ответственности интернет-посредников. Принята Комитетом министров 7 марта 2018 года на 1309-м заседании заместителей министров.

Комитет министров. 2018. Рекомендация CM/Rec(2018)4 Комитета министров государствам-членам об участии граждан в местной общественной жизни. Принята Комитетом министров 21 марта 2018 года на 1311-м заседании заместителей министров.

Комитет министров. 2019. Рекомендация CM/Rec(2019)2 Комитета министров государствам-членам о защите данных, связанных со здоровьем. Принята Комитетом министров 27 марта 2019 года на 1342-м заседании заместителей министров.

Комитет министров. 2019. Рекомендация CM/Rec(2019)3 Комитета министров государствам-членам о надзоре за деятельностью местных органов власти. Принята Комитетом министров 4 апреля 2019 года на 1343-м заседании заместителей министров.

Комитет министров. 2020. Рекомендация CM/Rec(2020)1 Комитета министров государствам-членам о воздействии алгоритмических систем на права человека. Принята Комитетом министров 8 апреля 2020 года на 1373-м заседании заместителей министров.

Совет адвокатов и юридических обществ Европы. 2020. Соображения ССВЕ по правовым аспектам искусственного интеллекта. Брюссель.

Совет Европы – Венецианская комиссия, ОБСЕ/БДИПЧ. 2011. Совместные руководящие принципы регулирования политических партий. <https://www.osce.org/odihr/77812>, accessed 20.12.2019.

Совет Европы, Консультативный комитет Конвенции о защите физических лиц при автоматизированной обработке персональных данных. 2019. Руководящие принципы по искусственному интеллекту и защите данных. T-PD(2019)01. <https://rm.coe.int/2018-lignes-directrices-sur-l-intelligence-artificielle-et-la-protecti/168098e1b7>, accessed 16.11.2019.

Совет Европы, Консультативный комитет Конвенции о защите физических лиц при автоматизированной обработке персональных данных. 2017. Руководящие принципы по защите физических лиц в отношении обработки персональных данных в мире Больших Данных. T-PD(2017)1. <http://rm.coe.int/t-pd-2017-1-bigdataguidelines-en/16806f06d0>, accessed 16.11.2019.

Совет Европы, Консультативный комитет Конвенции о защите физических лиц при автоматизированной обработке персональных данных. 2019. Профилирование и Конвенция 108+: предложения по обновлению. T-PD(2019)07BISrev.

Совет Европы, Генеральный директорат по демократии – Европейский комитет по демократии и управлению. 2016. Сборник наиболее актуальных текстов Совета Европы в области демократии.

<http://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806b5f2c>, accessed 18.11.2019.

Совет Европы, Генеральный директорат по демократии и политическим вопросам – Директорат демократических институтов. 2011. Руководящие принципы прозрачности выборов с использованием электронных средств.

Совет Европы, Генеральный директорат по демократии и политическим вопросам – Директорат демократических институтов. 2009. Проект «Надлежащее управление в информационном обществе», СМ(2009)9.

Совет Европы, Парламентская ассамблея. 2019. Резолюция 2254 (2019)1. Свобода СМИ как условие демократических выборов.

Совет Европы. 2008. The 12 Principles of Good Governance <https://www.coe.int/en/web/good-governance/12-principles> (%2225565951%22:[0]), accessed 16.03.2020.

Совет Европы. 2009. Дополнительный протокол к Европейской хартии местного самоуправления о праве на участие в делах местного органа власти, Утрехт, 16.

Совет Европы. 2017. Руководящие принципы гражданского участия в принятии политических решений, СМ(2017)83-final. Принято Комитетом министров 27 сентября 2017 года на 1295-м заседании заместителей министров.

Совет Европы. 2019. Декларация Комитета министров о манипулятивных возможностях алгоритмических процессов. Принята Комитетом министров 13 февраля 2019 года на 1337-м заседании заместителей министров.

Совет Европы. 2019. Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes (Принята Комитетом министров 13 февраля 2019 года на 1337-м заседании заместителей министров).

Совет Европы. Руководящие принципы гражданского участия в принятии политических решений. СМ(2017)83-final. <https://www.coe.int/en/web/youth/-/guidelines-for-civil-participation-in-political-decision-making>, accessed 15.03.2020.

Совет Европы-Комитет экспертов по интернет-посредникам(MSI-NET). 2018. Алгоритмы и права человека. Исследование по аспектам прав человека

автоматизированных методов обработки данных (в частности, алгоритмов) и возможным регулятивным последствиям. <https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5>, доступ 28.11.2019.

Crawford, K., and Joler, V. 2018. Anatomy of an AI System: Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources. AI Now Institute and Share Lab. <http://www.anatomyof.ai>, accessed 27.12.2019.

Cummings, M. L., Roff H. M., Cukier K., Parakilas J. and Bryce H. 2018. Chatham House Report. Artificial Intelligence and International Affairs Disruption Anticipated. Лондон: Chatham House. The Royal Institute of International Affairs. <https://www.chathamhouse.org/artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf>, accessed 21.03.2020.

Дюпон, Б. и др. 2018. Искусственный интеллект в контексте преступности и уголовного правосудия. Корейский институт криминологии 2018. <https://www.cyberjustice.ca/publications/lintelligence-artificielle-dans-le-contexte-de-Ja-criminalite-et-de-la-justice-penale/>, accessed 30.05.2020.

Кодекс практики ЕС по дезинформации, 2018 <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation>, 23.03.2020.

Европейская комиссия за демократию через право (Венецианская комиссия). 2002. Кодекс надлежащей практики в избирательных вопросах. Руководящие принципы и пояснительный отчет. Принято Венецианской комиссией на 51-й и 52-й сессиях (Венеция, 5-6 июля и 18-19 октября 2002 года).

Европейская комиссия за демократию через право (Венецианская комиссия). 2019. Совместный доклад Венецианской комиссии и Директората информационного общества и действий против преступности Генерального директората по правам человека и верховенству права (DGI) о цифровых технологиях и выборах.

Европейская комиссия по эффективности правосудия (CEPEJ). 2018. Европейская этическая хартия по использованию искусственного интеллекта в судебных системах и их окружении. Принята CEPEJ в ходе 31-го пленарного

заседания (Страсбург, 3-4 декабря 2018 года) <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c>, accessed 04.12.2018.

Европейская комиссия, Сети, контент и технологии – Генеральный директорат по коммуникациям. 2018. A Multi-Dimensional Approach to Disinformation Report of the Independent High Level Group on Fake News and Online Disinformation. <https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation>, accessed 22.03.2018.

Европейская комиссия. 2014. Green paper on mobile health. <http://ec.europa.eu/newsroom/dae/document.cfm?id=5147>, accessed 12.01.2020.

Европейская комиссия. 2020. A European strategy for data, COM(2020) 66 final. <https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy-en>, accessed 20.02.2020.

Европейская комиссия. 2020. Отчет о последствиях искусственного интеллекта, Интернета вещей и робототехники для безопасности и ответственности, COM(2020) 64final. <https://ec.europa.eu/info/publications/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics-0-en>, accessed 20.02.2020.

Европейская комиссия. 2020. White Paper on Artificial Intelligence – A European approach to excellence and trust, COM(2020) 65 final. <https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust-en>, accessed 20.02.2020.

Агентство Европейского союза по основным правам. 2018. #BigData: Discrimination in Data-Supported Decision Making. <https://fra.europa.eu/en/publication/2018/bigdata-discrimination-data-supported-decision-making>, accessed 20.05.2020.

Агентство Европейского союза по фундаментальным правам. 2018. Предотвращение незаконного профилирования сегодня и в будущем: A Guide. <https://fra.europa.eu/en/publication/2018/preventing-unlawful-profiling-today-and-future-guide>, accessed 20.05.2020.

Фей Якобсен, А. 2013. Право на участие в общественной жизни. Обновление законодательства о правах человека. Выпускной документ. Датский институт по правам человека.

Ферриман, К. и Питкан, М. 2018. Справедливость в точной медицине. *Data & Society*, февраль, <https://datasociety.net/wp-content/uploads/2018/02/Data.Society.Fairness.In.Precision.Medicine.Feb2018.FINAL-2.26.18.pdf>, accessed 20.12.2019.

Gonzalez Fuster, G. 2014. Появление защиты персональных данных в качестве основного права ЕС. Cham-New York: Springer International Publishing.

Мейсли, Н. 2017. Международное право прав? Статья 25(a) МПГПП как право человека на участие в международном правотворчестве. 28 *European Journal of International Law* 89.

Манхейм, К., Каплан, Л. 2019. Искусственный интеллект: Риски для частной жизни и демократии. 21 *Yale J.L. & Tech.* 106.

Mantelero, A. 2018. AI and Big Data: План оценки воздействия на права человека, социальные и этические аспекты. 34 *Computer Law & Security Review* 754.

Mantelero, A. 2019. Искусственный интеллект и защита данных: Вызовы и возможные средства защиты. Доклад об искусственном интеллекте. Консультативный комитет Конвенции о защите физических лиц при автоматизированной обработке персональных данных: Страсбург. T-PD(2018)09Rev, <https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6>, accessed 20.02.2020.

Майер-Шонбергер, В. 1997. Поколение развития защиты данных в Европе? In Agre, P.E., Rotenberg, M. (eds). *Technology and privacy: Новый ландшафт*. Cambridge, MA: MIT Press.

Мехр, Х. 2017. Искусственный интеллект для обслуживания граждан и правительства. Гарвардская школа Кеннеди. Ash Center for Democracy and Innovation.

- Мейер, А., Весселс, М. 2019. Predictive Policing: Обзор преимуществ и недостатков. 42 *International Journal of Public Administration* 1031.
- Нуньес, К. 2017. Искусственный интеллект и юридическая этика: Могут ли юристы с искусственным интеллектом принимать этические решения. 20 *Tul. J. Tech. & Intell. Prop.* 189-204.
- OECD. 2013. Рекомендация Совета относительно руководящих принципов, регулирующих защиту частной жизни и трансграничные потоки персональных данных, C(80)58/FINAL, с поправками от 11 июля 2013 года C(2013)79.
- ОБСЕ. 2020. Неофициальный документ о влиянии искусственного интеллекта на свободу выражения мнений. <https://www.osce.org/representative-on-freedom-of-media/447829>, accessed 11.06.2020.
- Osoba, O.A., Welser, W. 2017. Интеллект по нашему образу и подобию: The Risks of Bias and Errors in Artificial Intelligence. RAND Corporation. https://www.rand.org/pubs/research_reports/RR1744.html, accessed 20.05.2020.
- Парламентская ассамблея. 2019. Резолюция 2254 (2019)1. Свобода СМИ как условие демократических выборов.
- Парламентская ассамблея. 2019. Резолюция 2300 (2019)1, Улучшение защиты информаторов по всей Европе.
- Перри, У.Л. и др. 2013. Predictive Policing: Роль прогнозирования преступности в операциях правоохранительных органов. RAND Corporation 2013. https://www.rand.org/pubs/research_reports/RR233.html, accessed 30.03.2020.
- Privacy International. 2017. Smart Cities: Utopian Vision, Dystopian Reality. <https://privacyinternational.org/report/638/smart-cities-utopian-vision-dystopian-reality>, accessed 21.03.2020.
- Re, R.M., Solow-Niederman, A. 2019. Developing Artificially Intelligent Justice. 22 *Stan. Tech. L. Rev.* 242.
- Розенбаум, Д., 2006. The limits of hot spots policing. In: D. Weisburd and A. Braga, eds. *Инновации в полиции: контрастирующие перспективы*. Нью-Йорк, штат Нью-Йорк: Cambridge University Press, 245-266.

Рувруа, А. 2016. «О данных и людях» – Основные права и свободы в мире больших данных. Консультативный комитет Конвенции о защите частных лиц при автоматизированной обработке персональных данных: Страсбург. T-PD-BUR(2015)09Rev,

<http://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806a6020>, accessed 04.11.2019.

Savaget, P., Chiarini, T., Evans, S. 2019. Empowering Political Participation through Artificial Intelligence. 46 Science and Public Policy 369.

Seatzu, F. 2015. Опыт Европейского суда по правам человека с Европейской конвенцией о правах человека и биомедицине. 31(81) Utrecht Journal of International and European Law 5, DOI: <http://dx.doi.org/10.5334/ujiel.da>, accessed 07.12.2019.

Sunstein, C.R. 2015a. The Ethics of Nudging. Йельский журнал по регулированию 32: 413

Sunstein, C.R. 2015b. Why Nudge? The Politics of Libertarian Paternalism. Нью-Хейвен: Издательство Йельского университета.

Санстейн, К.Р., Талер, Р. 2003. Либертарианский патернализм не оксюморон. Юридическое обозрение Чикагского университета 70 (4): 1159.

Талер, Р., Санстейн, К.Р. 2008. Nudge. Нью-Хейвен, КТ: Издательство Йельского университета.

Специальный докладчик Организации Объединенных Наций (ООН) по вопросам свободы мнений и их выражения, Представитель Организации по безопасности и сотрудничеству в Европе (ОБСЕ) по вопросам свободы СМИ, Специальный докладчик Организации американских государств (ОАГ) по вопросам свободы выражения мнений и Специальный докладчик Африканской комиссии по правам человека и народов (АКПЧН) по вопросам свободы выражения мнений и доступа к информации. 2017. Совместная декларация о «фальшивых новостях», дезинформации и пропаганде. <http://www.osce.org/fom/302796?download=true>, accessed 02.02.2020.

T-PD(2019)07BISrev. Консультативный комитет Конвенции о защите физических лиц при автоматизированной обработке персональных данных. 2019. Профилирование и Конвенция 108+: предложения по обновлению Управление Верховного комиссара ООН по правам человека. 1981. Замечание общего порядка № 1 КЭСКП: Представление отчетности государствами-участниками. Принято на тринадцатой сессии Комитета по экономическим, социальным и культурным правам 27 июля 1981 года (Содержится в документе E/1989/22).

Управление Верховного комиссара ООН по правам человека. 1996. Замечание общего порядка № 25: Право на участие в ведении государственных дел, право голоса и право равного доступа к государственной службе (ст. 25). CCPR/C/21/Rev.1/Add.7.

ЮНЕСКО. 2019. Предварительное исследование возможного нормативного документа по этике искусственного интеллекта. <https://unesdoc.unesco.org/ark:/48223/pf0000369455>, accessed 21.11.2019.

ЮНЕСКО. Декларация о геноме человека и правах человека (11 ноября 1997 года). Конвенция ООН против коррупции, 2003.

CM/Rec(2018)2 о роли и ответственности интернет-посредников CM/Rec(2020)х о воздействии алгоритмических систем на права человека Decl(13/02/2019) о манипулятивных возможностях алгоритмических процессов CM/Rec(2018)1 о плюрализме СМИ и прозрачности владения СМИ CM/Rec(2020)х о содействии созданию благоприятных условий для качественной журналистики в цифровую эпоху Выбора

Всеобщая декларация прав человека

Международный пакт о гражданских и политических правах

Конвенция Организации Объединенных Наций о ликвидации всех форм расовой дискриминации

Конвенция Организации Объединенных Наций о ликвидации всех форм дискриминации в отношении женщин

Конвенция Организации Объединенных Наций о правах инвалидов

Конвенция Организации Объединенных Наций против коррупции

Конвенция о защите прав человека и основных свобод (ETS № 5)

Протокол к Конвенции о защите прав человека и основных свобод (ETS № 9)

Европейская хартия местного самоуправления (ETS № 122)

Европейская хартия региональных языков или языков меньшинств (ETS № 148)

Конвенция о киберпреступности (ETS № 185)

Конвенция о защите физических лиц при автоматизированной обработке персональных данных (ETS № 108)

Кодекс надлежащей практики в избирательных вопросах, принятый Советом по демократическим выборам Совета Европы и Европейской комиссией за демократию через право (Венецианская комиссия)

Рекомендация Rec(2003)3

Комитета министров государствам-членам о сбалансированном участии женщин и мужчин в принятии политических и общественных решений

Конвенция о стандартах демократических выборов, избирательных прав и свобод в государствах-участниках Содружества Независимых Государств (CDL-EL(2006)031rev)

Рекомендация Rec(99)5 Комитета министров государствам-членам о защите частной жизни в Интернете

Рекомендация Rec(2004)15 Комитета министров для государств-членов по электронному управлению (e-governance)

Рекомендация CM/Rec(2007)15 Комитета министров государствам-членам о мерах по освещению избирательных кампаний в средствах массовой информации

Приложение 2. Затронутые территории

Воздействующие области:

Биомедицина.

Наблюдение, профилактика, диагностика и вмешательство в медицинских учреждениях на основе ИИ.

Наблюдение, диагностика, мониторинг, финансирование (страхование) лечения на основе прогнозирования (например, приложения и онлайн-сервисы, ориентированные на пользователя, вне медицинских учреждений).

Антидискриминация.

Автоматизированное принятие решений, охватывающее различные области, как в государственном, так и в частном секторе (например, прием на работу, социальное обеспечение/социальные пособия, доступ к товарам и услугам, таким как банковские кредиты, страхование).

Предиктивное полицейское управление (в котором высок риск расового профилирования) Предиктивное правосудие Распознавание лиц.

Технологии прогнозирования поведения, такие как распознавание эмоций и детекция лжи на основе ИИ.

Инструменты персональной помощи (например, Siri) Модерация контента Защита данных.

Киберпреступность и электронные доказательства.

Автоматизированные киберпреступления и кибератаки, такие как: Распределенные атаки типа «отказ в обслуживании» (DDOS) Атаки на критическую информационную инфраструктуру Атаки типа «человек посередине».

Фишинг и аналогичные методы социальной инженерии.

Сканирование на предмет уязвимостей.

и т.д.

Расследование киберпреступлений и компьютерная криминалистика:

Сбор и анализ электронных доказательств (в связи с любым преступлением).

Атрибуция.

Обратная разработка.

Кибербезопасность и предотвращение киберпреступлений: Обнаружение вредоносных программ, вторжений и т.д. Автоматизированное устранение уязвимостей.

Сектор правосудия.

Обработка судебных решений и данных:

для поддержки принятия судебных решений или судебных исследований).

Урегулирование споров в режиме онлайн.

Предоставление юридических консультаций участникам судебного процесса.

Предиктивное полицейское управление.

Конгресс местных и региональных властей.

Предоставление местных общественных услуг. Инструменты для поощрения участия граждан.

Широкое разнообразие цифровых и электронных приложений в городах и местных сообществах.

Применение информационно-коммуникационных технологий (ИКТ) для улучшения качества жизни и рабочей среды в городах.

Умное городское управление. Внедрение ИКТ в государственные системы.

Внедрение на местах практик, объединяющих людей и ИКТ для стимулирования инноваций и расширения знаний, которые они предлагают.

Свобода выражения мнений.

Индивидуальная коммуникация (посредством автоматизированной модерации и ограничения контента – алгоритмическая сортировка, классификация, оптимизация и рекомендательные системы).

Производство и распространение СМИ (робо-журналистика, журналистика данных, НЛП, микро-таргетинг читательской базы, автоматизированные новостные ленты на основе профиля читателя).

Социальная и политическая коммуникация/фрагментация/поляризация общественного дискурса, политическая редирекция (микро-таргетинг базы избирателей, подтасовка мнений с помощью ботов, распространение автоматизированных сайтов местных СМИ).

Выборы.

Предвыборный период:

Планирование календаря выборов.

Обучение участников избирательного процесса.

Делимитация избирательных округов.

Регистрация избирателей и кандидатов.

Аккредитация наблюдателей (международных и внутренних).

Обновление списка избирателей.

Обновление законодательной базы.

Финансирование политических партий.

Предвыборная пропаганда со стороны администрации и политических партий/кандидатов

Предвыборный период:

Финансирование избирательных кампаний.

Доступ к средствам массовой информации.

Голосование.

Подсчет бюллетеней Подсчет результатов.

Послевыборный период: Публикация результатов выборов Разрешение избирательных споров.

Демократия (за исключением вопросов, связанных с выборами и избирательным циклом).

Разделение властей. Участие гражданского общества. Участие граждан.

Частная жизнь. Гражданство.

Защита меньшинств Плюрализм и разнообразие Легитимность.

Эффективное управление.

Местные органы власти Региональная администрация Предоставление услуг

Бюджетные ассигнования.

Системы социального обеспечения и социальных пособий Полиция и судебная система Умные города.

Государственные тендеры и закупки Институциональный потенциал.

Гендерное равенство, включая насилие в отношении женщин (НОЖ).

Общая проблема унаследованной гендерной предвзятости от систем данных, на которых обучаются алгоритмы (актуально для многих областей), что может привести к усугублению гендерного и социального неравенства.

Общие вопросы, связанные с ИИ как сектором занятости:

недостаточное участие / недопредставленность женщин усугубляет потенциальные гендерные предубеждения и исключает их из мощного сектора.

Эксплуатация «кликеров работников» в Европе и по всему миру (низкие зарплаты, отсутствие социальной защиты, отсутствие трудовых прав, длительное воздействие вредного контента для модераторов контента и т.д.).

Конкретные проблемы Дискриминационный отбор при приеме на работу.

Автоматизированное принятие решений для государственных и частных служб.

Распознавание лиц и речи (хуже работает для женщин, особенно для некоторых групп) Наблюдение / преследование, облегченное инструментами ИИ, например, в контексте домашнего насилия.

Автоматизированное принятие решений усугубляет возможность множественной дискриминации по признаку пола/гендера, расы и социального происхождения путем объединения вторичных данных, таких как уровень образования, адрес, уровень дохода.

Предиктивное правосудие (ex VAW).

Прогнозирование здоровья на основе гендерно предвзятых данных (например, некоторые заболевания характеризуются как «женские» или «мужские»).

Унаследованные предубеждения при машинной модерации контента (высокая толерантность к сексизму, сексистским высказываниям ненависти и НОЖ).

Гендерные виртуальные помощники / роботы, увековечивающие гендерные стереотипы Гендерный маркетинг, увековечивающий гендерные стереотипы

Дифференцированное ценообразование в зависимости от пола/гендера.

Положительные последствия.

Использование устройств GPS слежения для обеспечения соблюдения охранных приказов.

Использование ИИ правоохранительными органами для проведения оценки риска в случаях ДВ.

Использование ИИ для выявления и отслеживания гендерной предвзятости и возможность карантина или устранения распространения (сексистского) языка ненависти на платформах.

Разработка приложений для поддержки и информирования жертв НОЖ.

Использование инструментов на основе ИИ для анализа контента и отслеживания гендерной предвзятости / анализа репрезентации (например, в кино или других СМИ).

Культура, творчество и наследие.

Доступ и участие в общественной / культурной жизни;

ФoЕ (в т.ч. свобода художественного выражения).

Доступ к беспристрастной информации.

Автоматизированное принятие решений, таргетинг, профилирование.

Автоматизированное принятие решений, таргетинг, профилирование.

Но также изучение языков, находящихся под угрозой исчезновения, с целью их сохранения/защиты Автоматизированная помощь в управлении, здравоохранении и т.д. для носителей групп меньшинств/языков.

Геолокализация, предиктивная полиция, криминальная аналитика (уничтожение, разграбление, незаконный оборот культурных ценностей; таргетинг; изучение наследия, находящегося под угрозой исчезновения, может помочь в его защите).

Автоматизированное создание контента, таргетинг, профилирование (для создания, обмена, потребления культурных ценностей).

Разработка и производство аудиовизуального контента: Прогнозируемый анализ аудитории Автоматизированный анализ сценария Помощник или автоматическое написание сценария Компьютерная генерация изображений (SFX, анимация...).

Автоматизированное определение местоположения, планирование и бюджетирование (влияние еще предстоит оценить) Распространение контента.

Алгоритмы рекомендаций Целевая реклама.

Автоматизированный контроль контента (соответствие нормативным требованиям) / цензура (см. исследование «Вступление в новую парадигму искусственного интеллекта и сериалов», проведенное по заказу DG2 и Eurimages).

Доступ и участие в общественной / культурной жизни; FoE (включая свободу художественного выражения) Доступ к беспристрастной информации.

GRECO.

Борьба с коррупцией.

Уголовная ответственность, связанная с использованием автоматизированных транспортных средств.

Статья 8: Право на уважение частной и семейной жизни.

ESC.

Все области социальных прав, социального обеспечения, социальной сплоченности и т.д. Включая, но не ограничиваясь: многие аспекты занятости (в том числе, но не ограничиваясь, мониторинг и наблюдение, проверка рабочих мест и работа в платформенной экономике и др;).

Так же различные аспекты здравоохранения (право на наивысший достижимый уровень здоровья); так же образование. В равной степени на социальную защиту, интеграцию и участие, не говоря уже о недискриминации, жилье и защита от социальной изоляции.

Например:

правосудие (как в отношении отправления правосудия, так и в отношении уголовного правосудия и тюрем;

торговля людьми (принудительный труд и эксплуатация, ...);

миграция и беженцы;

гендерное равенство, плюс насилие в отношении женщин;

дети и молодежь, плюс образование;

биоэтика;

недискриминация, рома и тревеллеры, СОГИ; наркополитика;

участие и культура; спорт;

Приложение 3. Принципы

Человеческий контроль

Приложения ИИ должны обеспечивать значимый контроль со стороны субъектов данных над обработкой данных и соответствующими последствиями для человека и общества.

Ценностно-ориентированный дизайн

Разработчики, производители и поставщики услуг ИИ должны применять ценностно-ориентированный подход при разработке своих продуктов и услуг в соответствии с Конвенцией 108+, в частности со статьей 10.2, и другими соответствующими документами Совета Европы.

Предостерегающий подход.

Разработчики, производители и поставщики услуг ИИ должны оценивать возможные негативные последствия применения ИИ для прав человека и основных свобод и, учитывая эти последствия, применять осторожный подход, основанный на соответствующих мерах по предотвращению и снижению рисков.

Подход, основанный на проектировании прав человека, и выявление предвзятости.

На всех этапах обработки, включая сбор данных, разработчики, производители и поставщики услуг ИИ должны применять подход, основанный на правах человека, и избегать любых потенциальных предубеждений, в том числе непреднамеренных или скрытых, и риска дискриминации или других неблагоприятных последствий для прав человека и основных свобод субъектов данных.

Качество и минимизация данных.

Разработчики ИИ должны критически оценивать качество, характер, происхождение и объем используемых персональных данных, сокращая

ненужные, избыточные или маргинальные данные на этапах разработки и обучения, а затем отслеживая точность модели по мере поступления в нее новых данных. Использование синтетических данных может рассматриваться как одно из возможных решений для минимизации объема персональных данных, обрабатываемых приложениями ИИ.

Риск деконтекстуализации

При разработке и использовании приложений ИИ следует адекватно учитывать риск негативного воздействия на человека и общество из-за деконтекстуализированных данных и деконтекстуализированных алгоритмических моделей.

Независимые комитеты экспертов

Консультативный комитет Конвенции о защите физических лиц при автоматизированной обработке персональных данных (Конвенция 108). 2019. Руководство по последствиям искусственного интеллекта для защиты данных.

Разработчикам, производителям и поставщикам услуг ИИ рекомендуется создавать независимые комитеты экспертов из различных областей, а также взаимодействовать с независимыми академическими институтами, которые могут способствовать разработке приложений ИИ, основанных на правах человека, этически и социально ориентированных, а также выявлению потенциальной предвзятости. Такие комитеты могут играть особенно важную роль в областях, где прозрачность и взаимодействие с заинтересованными сторонами могут быть более сложными из-за конкурирующих интересов и прав, например, в области предиктивного правосудия, предотвращения преступлений.

Конвенция 108 и обнаружения преступлений.

Приложение 4. Защита данных

Обязательные и необязательные документы в области защиты данных

III.7 Необходимо создать соответствующие механизмы для обеспечения независимости комитетов экспертов.

Участие и демократический надзор за разработкой ИИ

Следует поощрять партисипативные формы оценки рисков, основанные на активном вовлечении отдельных лиц и групп, на которые потенциально может повлиять применение ИИ.

III. 8. Отдельные лица, группы и другие заинтересованные стороны должны быть информированы и активно вовлечены в дебаты о том, какую роль ИИ должен играть в формировании социальной динамики и в процессах принятия решений, затрагивающих их.

Человеческий надзор

Все продукты и услуги должны быть разработаны таким образом, чтобы обеспечить право человека не быть объектом решения, существенно влияющего на него, основанного исключительно на автоматизированной обработке, без учета его мнения.

Свобода выбора

В целях повышения доверия пользователей разработчикам, производителям и поставщикам услуг ИИ рекомендуется разрабатывать свои продукты и услуги таким образом, чтобы гарантировать свободу выбора пользователей в отношении использования ИИ, предоставляя им реальные альтернативы применения ИИ.

Бдительность алгоритмов

Разработчики ИИ, производители и поставщики услуг должны внедрять такие формы бдительности в отношении алгоритмов, которые способствуют подотчетности всех соответствующих заинтересованных сторон на протяжении всего жизненного цикла этих приложений, чтобы обеспечить соблюдение законов и принципов в области защиты данных и прав человека.

Прозрачность и расширяемость

Субъекты данных должны быть проинформированы, если они взаимодействуют с приложением ИИ, и имеют право на получение информации

о причинах, лежащих в основе применяемых к ним операций по обработке данных ИИ. Это должно включать последствия таких рассуждений.

Право на возражение

Право на возражение должно быть обеспечено в отношении обработки, основанной на технологиях, которые влияют на мнения и личностное развитие человека.

Подотчетность и бдительность

Без ущерба для конфиденциальности, гарантированной законом, процедуры государственных закупок должны налагать на разработчиков, производителей и поставщиков услуг ИИ особые обязанности по обеспечению прозрачности, предварительной оценке влияния обработки данных на права человека и основные свободы, а также бдительности в отношении потенциальных негативных эффектов и последствий применения ИИ (далее – бдительность в отношении алгоритмов).

Свобода человека, принимающего решения

4. Чрезмерная зависимость от решений, предлагаемых приложениями ИИ, и боязнь оспаривать решения, предлагаемые приложениями ИИ, могут привести к изменению автономности человеческого вмешательства в процессы принятия решений. Поэтому следует сохранить роль человеческого вмешательства в процессы принятия решений и свободу людей, принимающих решения, не полагаться на результаты рекомендаций, предоставляемых с помощью ИИ.

НАЦИОНАЛЬНЫЕ ПЕРСПЕКТИВЫ РЕГУЛИРОВАНИЯ СИСТЕМ ИИ

ГЛАВА I. Использование инноваций: израильские перспективы этики и управления ИИ

Профессор Исаак Бен-Исраэль, профессор Эвиатар Матания, Лихе Фридман

I. Резюме

В данной статье излагается текущее состояние дел в Израиле в области разработки политики, касающейся возможностей и проблем, связанных с искусственным интеллектом (ИИ) в отношении прав человека и этики. В значительной степени она основана на отчете Национальной инициативы Израиля по безопасным интеллектуальным системам, который был недавно представлен израильскому правительству. В настоящем обзоре описывается уникальный подход Израиля к попытке использовать возможности, предоставляемые ИИ, и одновременно решить проблемы, которые он создает. В статье описывается, как израильский подход к управлению на сегодняшний день стремится сбалансировать необходимость создания условий для инноваций, как в государственном, так и в частном секторе, с моральными и правозащитными императивами, которые вездесущи в разработках ИИ.

Израильские политики склонны рассматривать развитие ИИ не только как разрушительное, но и как преобразующее: Технология ИИ рассматривается как критически важная для благосостояния, экономики и безопасности граждан Израиля. Принимая это за отправную точку, приоритетной задачей для Израиля стало создание целостной и устойчивой экосистемы ИИ, движущей силой которой является частный сектор, но в которой участвуют правительство, частная промышленность и научные круги, и которая поддерживает использование ИИ на всех уровнях. Принимая это во внимание, в данной статье освещаются основные проблемы, которые были определены политиками в Израиле и за рубежом в области прав человека, демократии и верховенства

закона – безопасность, конфиденциальность, автономия, гражданские и политические права, безопасность, справедливость – включая честную конкуренцию – и подотчетность. Предлагаемый Израилем подход к решению этих проблем, согласно отчету Национальной инициативы, открывает новые горизонты. Хотя он прочно опирается на установленные принципы управления и передовую международную практику в области ИИ, он, тем не менее, представляет собой новый подход к управлению, сфокусированный на сбалансированном регулировании для стимулирования инноваций. Для этого в нем предлагаются оригинальные инструменты политики, такие как инструмент оценки рисков, позволяющий подобрать различные подходы к регулированию в зависимости от уровня риска, связанного с конкретной деятельностью, и динамическая карта частот, которая помогает найти проблемные области с точки зрения применения этических ценностей к разработке конкретной системы ИИ.

Это также сигнализирует о необходимости взаимодействия со странами и международными форумами, чтобы учиться и вносить свой вклад в международные процессы, связанные с вопросами ИИ, этики, права и управления.

Данная статья не является официальным правительственным документом и не обязательно отражает политику правительства Израиля. Ее авторы пишут в своем личном качестве, хотя соответствующую информацию они получили от различных правительственных чиновников.

Благодарности

Авторы выражают благодарность Седрику Саббаху, директору отдела международной кибербезопасности и ИТ-права в офисе заместителя генерального прокурора (международное право) Министерства юстиции Израиля, профессору Карин Нахон, главе рабочей группы по этике и регулированию Национальной инициативы по безопасным интеллектуальным системам, доктору Рою Шондорфу, заместителю генерального прокурора

(международное право), и Дрору Бен Моше, руководителю отдела больших данных Министерства здравоохранения, за их ценный вклад, профессиональное руководство и проницательные комментарии.

II. Введение

Израиль рассматривает ИИ как одну из основных развивающихся технологий и как «инфраструктуру инфраструктур, которая имеет решающее значение для будущего государства Израиль – для его безопасности, экономики и благосостояния его населения». Приложения ИИ, благодаря своему потенциалу повышения доступности, надежности и эффективности национальных инфраструктур, услуг и систем при меньших затратах для государства и его граждан, играют ключевую роль в способности Израиля решить некоторые из национальных задач в 21 веке.

В перспективе ИИ, вероятно, коренным образом изменит все аспекты частной и общественной жизни. Чтобы использовать положительный потенциал технологий ИИ, Израиль стремится создать целостную и устойчивую экосистему ИИ, включающую правительство, частную промышленность и научные круги. Петля обратной связи с участием этих трех секторов принесет пользу всему обществу, поскольку: (а) расширения использования приложений ИИ; (б) улучшения работы правительства и предоставляемых им услуг; (в) стимулирования экономики и инноваций новых научно-технических разработок, что в свою очередь вновь повысит спрос на новые приложения ИИ.

Соответственно, израильский подход к ИИ основан на двух взаимодополняющих усилиях:

I. Содействие широкому и справедливому использованию приложений ИИ как в государственном, так и в частном секторе.

II. Содействие развитию ведущей технологической отрасли, которая будет разрабатывать решения на основе ИИ для решения возникающих проблем в Израиле и во всем мире.

Национальная инициатива Израиля по созданию безопасных интеллектуальных систем

Для того чтобы понять текущий ландшафт ИИ в Израиле и изучить, как можно достичь широких политических целей с учетом особенностей израильского общества, премьер-министр Израиля в 2018 году запустил Национальную инициативу по безопасным интеллектуальным системам и назначил двух авторов этой статьи, профессора Исаака Бен-Исраэля и профессора Эвиатара Матанию, соруководителями инициативы. Ее мандат заключался в разработке национального плана в области ИИ и связанных с ним интеллектуальных технологий. В рамках рабочей инициативы был использован многосторонний подход: сотни израильских экспертов в различных областях из академического, промышленного и правительственного секторов вызвались принять участие в этом начинании. Эксперты были разделены на 15 рабочих групп, занимающихся различными технологическими, отраслевыми и межотраслевыми аспектами интеллектуальных систем, в соответствии со следующей моделью:



Внедрения

Здравоохранение и медицина

Транспорт

Безопасность

Финтех

Сельское хозяйство

Отношения между промышленностью и академией и здравоохранение и медицина

Транспорт

Безопасность

Финтех

Сельское хозяйство

Отношения между промышленностью и академией



Кросс-секторный

Кибербезопасность для ИИ

Этика и регулирование

Исследовательские центры

Правительство



Технологии



ИИ, машинное обучение и наука о данных

Робототехника и автономные системы

Распределенный интеллект

Вычислительные мощности и квантовые вычисления

Сенсоры

Каждая рабочая группа проанализировала израильскую среду ИИ в соответствии со своими тематическими перспективами. Рабочие группы использовали в качестве сравнительной точки предыдущие и передовые работы, проведенные в других юрисдикциях, чтобы представить национальный план, адаптированный к специфическим особенностям Израиля. Выводы и рекомендации были включены в итоговый отчет, в котором предлагается Национальная стратегия Израиля в области защищенных интеллектуальных систем («Отчет о национальной инициативе»). В нем интеллектуальные технологии определены как национальный приоритет и разработан оперативный национальный план по созданию устойчивой экосистемы в области защищенных интеллектуальных систем. Национальный план основан на трех уровнях: (1) критические инфраструктуры; (2) благоприятные

инфраструктуры; (3) наращивание потенциала; и состоит из следующих структурных блоков:¹⁹²

Критические инфраструктуры

Развивающие инфраструктуры

Национальные проекты

(Сельское хозяйство, здравоохранение, чрезвычайные ситуации, транспорт)

Вычисления, хранение и связь

Кибербезопасность

Этика и саморегулирование

(сбалансированное и стимулирующее)

Цифровизация государственных услуг + приложения искусственного интеллекта

Человеческий капитал

(Все, студенты, STEM, НИОКР)

Поощрение промышленности

(В целом + монетизация академических исследований)

Исследования

Данные

(Открытые государственные данные)

Отчет о национальной инициативе был недавно представлен премьер-министру Израиля.

III. Применение ИИ в Израиле – возможность для государственной политики

«Израиль сейчас занимает третье место в мире по количеству решений в области ИИ. Имея всего 8,5 миллионов граждан, Израиль занимает долю рынка в 11% и по этому показателю равен Китаю. В Израиле в 40 раз больше ИИ-компаний на душу населения, чем в США, являющихся лидером рынка, и это

192 Ibid. с.23.

делает Израиль явным скрытым чемпионом в области искусственного интеллекта».¹⁹³

Израиль обладает мощной экосистемой высоких технологий и инноваций в сочетании с культурой, которая принимает и адаптируется к технологическому развитию. Преобладающая атмосфера в «стране стартапов» поощряет как государственный, так и частный сектор к изучению и использованию приложений ИИ в различных областях. Однако ландшафт применения ИИ в Израиле формируется, прежде всего, на частном рынке.

1. Частный сектор

По данным Управления инноваций Израиля, инвестиции в израильские высокотехнологичные проекты в области ИИ выросли в 2011-2019 годах в 12,5 раз, с 305 миллионов долларов до 4 миллиардов долларов. В 2019 году 42% от общей суммы, инвестированной в израильский хай-тек, пойдут на технологии ИИ.¹⁹⁴ В настоящее время в Израиле работают около 1 400 компаний, разрабатывающих и использующих технологии ИИ в различных областях, таких как бизнес-анализ, кибернетические приложения, здравоохранение и многое другое. Более 40% компаний занимаются информационными технологиями и организационным программным обеспечением, а 30% – интернет-услугами и коммуникациями.¹⁹⁵

Распределение компаний ИИ по секторам

1024 из этих компаний являются стартапами. Несмотря на небольшие размеры и ограниченные ресурсы, Израиль занимает третье место в мире по количеству ИИ-стартапов после США и Китая и первое место по количеству ИИ-компаний на душу населения.¹⁹⁶ За последние пять лет ежегодно появлялось в среднем 140 новых стартапов, предлагающих приложения и

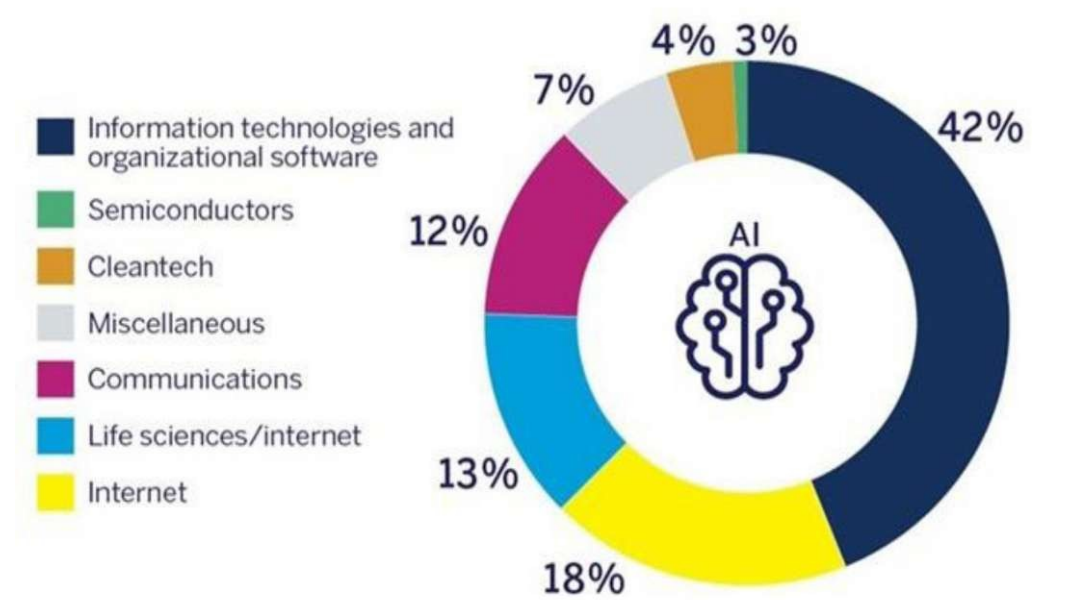
193 ASGARD. Глобальный ландшафт искусственного интеллекта. Получено с сайта <https://asgard.vc/global-ai/7>.

194 Управление инноваций Израиля. (2020). Укрепление искусственного интеллекта. Получено с сайта <https://innovationisrael.org.il/en/reportchapter/bolstering-artificial-intelligence-0#footnoteref2 bscx727>

195 Там же; Израильское управление инноваций. (2020). Инновации в Израиле – Отчет об инновациях за 2019 год. стр.62 (иврит).

196 ROLAND BERGER GMBH & ASGARD. (2018). Искусственный интеллект – стратегия для европейских стартапов. Р. 17

продукты, охватывающие все отрасли и сферы жизни. Однако ведущим сектором является здравоохранение: 188 стартапов (18%) предлагают решения на основе ИИ в таких областях, как диагностика, мониторинг, управление заболеваниями, персонализация и клинический документооборот. Далее следует корпоративное программное обеспечение: 152 стартапа (15%) разрабатывают и используют продукты и услуги ИИ в таких областях, как продажи и обучение, управление персоналом, данные и разведка, клиенты и поддержка, разработка и ИТ, управление и командная работа, безопасность, конфиденциальность и финансы.¹⁹⁷



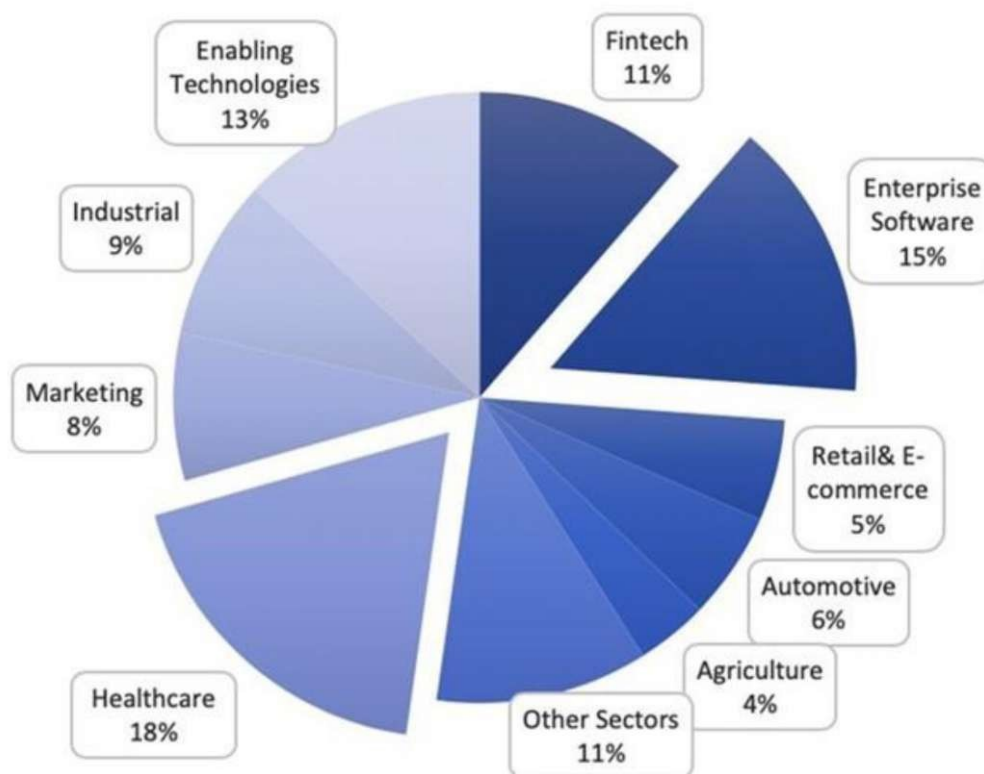
Источник: Израильское управление инноваций (2020). Укрепление искусственного интеллекта

Израиль также занимает первое место по числу компаний, разрабатывающих инфраструктурные технологии для искусственного интеллекта, такие как специализированные чипы, инфраструктурные алгоритмы и сложные системы для ускорения вычислений.¹⁹⁸

197 Cardumen Capital (9 июня 2020) Израильские стартапы в области искусственного интеллекта, июнь 2020. Medium. получено с <https://medium.com/@cardumencapital/israels-artificial-intelligence-startups-june-2020-81e27d9332d8>.

198 Управление инноваций Израиля. (2020). Укрепление искусственного интеллекта. Получено с сайта <https://innovationisrael.org.il/en/reportchapter/bolstering-artificial-intelligence-0#footnoteref2> bscx727.

Israel's AI startups distribution by sectors



Источник: Cardumen Capital. Израильские стартапы в области искусственного интеллекта, июнь 2020 г.

Полную карту израильского ландшафта стартапов в области искусственного интеллекта в 2020 году, сегментированную по секторам и сферам применения, см. в Приложении I.

Государственные инициативы и политика

В дополнение к деятельности частного сектора, правительство Израиля играет важную роль в продвижении приложений ИИ. Оно может сделать это, иницируя проекты самостоятельно или создавая благоприятные условия для частного сектора в целях дальнейшего развития и использования технологий ИИ, например, с помощью благоприятного регулирования, стимулов для отрасли и т. д. Большой потенциал, который правительство приписывает технологиям ИИ и их применению, обусловлен двумя особенностями израильского общества:

Высокие темпы роста населения – демографические показатели Израиля являются исключительными для развитых стран, и поэтому они ставят перед правительством страны некоторые уникальные задачи. С 3,1 ребенка на одну женщину, Израиль имеет самый высокий уровень рождаемости среди стран ОЭСР, где средний уровень рождаемости составляет 1,6 ребенка.¹⁹⁹ Кроме того, средняя продолжительность жизни в Израиле составляет 82,9 года, что является пятым показателем в ОЭСР.²⁰⁰ Следовательно, ежегодный прирост населения Израиля – 1,9%²⁰¹ – почти в четыре раза превышает средний показатель прироста населения стран-членов ОЭСР (0,54%).²⁰² Постоянный быстрый рост населения требует от государства Израиль корректировки государственных и социальных услуг, поддержания и соответствующего увеличения инфраструктуры во всех сферах жизни (например, здравоохранение, транспорт, образование, энергетика и т.д.).

Плотность населения и перегрузка инфраструктур – большинство населения Израиля живет в центре страны, а более 40% населения распределено менее чем на 7% территории страны.²⁰³ Это приводит к серьезным перегрузкам инфраструктуры и услуг в густонаселенных районах. Одним из ярких примеров является растущая загруженность дорог в центре страны, которая негативно сказывается на производительности труда, состоянии окружающей среды и количестве аварий и жертв.

Высокие темпы роста населения и его плотность определяют подход Израиля к ИИ, поскольку они значительно увеличивают потребность Израиля в национальных инфраструктурах и услугах. Мотивация правительства к расширению использования ИИ, помимо тех областей, которые уже охвачены частным сектором, заключается в потенциале этой технологии для

199 ОЭСР (2020), Уровень рождаемости (показатель). Получено с сайта <https://data.oecd.org/pop/fertility-rates.htm#indicator-chart>.

200 ОЭСР (2020), Ожидаемая продолжительность жизни при рождении (показатель). Получено с сайта <https://data.oecd.org/healthstat/life-expectancy-at-birth.htm>.

201 Центральное бюро статистики. (2020). Israel in Figures Selected Data From the Statistical Abstract of Israel 2019. Retrieved from https://www.cbs.gov.il/he/publications/DocLib/isr_in_n/isr_in_n19e.pdf p.6.=

202 Всемирный банк, Рост численности населения для членов ОЭСР [SPPOPGRWOED]. Получено из FRED, Федеральный резервный банк Сент-Луиса <https://fred.stlouisfed.org/series/SPPOPGRWOED>

203 Центральное бюро статистики. (2020). Статистическая аннотация Израиля 2019 – №70. Retrieved from https://www.cbs.gov.il/he/publications/DocLib/2019/Shnaton70_mun.pdf (иврит). p.21.

удовлетворения растущей потребности в повышении доступности, надежности и эффективности государственных услуг и национальных инфраструктур при меньших затратах для государства и его граждан. По этой причине одними из первых секторов, для которых правительство продвигает решения на основе ИИ, являются здравоохранение и транспорт, где перегрузка существующих инфраструктур ощущается наиболее остро.

Однако правительственные министерства Израиля различаются по степени готовности к внедрению приложений ИИ из-за различий в уровне цифровой зрелости, а в некоторых случаях даже из-за пробелов в цифровых технологиях. Национальная цифровая инициатива Израиля, также называемая «Цифровой Израиль», является государственным органом, ответственным за предоставление услуг электронного правительства. Digital Israel возглавляет усилия правительства по цифровой трансформации, направленные на сокращение социально-экономических разрывов, содействие экономическому процветанию и создание более умного и дружелюбного правительства. Сфера его деятельности охватывает широкий спектр услуг электронного правительства на всех уровнях власти (включая муниципалитеты). Так, Digital Israel сотрудничает с другими правительственными министерствами, помогая им в разработке и внедрении планов оцифровки. Digital Israel также возглавляет правительственные инициативы по созданию «умных городов» и Национальный план по цифровой грамотности.

В значительной степени выполнение мандата Digital Israel зависит от доступности и возможности передачи данных. Действительно, для того чтобы максимально использовать весь потенциал цифровой трансформации, государственные органы должны иметь возможность собирать большие объемы информации, объединять ее с данными из других источников и использовать технические средства для анализа данных и выводов. Во многих случаях данные, которые необходимо собирать и передавать, включают персонально идентифицируемую информацию («PII»). Кроме того, программные инструменты, которые могут быть использованы, включают анализ больших

данных, значительная часть которых может быть усовершенствована с помощью машинного обучения. Таким образом, проекты, которые Digital Israel хочет реализовать, сталкиваются с ключевой проблемой – как найти баланс между потребностями в данных, с одной стороны, и юридическими и этическими соображениями – с другой.

На сегодняшний день не существует всеобъемлющей государственной политики для решения этой проблемы. В настоящее время работа над такой политикой завершается. До ее принятия деятельность Digital Israel основывается на существующей правовой базе, которая включает конституционные меры защиты прав человека (неприкосновенность частной жизни, недискриминация, свобода слова), а также нормы административного права и принципы, применимые в целом ко всем государственным органам (прозрачность, подотчетность, справедливость, надлежащий процесс и разумность).

Ниже приводится описание основных проектов с использованием ИИ, которые находятся в процессе разработки (в каждом случае совместно с соответствующим государственным министерством). При разработке каждого из этих проектов Digital Israel сотрудничала с корпоративным юристом и конституционным юристом Министерства юстиции, чтобы обеспечить соответствие разработки и внедрения каждого проекта применимым ограничениям конституционного и административного права.

Цифровое здоровье

Проект имеет амбициозные цели, включая:²⁰⁴

Индивидуальное лечение: содействие исследованиям, разработке и внедрению инструментов, позволяющих пациенту получить наилучшее и наиболее персонализированное лечение;

204 Домашняя страница цифровых услуг Министерства здравоохранения (иврит)
<https://www.health.gov.il/About/projects/DigitalHealth/Pages/default.aspx>.

Укрепление здоровья и профилактика пациентов путем использования цифровых инструментов таким образом, чтобы сместить акцент с обслуживания пациентов на профилактическую медицину;

Устойчивое здоровье: содействие разработке и внедрению систем, повышающих операционную и управленческую эффективность системы здравоохранения, таким образом, чтобы высвободить имеющиеся ресурсы;

Разработка и внедрение цифровых инструментов, упрощающих коммуникацию между Министерством здравоохранения и теми, кого оно обслуживает;

Предоставление услуг неотложной помощи с помощью системы управления записью на прием и приложения, информирующего пациента о ходе лечения. Собранная информация позволит лучше управлять ресурсами, чтобы избежать перегруженности отделений неотложной помощи;

Расширен обмен клинической информацией между платформами поставщиков услуг, что позволяет связать различных поставщиков услуг и дает им возможность просматривать методы лечения и диагнозы, поставленные другими медицинскими работниками в различных организациях.

Некоторые из этих целей уже реализуются. Постановлением правительства 243 Министерство здравоохранения создало платформу под названием «TIMNA», которая предоставляет третьим сторонам контролируемый доступ к медицинским данным с целью содействия прикладным исследованиям. Эти данные включают огромное количество медицинских записей, собранных больницами и клиниками со всей страны, что представляет собой бесценный ресурс. Доступ к этим данным регулируется строгими ограничениями в отношении конфиденциальности и этики. Во-первых, учреждения, исследователи и стартапы, желающие получить доступ к этим данным, должны предоставить одобрение Хельсинкского комитета на свой проект. Они должны определить конкретные типы данных, которые им нужны, и только эти данные предоставляются. Они должны подписать обязательства о конфиденциальности. Все персональные данные

анонимизируются. Исследования проводятся полностью в рамках оцифрованной платформы – никакие личные данные не могут быть извлечены из платформы, что еще больше защищает конфиденциальность. Кроме того, перед публикацией исследования и использованием алгоритма Министерство здравоохранения проверяет его, чтобы убедиться, что никакая личная информация не используется и не раскрывается. На сегодняшний день проекты TIMNA, использующие ИИ, выглядят следующим образом:

Израильский центр контроля заболеваний применяет алгоритм ИИ для проверки диагностических форм пациентов и подтверждения диагноза рака. Это экономит значительное количество времени, поскольку автоматизирует процесс проверки более 100 000 бланков в год. Проводятся проверки, чтобы убедиться в отсутствии ложноотрицательных результатов.

Аналогичным образом разрабатывается алгоритм для помощи в медицинском наблюдении в двух областях: развитие ребенка и беременность. В обоих случаях алгоритм в режиме реального времени анализирует отчеты о состоянии, диагнозы, записи и рекомендации врачей и медсестер, сравнивая их со стандартными протоколами. Затем он предупреждает больницу или клинику о возможных ошибках, неправильном диагнозе или проблемах, которые могут потребовать дополнительного тестирования или последующего наблюдения. Система направлена на оказание помощи медицинским работникам в выявлении ошибок и не влечет за собой значительных этических рисков для пациента.

Другая область исследований – использование ИИ для анализа медицинских изображений (МРТ, КТ и т.д.). Пока результаты многообещающие, поскольку алгоритмы смогли обнаружить случаи, которые были пропущены врачами. Цель состоит не в том, чтобы заменить врача в принятии решений, а в том, чтобы упростить процесс и помочь ему/ей в анализе изображений.

Наконец, Министерство здравоохранения задействовало алгоритмы искусственного интеллекта, чтобы помочь в работе по замедлению

распространения COVID-19. Часто эпидемиологическое исследование, основанное на беседах с инфицированным человеком, оказывается неполным из-за того, что он не помнит все посещенные места, интервьюер не смог правильно ввести всю информацию или из-за других ошибок человека. Алгоритмы используются для формирования более полной картины вероятного течения предыдущих инфекций и прогнозирования будущих инфекций. Эта информация затем используется для формирования государственной политики в отношении мер по закрытию в общем масштабе. Следует отметить, что информация не используется для принятия решений в отношении конкретных людей или сообществ.

Перевозка

Министерство транспорта создает пилотный проект для проведения испытаний автономных транспортных средств. Проект позволит производителям подавать заявки на получение специальной лицензии, в соответствии с которой они смогут тестировать свой продукт в реальных условиях, в среде вождения с низким уровнем риска. С этой целью Министерство опубликовало законопроект, который открыт для публичных комментариев.²⁰⁵ Законопроект еще не содержит всех правил, которые будут применяться к испытаниям или к требованиям, предъявляемым к системе автономного вождения (возможности, безопасность и надзор). Эти вопросы, а также этические проблемы, которые могут возникнуть, будут рассмотрены на более позднем этапе. Этот подход отражает философию осторожных и постепенных инноваций: чтобы понять воздействие технологии и учитывая риск для человеческой жизни, испытания разрешены в среде, которая не является «контролируемой», но представляет относительно низкий риск. Это должно позволить политикам внести коррективы, прежде чем приступить к более масштабным экспериментам.

²⁰⁵ С текстом законопроекта можно ознакомиться здесь: https://www.nevo.co.il/law_word/law11/200820-2.doc (иврит).

В области общественного транспорта также наблюдается прогресс, поскольку разрабатываются инновационные решения для снижения загруженности дорог.²⁰⁶

Налогообложение

Налоговое управление Израиля запустило проект, призванный помочь следователям в выявлении налогового мошенничества. В проекте используются инструменты искусственного интеллекта для прогнозирования вероятности налогового мошенничества на основе определенных показателей. При разработке проекта основное внимание уделялось вопросам конфиденциальности. Действительно, проект основан на многоуровневом подходе к доступу к информации: на начальном этапе для выявления риска мошенничества используется небольшое количество показателей и мало информации; если первоначальное расследование предполагает более высокий риск мошенничества, только после этого требуется личная информация, чтобы с большей уверенностью установить личность потенциального преступника.

Предложения по национальным проектам

В дополнение к вышеупомянутым реализуемым проектам, Отчет о национальной инициативе рекомендует правительству в сотрудничестве с промышленным и академическим секторами запустить еще четыре национальных проекта в области здравоохранения, транспорта, безопасности и сельского хозяйства. Все четыре проекта были разработаны для удовлетворения реальных национальных потребностей, обусловленных изменениями в израильском обществе, которые вызывают потребность в улучшенных и новых инфраструктурах и услугах.

Здравоохранение – Реформирование национальной системы здравоохранения путем повышения качества и доступности медицинских услуг, а также снижения перегрузки больниц путем запуска национальной системы,

206 Бен Дроп, М. и Азария, М. (24 июля 2020 года). Израильские «умные поездки» показывают, каким может быть общественный транспорт после COVID-19. Всемирный экономический форум. Получено с сайта <https://www.weforum.org/agenda/2020/07/israel-smart-commuting-after-covid-public-transport-innovation/>.

основанной на интеллектуальных технологиях для: (I) дистанционного управления пациентами; (II) более эффективной сортировки и лечения в неотложной медицине; (III) создания сравнительных индексов качества для оценки клинических результатов.²⁰⁷

Транспорт – «Из множества способов, которыми интеллектуальные системы могут решить острые проблемы в области транспорта, мы решили рекомендовать на первом этапе установку «умных» светофоров во всей пилотной столичной области с целью решения проблемы пробок, которая является самой серьезной и острой транспортной проблемой в Израиле».²⁰⁸

Безопасность – создание национальной двойной системы, которая будет использовать потенциал интеллектуальных технологий для улучшения способности прогнозирования и процессов принятия решений, для более эффективного управления на национальном уровне. Собирая и анализируя данные для гражданского применения командования и управления в обычное время, она повысит национальный потенциал для подготовки к чрезвычайным ситуациям (стихийные бедствия, эпидемии и угрозы безопасности со стороны врагов) и принятия решений во время кризисов.²⁰⁹

Сельское хозяйство – «В эпоху, когда продовольственная безопасность, управление водными ресурсами и другие области сельского хозяйства становятся острыми глобальными проблемами, [...] Мы рекомендуем продвигать национальный проект по разработке системы раннего обнаружения вредителей и болезней сельскохозяйственных культур на основе интеллектуальных технологий; наряду с интеграцией интеллектуальных систем в сельскохозяйственный сектор для оптимизации использования природных ресурсов и [других] факторов производства для обеспечения оптимального производства продовольствия».²¹⁰

207 Бен-Израэль, И., Матания, Е. и Фридман, Л. (ред.) (сентябрь 2020 г.). Национальная инициатива по безопасным интеллектуальным системам для укрепления национальной безопасности и научно-технической устойчивости: Национальная стратегия для Израиля. Специальный доклад премьер-министру. (Иврит) стр.35.

208 Ibid. с.37

209 Ibid. с.36

210 Ibid. с.36

Наконец, признавая различия в уровне цифровой зрелости в правительстве, в рамках Национальной инициативы была создана рабочая группа, занимающаяся правительственным сектором, с целью разработки руководящих принципов подготовки всего правительства к эпохе ИИ. Рабочая группа оценила необходимые организационные, технологические и регулятивные меры для содействия внедрению приложений ИИ в правительстве, чтобы улучшить как межведомственную работу, так и взаимодействие между правительством и гражданами.²¹¹

Риски и вызовы, создаваемые ИИ в области прав человека, демократии и верховенства закона

Темы прав человека и этики были рассмотрены в рамках Национальной инициативы специальной рабочей группой, которая глубоко изучила этот вопрос с нормативной и этической точек зрения. В отчете рабочей группы («Отчет по этике»)²¹² определены вопросы, которые являются новыми и уникальными для ИИ и тех способов, которыми, как ожидается, он повлияет на нашу жизнь. В свете этих вопросов в нем излагается ряд этических проблем, возникающих в связи с технологией, касающихся прав человека, демократии и верховенства закона.

Ниже приводится обзор вопросов, выявленных в Отчете по этике, а затем обзор рисков и этических проблем, которые предлагается решить.

I. Что нового и особенного в ИИ?

«Системы ИИ склонны радикализировать существующие социальные отношения. Например, если существует неравенство между различными

211 Шарвит, С. и др. (2020). Отчет правительственной рабочей группы. In Ben-Israel, I., Matania, E. & Friedman, L. (Eds.). The National Initiative for Secured Intelligent Systems to Empower the National Security and Techno-Scientific Resilience: Национальная стратегия для Израиля. Специальный доклад премьер-министру. стр. 214-227. (Иврит).

212 Nahon K., Ashkenazi A., Gilad Bachrach R., Ken-Dror Feldman D., Keren A. and Shwartz Altshuler T. (2020). Отчет рабочей группы по этике и регулированию искусственного интеллекта. In Ben-Israel, I., Matania, E. & Friedman, L. (Eds.). The National Initiative for Secured Intelligent Systems to Empower the National Security and Techno-Scientific Resilience: Национальная стратегия для Израиля. Специальный доклад премьер-министру. стр. 172-119. (Иврит).

социальными группами, системы ИИ могут воспроизводить и даже усугублять его. Это касается дискриминации, стереотипов, нарушений прав, политического экстремизма и т.д. Для удобства мы продемонстрируем это утверждение на примере неравенства. Существует несколько основных причин этого явления:

Поскольку системы ИИ зависят от предоставляемой им информации, их входные данные могут отражать уже существующее неравенство, а если введенными данными манипулировали, система узнает об этой манипуляции.

Системы ИИ становятся все более распространенными во все большем числе социальных контекстов. Поэтому их воздействие – и потенциальные предубеждения – затрагивают все большую аудиторию.

Существует ошибочная тенденция рассматривать продукты систем ИИ, которые быстро и масштабно анализируют данные, как научную истину. Следовательно, существует опасность, что такие системы не будут подвергаться контролю, применяемому к эквивалентным человеческим решениям, когда возникает подозрение в предвзятости.

Из-за сложности систем трудно заранее предвидеть и проверить их поведение. Следовательно, зачастую трудно отличить «истинный» диагноз, основанный на достоверном процессе анализа и мониторинга, как это делается в отношении классических алгоритмов или принятия решений человеком, от необъективного диагноза.»²¹³

Процедурная проблема: как «спроектировать» ценности. Этот вопрос возникает в тех областях, где системы ИИ разрабатываются для замены людей, принимающих решения, которые обладают необходимыми навыками и полномочиями для применения нормативных соображений. «При разработке систем ИИ, которые заменяют человеческое усмотрение, ответственность за эти нормативные соображения переходит от профессионалов, таких как врачи и юристы, к инженерам и информационным ученым, что не так часто происходит при работе с классическими алгоритмами».²¹⁴

213 Ibid. с. 179.

214 Ibid. с. 180.

Риски конфиденциальности и автономности беспрецедентного масштаба.

Сложность, подрывающая доверие общества. Отсутствие ясности и понимания общественностью принципов работы систем ИИ и их влияния на нашу жизнь часто приводит к недоверию, что может привести к нежеланию принимать технологию даже в тех областях, где системы ИИ предлагают очевидные преимущества для бизнеса и общества. В докладе упоминается в этой связи оценка экспертной группы высокого уровня ЕС по искусственному интеллекту.²¹⁵

«Несправедливая экономия на масштабе». [...] мощные игроки, обладающие большими данными, необходимыми для разработки систем ИИ, используют преимущества интернет-экономии от масштаба для формирования путей выхода на рынок новых игроков, что негативно сказывается на конкурентоспособности. Когда речь идет о совершенно новых игроках, тот факт, что у них нет необходимого объема данных, может означать, что им фактически закрыт доступ на рынок ИИ».²¹⁶

«Изменения в привычных категориях гарантии. Возможность сбора и обработки данных с помощью продуктов [IoTs] позволяет компаниям предлагать новые сопутствующие услуги, но при этом возникают новые вопросы о гарантии на эти услуги, а также о разделении ответственности между производителем и теми, кто предоставляет услуги на практике. Интегрированные ИИ продукты, в частности, также включают в себя сочетание физического продукта и удаленной вычислимости и работоспособности. Таким образом, классическое разделение между продуктом и услугой, а также гарантией на продукт и гарантией на услугу требует пересмотра. Ситуация становится еще более сложной, когда такие продукты и услуги используются другими хозяйствующими субъектами. Например, когда продуктовая сеть использует беспилотник для доставки товаров. Беспилотник способен летать,

215 Этические руководящие принципы для заслуживающего доверия ИИ, Группа экспертов высокого уровня по искусственному интеллекту, ЕС, 2019, <https://ec.europa.eu/futurium/en/ai-alliance-consultation>. (Здесь и далее – ЕС)

216 Нахон К. и др. (2020). с.180.

ориентироваться и работать с окружающей средой. Кроме того, он предоставляет услуги картографии и прогноза погоды. Все это приобретается сетью продуктовых магазинов для скромной цели доставки продуктов».²¹⁷

II. Этические риски и вызовы

В Отчете по этике рассматриваются следующие этические проблемы. Следует отметить, что при разработке различных этических принципов он в значительной степени опирается на экспертную группу высокого уровня ЕС по искусственному интеллекту.

Безопасность

Право на жизнь и безопасность часто упускается из виду в глобальных дискуссиях о правах человека. Тем не менее, оно остается самым фундаментальным правом из всех, без которого другие права человека не могут быть применены. На базовом уровне, применительно к ИИ, Отчет по этике подчеркивает необходимость обеспечения безопасности приложений ИИ и сетей и компьютеров, использующих ИИ. В докладе делается несколько важных замечаний по этому поводу. В нем отмечается, что информация – это «энергия, которая питает нынешнюю волну ИИ», поэтому безопасность приложений и сетей ИИ является необходимым условием для разработки и внедрения технологий ИИ.²¹⁸ В нем также отмечается, что информация, которая может быть собрана ИИ для создания и развертывания инструментов ИИ, включает огромные объемы личной и коммерческой информации, в том числе личной, медицинской, экономической и другой конфиденциальной информации. В документе отмечается, что даже информация, которая кажется «нечувствительной», может стать чувствительной при перекрестных ссылках с другой информацией.

²¹⁷ Ibid. с. 181.

²¹⁸ Ibid. с.190.

С точки зрения политики, это напрямую связывает безопасность с правами человека. Например, защита частной жизни требует, на фундаментальном уровне, защиты частной информации от злонамеренных кибер-операций. Аналогичным образом, помимо коммерческих и эксплуатационных соображений, защита целостности данных также является императивом прав человека: для защиты от предвзятости в конкретном приложении ИИ данные, а также алгоритмы, на которых основано приложение, не должны быть подделаны. Свобода слова и доступ к информации также в значительной степени зависят от безопасности приложений и сетей, связанных с ИИ.

Использование ИИ еще больше усиливает зависимость от компьютеров, а значит, создает новые уязвимости для кибератак. В рамках национальной инициативы специальная рабочая группа по кибербезопасности в эпоху интеллектуальных систем определила новые киберугрозы, связанные с технологиями ИИ. Во-первых, атаки на системы ИИ, которые могут привести к повреждению механизма принятия решений, что приведет к ложным, вводящим в заблуждение или предвзятым решениям, или к угрозам против ИС системы.

Во-вторых, злонамеренное использование возможностей ИИ в качестве кибероружия для проведения сложных «интеллектуальных атак». Эти уязвимости систем ИИ поднимают вопрос о том, можно ли и как подтвердить подлинность системы ИИ как защищенной и надежной.²¹⁹ Таким образом, безопасность ИИ — это базовый слой, поверх которого может происходить более богатое взаимодействие. Ее важность отражена в названии израильской национальной инициативы по ИИ — «Национальная инициатива по безопасным интеллектуальным системам».

219 Зак, Х. и др. (2020). Отчет рабочей группы по кибернетическим и интеллектуальным системам. In Ben-Israel, I., Matania, E. & Friedman, L. (Eds.). The National Initiative for Secured Intelligent Systems to Empower the National Security and Techno-Scientific Resilience: Национальная стратегия для Израиля. Специальный доклад премьер-министру. стр. 168-171. (Иврит). стр.168.

Конфиденциальность

Примечательно, что неприкосновенность частной жизни — это первое и главное из прав человека, рассматриваемых в Отчете по этике. В докладе подчеркивается, что приложения ИИ «в значительной степени основаны на информации о людях или на получении выводов о них на основе информации, позволяющей установить их личность».²²⁰ Защита частной жизни в значительной степени зависит от надежной законодательной базы. В Израиле эта база состоит в основном из Основного закона Израиля: Достоинство и свобода человека (1992),²²¹ Закон о неприкосновенности частной жизни (1981), несколько положений о неприкосновенности частной жизни, включая положения о передаче данных²²² и всеобъемлющие Положения о защите частной жизни (безопасности данных) 2017 года.²²³ Правовой режим дополняется обширным прецедентным правом и надежной судебной системой. Со временем сформировался ряд принципов защиты частной жизни: необходимость законных оснований для сбора и обработки информации (например, информированное согласие), ограничения на использование, право на просмотр и исправление личной информации, прозрачность по отношению к владельцу информации и обязательство защищать информацию.

На этом фоне Отчет по этике отмечает, что остаются пробелы между традиционными представлениями о защите частной жизни и проблемами, возникающими в связи с ИИ. Действительно, в той степени, в которой ИИ опирается на сбор и обработку, можно ожидать возникновения новых проблем в сфере защиты частной жизни, что может потребовать адаптации существующих законов о защите частной жизни в будущем.

220 Нахон К. и др. (2020). с.188.

221 <http://knesset.gov.il/laws/special/eng/BasicLawLiberty.pdf>. Это квазиконституционный закон, основополагающие принципы которого рассматриваются как конституционно закрепленные, даже при отсутствии официальной письменной конституции.

222 Положение о защите частной жизни (передача данных в базы данных за рубежом) (2001).

См. неофициальный перевод здесь: <https://www.gov.il/BlobFolder/legalinfo/legislation/en/PrivacyProtectionTransferofDataabroadRegulationsun.pdf>

223 Неофициальный перевод см. здесь:

<https://www.gov.il/BlobFolder/legalinfo/data%20PROTECTION%20OF%20PRIVACY%20REGULATIONS.pdf> security

Отчет по этике также отмечает, что в некоторых случаях может возникнуть конфликт между конфиденциальностью и справедливостью. Если люди, принадлежащие к определенной группе, воздерживаются от предоставления своей личной информации приложению ИИ, это приложение не сможет использовать данные, учитывающие эту группу, что может привести к усилению дискриминации.²²⁴ Таким образом, политический императив заключается в том, чтобы разрешить сбор данных, обеспечив при этом безопасность таких данных и надежную защиту конфиденциальности.

Автономия

Отчет по этике определяет автономию как «способность человека принимать разумные решения, включая предотвращение несправедливого или неосознанного влияния на поведение человека».²²⁵ В терминах прав человека это может относиться к таким понятиям, как человеческое достоинство и право на доступ к информации. В отчете по этике говорится²²⁶:

«Автономия основана не только на способности человека выбирать между вариантами, но и на доступности информации, позволяющей сделать обоснованный выбор, и оценке ее достоверности. Эти вопросы нельзя считать само собой разумеющимися в эпоху ИИ. Более того, способность проводить глубокий анализ информации о человеке с помощью ИИ позволяет разрабатывать высокоинтрузивные попытки убеждения, что опять же имеет потенциальные последствия, которые пока не до конца понятны.

Автономия также связана с диапазоном решений, принимаемых человеком при взаимодействии с технологией, который технология может сузить. Поэтому мы должны всегда исследовать, влияет ли данное приложение на автономию и каким образом. Обратите внимание, что в рамках данного обсуждения могут быть случаи, когда автономия изначально более узкая (в

224 Нахон К. и др. (2020). с.188.

225 Ibid. с.182.

226 Ibid. с.189.

силу определенных социально-экономических или нормативных характеристик), или когда более узкая автономия считается более уместной с нормативной точки зрения, что делает специальные шаги по защите свободы выбора необязательными.

Некоторые технологии ИИ, такие как «глубокая подделка», предназначены для получения ненадежной информации, которую трудно отличить от достоверной. Эти технологии способны снизить способность людей понимать реальность и принимать самостоятельные, обоснованные решения, а также подорвать доверие между людьми и между ними и их правительством. Например, мы недалеко от того дня, когда можно будет искусственно создать фильм, в котором лидер объявляет войну, что приводит к катастрофическим результатам. Комитет считает, что Государство Израиль должно изучить способы работы с этими технологиями в отдельном докладе.

Последней областью, имеющей отношение к автономности, является проникновение инструментов ИИ в средства массовой информации. Многие каналы связи используют ИИ для создания индивидуально адаптированных новостей. Этот инструмент имеет много преимуществ, но также представляет опасность избирательного воздействия: определенные группы населения получают стандартизированную информацию и не знают о доказательствах и аргументах, которые не соответствуют их мировоззрению. Это лишает такое население свободы выбора или свободы ознакомления с разнообразием мнений и делает его уязвимым для несправедливых и высокоэффективных кампаний влияния со стороны заинтересованных сторон. В частности, это может позволить иностранным правительствам вмешиваться в выборы».

Гражданские и политические права

Отчет по этике определяет гражданские и политические права как включающие «право избирать, свободу слова и свободу совести религии.»²²⁷ Они лежат в основе демократических ценностей и заслуживают особой защиты. В этом отношении Отчет по этике с озабоченностью отмечает, как происходит

227 Ibid. с.182.

автоматизированное манипулирование глобальным дискурсом, например, чрезмерное усиление одних взглядов при одновременном заглушении других, поляризация дискурса и придание легитимности взглядам, которые могут быть оскорбительными для определенных групп, а также распространение ложной информации в больших масштабах.²²⁸

Все это может нанести вред самому демократическому процессу, внести раскол в общество, подорвать веру в демократический процесс и произвести.

Справедливость

Это широкий этический принцип, который относится к необходимости достижения существенного равенства, предотвращения предвзятости (в информации, в процессе и в продукте), предотвращения дискриминации и избежания увеличения социально-экономического и образовательного разрыва. Отчет по этике отмечает: «Технология не является нейтральной, поскольку она основана на человеческом программировании и различных коммерческих интересах. Более того, системы ИИ основаны на информации, связанной с поведением человека, которая может отражать и усугублять различные виды социальных предубеждений».²²⁹ Он приводит следующие примеры систем искусственного интеллекта, которые вызывают вопросы о справедливости:

Система принимает решение о распределении ресурсов, таких как фонды и медицинское лечение.

Система оценивает кандидатов на рабочее место или высшее образование.

Система оценивает людей с целью уголовного наказания или его смягчения.

Система принимает решения, угрожающие имущественным и финансовым интересам пользователей.

228 Ibid. с.190.

229 Ibid. с.184.

Для устранения этих рисков в Отчете по этике подчеркивается важность проактивного изучения целевой аудитории и заблаговременного выявления групп, которые могут быть искажены или недопредставлены. Кроме того, в нем подчеркивается необходимость проведения консультаций с представителями самих целевых пользователей, чтобы помочь создать более справедливые системы.²³⁰

Подотчетность

Отчет по этике разделяет подотчетность на три категории: прозрачность, объяснимость и ответственность.

Прозрачность — это «предоставление информации о процессе и принятии соответствующих решений» и называется «ключевой ценностью в технологическом развитии и, в частности, в разработке продуктов ИИ». Это как самостоятельная ценность, так и аспект подотчетности. Она позволяет контролировать и реализовывать другие ценности, такие как справедливость. Прозрачность является основным компонентом общественного доверия.

Объясняемость — это способность системы ИИ объяснить процесс принятия решения, будь то отдельному конечному пользователю или на коллективном уровне, если решение затрагивает группу. Она также включает в себя способность системы предоставлять содержательные объяснения самим операторам системы. Отчет по этике принимает позицию Отчета группы экспертов ЕС о том, что объяснимость включает принцип «содержательной информации», то есть уровень предоставляемой информации должен быть достаточным, но не чрезмерно техническим или подробным.

Ответственность предполагает принятие соответствующих правил для предотвращения риска, исходя из контекста и предполагаемой серьезности риска, управление рисками и назначение сотрудника, ответственного за управление рисками. В Отчете по этике отмечается, что разнообразие заинтересованных сторон и сложность систем искусственного интеллекта

230 Ibid.

делают эту задачу особенно трудной. Это усугубляется тем, что системы ИИ также принимают свои собственные «решения».

Безопасность

Отчет по этике признает необходимость устранения рисков для безопасности, возникающих при использовании систем ИИ. Действительно, чем больше система ИИ наделена правом принимать решения, оказывающие непосредственное влияние на жизнь человека, тем более рискованным является ее использование. Риск возникает как при обычной эксплуатации системы ИИ, так и в экстремальных ситуациях. Таким образом, в Отчете по этике проводится различие между рисками безопасности, возникающими в результате неисправности, и рисками, возникающими, когда система работала без сбоев, но таким образом, что это, тем не менее, привело к смерти или физическому ущербу.

Риски безопасности могут быть уменьшены путем реализации ряда мер. Например, чтобы предотвратить принятие неверных решений, основанных на ошибочных предубеждениях, следует использовать разнообразный набор данных. Аналогичным образом, соображения безопасности должны учитываться на этапе проектирования. В отчете по этике приводится интересный и, возможно, контринтуитивный пример. При разработке автономного автомобиля важно, чтобы система тренировалась в различных условиях, включая условия, в которых вред может быть неизбежен. Для этого процесса обучения необходимо поместить людей в рискованные ситуации, которые затем могут стать частью набора данных. Конечно, хотя бы частично это можно сделать с помощью симуляций без участия физических людей, но это подчеркивает противоречия, которые существуют между различными этическими принципами и принципами прав человека, а также компромиссы, которые иногда необходимы для того, чтобы системы ИИ были как можно более «хорошими».

Справедливая конкуренция

Среди различных этических принципов, которые часто упоминаются в различных документах и стандартах по этике ИИ, честная конкуренция, вероятно, цитируется реже всего. Напротив, в Отчете по этике он является отдельным этическим принципом. Справедливая конкуренция подразумевает необходимость предоставления новаторам, предпринимателям, разработчикам программного обеспечения и другим участникам цепочки поставок возможности пользоваться равным доступом к данным и возможностями для создания и внедрения систем ИИ. На первый взгляд, этот принцип может быть охарактеризован как замаскированная экономическая цель, однако в «Докладе по этике» он рассматривается с учетом прав человека и этических принципов. Действительно, в докладе отмечается, что честная конкуренция необходима «для инноваций и социального благосостояния». Таким образом, поддержание свободного рынка с честной конкуренцией позволит всем участникам цепочки создания стоимости, особенно малым и средним предприятиям и стартапам, получать выгоду и прибыль от своей деятельности. Этический императив носит общественный, а не индивидуальный характер: для того чтобы общество смогло воспользоваться преимуществами трансформации ИИ, необходимы более активные инновации со стороны самых разных участников. Как отмечается в докладе, примеры проблем, возникающих в результате недобросовестной конкуренции, включают:

Система создает преимущество для конкурентов, обладающих большими данными.

Система основана на большой базе данных, доступной лишь немногим участникам рынка.

В процессе своей деятельности система создает большую и уникальную базу данных, недоступную для конкурентов.

Соглашения о неконкуренции и автоматическая координация между компаниями на основе систем искусственного интеллекта. В отчете по этике отмечается: «Концентрация экономической власти может также привести к

концентрации политической власти, позволяя технологическим гигантам диктовать правила игры на рынке». Опасение заключается в том, что влияние этих мега-игроков на рынок может затруднить выход на рынок новых технологий или приложений и поставить под угрозу инновации, столь необходимые для ИИ.²³¹ В нем также предлагается обновить законы о конкуренции, стандартизированные контракты и защиту прав потребителей, чтобы соответствовать ожидаемым вызовам. «К этому следует добавить международный вызов, обусловленный тем, что некоторые из ключевых игроков базируются в США».²³² Короче говоря, основная часть проблемы заключается в том, чтобы дать возможность МСП получить доступ к большим базам данных.

Подход Израиля к решению проблем

Израиль осознает потенциальные риски и проблемы, связанные с растущим использованием приложений ИИ, и намерен их решать. В то же время, как отмечается в Отчете о национальной инициативе, опыт показывает, что чрезмерное регулирование может подавлять инновации, особенно когда речь идет о новых технологиях.²³³ Национальная инициатива включала две рабочие группы, посвященные, соответственно, этике и регулированию, а также кибербезопасности систем ИИ. Их основной задачей было создание модели, которая позволила бы сбалансировать необходимость: (1) обеспечить этическую и безопасную разработку и внедрение приложений ИИ в соответствии с ценностями Израиля как демократического государства; (2) способствовать технологическим инновациям, научным исследованиям и разработкам, которые имеют фундаментальное значение для израильской экономики и национальной безопасности. Как будет показано ниже, подход, использованный в отчете о национальной инициативе, является новым с точки зрения сочетания этики, прав человека и инноваций.

231 Ibid. cc.190-191.

232 Ibid.

233 Бен-Исраэль, И., Матания, Е. и Фридман, Л. (ред.) (сентябрь 2020 г.). стр.32.

I. Шесть этических принципов для ИИ

Признавая, что права человека и этические соображения остаются первостепенными, в Докладе по этике перечислены «6 этических принципов для ИИ», которые должны лечь в основу государственной политики.²³⁴

Справедливость: Стремление к существенному равенству, предотвращение предвзятости (в информации, в процессе и в продукте), предотвращение дискриминации и недопущение увеличения социально-экономического и образовательного разрыва.

Подотчетность:

Прозрачность: Предоставление информации о процессе и принятии соответствующих решений.

Объяснимость: Возможность объяснить процесс принятия решений в системе (на уровне отдельных пользователей, а также на коллективном уровне, если система затрагивает группу, а также для самих операторов системы).

Этическая и юридическая ответственность — должна быть распределена между соответствующими участниками цепочки создания стоимости, наряду с управлением рисками. Определение ответственности за установление правил разумных мер по предотвращению риска в соответствии с контекстом и предполагаемой серьезностью риска, за управление рисками и за назначение сотрудника, ответственного за управление рисками.

Защита прав человека:

Телесная неприкосновенность: Предотвращение любого вреда жизни или конечностям.

Конфиденциальность: Предотвращение ущерба частной жизни в результате сбора, анализа и обработки информации, обмена информацией и использования информации в новых и иных целях.

234 Nahon K. et al. (2020). с.182.

Автономия: Сохранение способности человека принимать разумные решения, включая предотвращение несправедливого или неосознанного влияния на поведение человека.

Гражданские и политические права: Включая право избирать, свободу слова и свободу совести религии.

Кибер и информационная безопасность: Поддержание систем в рабочем состоянии, защита информации, которую они используют, и предотвращение злоупотреблений со стороны злоумышленника.

Безопасность: Предотвращение опасности для отдельных лиц и общества и уменьшение любого ущерба.

Внутренняя безопасность: при разработке инструмента ИИ.

Внешняя безопасность: для окружающей среды и клиентов, при использовании инструмента.

Поддержание конкурентного рынка и правил поведения, способствующих конкуренции.

В свете этих руководящих этических принципов Отчет о национальной инициативе предлагает сбалансированную модель регулирования, основанную на применении минимального регулятивного вмешательства, необходимого для поддержания адекватной этической среды, с одной стороны, и воздержания от любых ненужных ограничений инноваций и научного прогресса, с другой стороны. Соответственно, Отчет о национальной инициативе призывает «поощрять саморегулирование путем использования инструментов, разработанных в рамках [национальной] инициативы, для оценки рисков и заблаговременного выявления этических проблем на этапах разработки и производства. Этические ограничения должны быть интегрированы в интеллектуальные системы, должно быть определено запрещенное поведение, а этические принципы должны внедряться в процессе обучения и подготовки тех, кто имеет дело с системами ИИ.»²³⁵

235 Бен-Исраэль, И., Матания, Е. и Фридман, Л. (ред.) (сентябрь 2020 г.). с.31.

II. Сбалансированное регулирование для стимулирования инноваций

Одной из ключевых особенностей Отчета по этике является его подход к регулированию. Вместо того, чтобы излагать перечень видов деятельности, которые должны регулироваться, он использует системный подход к этому вопросу, состоящий из трех этапов: (1) составление карты различных типов подходов к регулированию, а также их соответствующих преимуществ и недостатков; (2) определение основных областей и видов деятельности ИИ, которые могут выиграть от определенного уровня регулирования, и рисков, связанных с каждым из них;

(3) согласование различных подходов к регулированию с различными видами деятельности в области ИИ. Это дает правительству «дорожную карту» для разработки специальных отраслевых нормативных актов.

Отчет по этике выделяет следующие широкие подходы к регулированию:

Законодательство или регулирование

Принятие судебных решений для толкования существующего законодательства или заполнения пробелов

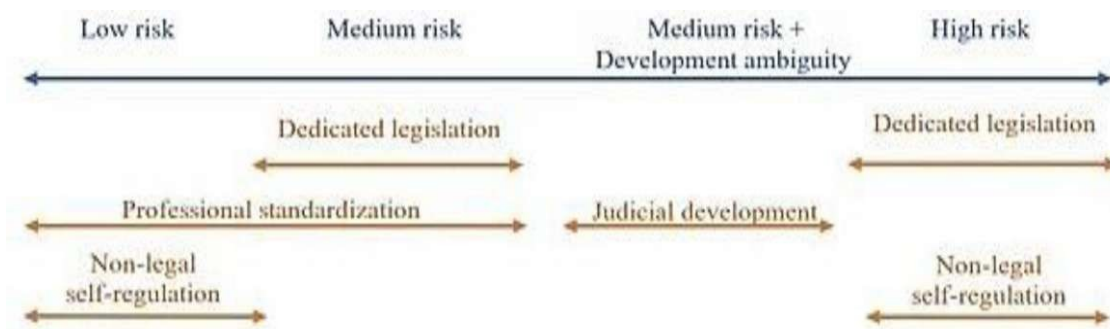
Профессиональные стандарты (правительством, промышленностью, научными кругами или гражданским обществом)

Саморегулирование с помощью этических правил или профессиональных стандартов, обычно разрабатываемых соответствующим профессиональным сообществом.

Далее в отчете освещаются преимущества и недостатки каждого подхода.

Далее в Отчете по этике предлагается следующая модель, позволяющая подобрать различные подходы к регулированию в зависимости от уровня риска, связанного с конкретным видом деятельности:

Так, например, деятельность с высоким уровнем риска лучше решать с помощью законодательства и саморегулирования *ex ante*, а не судебного вмешательства *post facto*. С другой стороны, деятельность с низким уровнем риска не обязательно требует специального законодательства и может быть решена с помощью стандартов и саморегулирования.



Эта модель, конечно, не предназначена для жесткого применения. Скорее, она представляет собой основу, которая позволяет политикам и регуляторам оценить подходящие средства для той или иной деятельности, принимая во внимание множество переменных. Далее отмечается, что вопрос «кто регулирует» не менее важен: регулирование центральным органом ИИ позволяет разработать последовательную политику; однако существует риск чрезмерного регулирования и подавления инноваций, если регулирование будет принято повсеместно. И наоборот, регулирование может быть возложено на различные отраслевые органы, что позволит проводить больше экспериментов в ущерб единообразию правил.

В свете вышесказанного Рабочая группа по этике предлагает 11 принципов регулирования:

Приведение израильского регулирования в соответствие с международным законодательством и стандартизацией, а также продвижение израильской политики на глобальных аренах — речь идет об участии в международной дискуссии о регулировании ИИ, чтобы быть в курсе возникающих международных стандартов и одновременно участвовать в формировании этих стандартов в будущем.

Сопоставление участников для создания адаптированной системы ответственности и стимулов — это требует подхода с участием многих заинтересованных сторон, чтобы дать возможность политикам понять их соответствующие роли в цепочке создания стоимости, их стимулы и ответственность.

Приведение принципа подотчетности в соответствие с динамизмом области ИИ – здесь предлагается требовать от организаций, внедряющих технологии ИИ, создания тестовой среды и контрольных периметров до внедрения технологии.

Здесь предлагается потребовать от организаций, внедряющих технологии ИИ, создать тестовую среду и контрольные периметры до внедрения технологии, чтобы определить, как лучше применить принцип подотчетности в конкретном случае с учетом ожидаемых эффектов технологии.

Содействие нормативной ясности на критических этапах цепочки создания стоимости продуктов ИИ — это подчеркивает важность руководства на ранних этапах разработки ИИ. Инструмент оценки рисков ИИ и, возможно, в некоторых случаях нормативное требование об оценке воздействия ИИ будут полезны для снижения рисков и для того, чтобы разработчики могли реализовать различные этические принципы и правовые нормы.

Постоянный пересмотр регулятором политики регулирования – помимо мониторинга реализации существующего регулирования и обновления правовых текстов, этот принцип требует экспериментов в области регулирования. Он требует от регулирующих органов гибкого и подвижного подхода, поощрения инноваций при учете рисков.

Регуляторные песочницы — концепция регуляторных песочниц хорошо известна. Контролируемое тестирование особенно полезно в контексте ИИ «из-за необходимости разрешить инновации, с одной стороны, и устранить непредсказуемые риски для социальных интересов, с другой».

Взаимосвязь между предлагаемыми принципами и существующими нормативными актами — учитывая, что законы и нормативные акты уже применяются во многих сферах деятельности (здравоохранение, транспорт, финансы, образование и т.д.), при определении необходимости нового регулирования и его направленности и сферы действия необходимо учитывать существующий правовой ландшафт, а также конкретные ценности, интересы и потенциальные социальные выгоды регулирования. На базовом уровне каждый

государственный орган уже несет ответственность за проведение такого анализа в рамках той области, которую он регулирует.

Роль Управления по защите конфиденциальности — Конфиденциальность является межсекторальным вопросом, поэтому Управление по защите конфиденциальности должно играть основную роль в оценке последствий систем ИИ для конфиденциальности и в принятии необходимых нормативных актов в координации с другими государственными органами. Кроме того, важно, чтобы Управление по защите частной жизни получило ресурсы, необходимые для разработки современной правовой и технологической базы в области анонимизации информации, поскольку она является основополагающей для развития ИИ.

Роль антимонопольного органа, как уже упоминалось выше, конкуренция и справедливый рынок — это не только экономика, но и этический вопрос. Таким образом, перед антимонопольным органом должна быть поставлена задача «сформулировать нормативные акты, направленные на поддержание честной конкуренции в области ИИ, защиту потребителей и обеспечение доступности технологий; а также предотвратить перекладывание технологических рисков и затрат на более слабых игроков, находящихся в нижней части цепочки создания стоимости, что является социально неэффективным».

Необходимость межведомственной координации — для обеспечения согласованной разработки политики и регулирования следует внедрить механизм межведомственной координации.

Органы, ответственные за информационные ресурсы.

Органы, отвечающие за информационные ресурсы, используемые для технологий ИИ, «играют ключевую роль в изучении того, подходит ли нормативно-правовая база, которую они применяют, для достижения целей общества.

Поэтому Комитет рекомендует, чтобы органы власти, ответственные за сферы деятельности, затрагиваемые продуктами обработки информации, должны были пройти оценку в свете принципов, подробно описанных выше. В частности, органам власти необходимо изучить, существует ли при внедрении технологий ИИ или их использовании в регулируемых ими областях деятельности необходимость корректировки применимых рамок с целью содействия защите регулируемых интересов.²³⁶

III. Оригинальный инструмент оценки этических рисков

Поскольку израильский подход поощряет саморегулирование, в Отчете по этике подчеркивается ответственность всех, кто вовлечен в ИИ, за то, чтобы оставаться в курсе рисков, связанных с этой динамичной технологией. Чтобы помочь им в выполнении этой ответственной обязанности, рабочая группа по этике и регулированию разработала оригинальный инструмент для лиц, принимающих решения, для оценки этических проблем. Этот инструмент предназначен для того, чтобы специалисты по ИИ могли выявлять этические риски на всех этапах разработки и производства и правильно реагировать на них. Он состоит из двух частей:

Набор предварительных вопросов, которые должны быть адресованы разработчикам продуктов ИИ, чтобы оценить влияние продукта.²³⁷

Каков уровень потенциального индивидуального вреда?

Какова степень потенциального воздействия на восприятие?

Какова степень потенциального ущерба для общества?

Есть ли влияние на распределение общественных ресурсов?

Достаточно ли разнообразна команда разработчиков?

Какова ожидаемая степень ущерба в результате неправильного использования продукта или потери контроля над ним?

236 Ibid. с.205.

237 Ibid. pp.192-193.

Существует ли быстрый способ выявления непрогнозируемых этических сбоек?

Динамическая карта частот, которая помогает найти сложные области с точки зрения применения этических ценностей при разработке системы. На карте представлены шесть этических принципов в сопоставлении с кардинальными этапами процесса разработки. Она показывает частоту возникновения этических проблем в цепочке разработки продукта, выделяя области, где в прошлом были обнаружены неудачи, и предоставляя информацию о частоте их возникновения (образец карты частот см. в Приложении II). Карта основана на оценке реальных прошлых случаев использования систем ИИ, в которых возникали этические проблемы или конфликты, что позволяет повысить осведомленность о тех областях, в которых организации ИИ сталкивались с проблемами в прошлом, и областях, требующих особого внимания со стороны лиц, принимающих решения. Важно отметить, что поскольку карта формируется на основе тестовых примеров, используемых для ее создания, каждая организация должна выбрать набор примеров, которые имеют отношение к разрабатываемому ею продукту. Кроме того, чтобы карта оставалась актуальной, она должна часто обновляться новыми тестовыми примерами. В Отчете по этике подробно объясняется, как организация ИИ может создать и обновлять карту частот, соответствующую ее продукту. Дополнительную информацию см. в Приложении II.

IV. Международная деятельность и сотрудничество

Израиль участвует в международных форумах, посвященных этике ИИ и правам человека. Израильские представители принимали активное участие в разработке Рекомендаций и руководящих принципов ОЭСР по ИИ. Кроме того, Израиль является членом «Цифровых наций» («DN»), объединяющих 10 ведущих мировых цифровых экономик. В 2018 году Израиль принимал ежегодную встречу DN, на которой была принята декларация об ответственном

ИИ.²³⁸ В 2019 году DN также приняла декларацию об управлении данными.²³⁹ Хотя эти декларации не являются юридически обязательными, они отражают приверженность «цифровых наций» соблюдению высоких стандартов прав человека, этики и подотчетности при использовании цифровых технологий. Израиль также сотрудничает с проектом C4IR Всемирного экономического форума в проведении исследовательских проектов в области транспорта и здравоохранения.²⁴⁰

Карта частот показывает частоту возникновения этических проблем на протяжении всей цепочки разработки продукта. Она указывает на области, где в прошлом были обнаружены сбои, и предоставляет информацию о частоте их возникновения. Поскольку частота может меняться с течением времени и появлением новых событий, мы рекомендуем регулярно обновлять карту.

Для создания карты частот мы использовали десять тестовых случаев, отобранных из реальных прошлых дел, которые представляют собой различные проблемы. Карта иллюстрирует все этические принципы, перечисленные в разделе «Этические принципы для ИИ».

Этические этапы процесса разработки

Ниже мы приводим примеры этических проблем, возникающих в процессе разработки, и отслеживаем их по мере развития, чтобы выявить особенно чувствительные этапы разработки. Для этого мы представим типичный процесс разработки ИИ.

238 Общий подход к ответственному использованию искусственного интеллекта, https://fd812d-4234-49d8-8755-ff45ad565157.filesusr.com/ugd/189d02_ef802d92ba5147d2901bde25c6e954a3.pdf, ноябрь 2018, Написанный рабочей группой по искусственному интеллекту, этот рамочный подход был принят на министерском саммите D9 в Израиле в 2018 году.

239 Декларация Data 360, https://fd812d-4234-49d8-8755-ff45ad565157.filesusr.com/ugd/189d02_abce8f2b8cc140e4baeec7dcab7bee97.pdf, Ноябрь 2019 года, Разработанная рабочей группой Data 360, эта общая декларация была представлена на министерском саммите D9 в Уругвае в 2019 году.

240 Израильское управление инноваций. Создание Израильского центра четвертой промышленной революции – Всемирный экономический форум. Получено с сайта <https://innovationisrael.org.il/en/contentpage/establishment-israeli-center-fourth-industrial-revolution-world-economic-forum>.

Определение продукта

Понимание бизнес-потребности или проблемы, которую пытается решить система, и создание научно-исследовательской организации

Сбор данных – определение источников информации внутри и вне организации, которые будут использоваться для создания системы и оценки ее работы

Подготовка продукта

Обработка и фильтрация исходных данных в форму, которая позволит алгоритмам ИИ получать данные и выполнять с ними вычисления

Моделирование — применение алгоритма ИИ к обработке информации в попытке выявить обобщающие закономерности

Интеграция

Оценка точности модели

Соединение компонентов ИИ с остальной частью системы и распространение ее для широкого использования

Управление рынком

Мониторинг производительности, чтобы убедиться, что система работает так, как ожидается

Экосистема — наряду с процессом внутри организации, необходимо также рассмотреть этические соображения, возникающие в связи с тем, что процесс происходит в израильской экосистеме. Интеграция ИИ может повлиять на социально-экономические, нормативные и другие системы, и это необходимо постоянно отслеживать после запуска.

Проанализируйте список тестовых случаев, а также последствия и зарегистрированные события с учетом списка этических ценностей на стр. 8 выше. Заполните таблицу в соответствии с возникающими этическими проблемами. Цифры в ячейках таблицы означают номер события. Затем проверьте накопленное количество событий. Ячейки с низкой, средней и высокой частотой событий окрашены в бежевый, желтый и красный цвета соответственно. Обратите внимание, что данная таблица не указывает на

степень и масштаб потенциального вреда. Более сложный инструмент может учитывать и эти факторы. Решение о том, что считать низкой или высокой частотой, должно приниматься при выборе количества событий, к которым относится организация.

Комитет рекомендует лицам, принимающим решения, обсуждать и предлагать решения возникающих проблем в соответствии с картой частот на протяжении всего процесса разработки. Поскольку карта зависит от списка тестовых случаев, каждой организации необходимо выбрать набор тестовых случаев, относящихся к разрабатываемому продукту, предполагая, что этот набор меняется со временем.

Низкая частота проблемных случаев (единичный случай)

Средняя частота проблемных случаев (два случая)

Высокая частота проблемных случаев (три случая и более)

ГЛАВА II. Управление искусственным интеллектом в Японии

I. Вступление

В процесс исследования и разработки новых технологий и их использования в обществе вовлечены различные субъекты. На каждом этапе исследования, разработки и использования технологии каждый субъект принимает решения и политику для ее социальной реализации на основе оценки ее социального воздействия. Совокупность ролей различных субъектов в вышеупомянутой оценке социального воздействия, принятии решений, разработке политики и реализации можно назвать управлением технологиями. В настоящее время в Японии и за рубежом существуют различные подходы к управлению технологиями искусственного интеллекта (ИИ). Не существует единственно правильного ответа в отношении систем и средств внедрения технологий в общество, и на самом деле в этом вопросе переплетается

комбинация различных факторов, включая технологию, культуру, политику, ценности людей, существующие правовые системы, окружающую среду и экономику. Как следствие, дебаты в рамках этих попыток управления носят разнообразный характер.

Хотя трудно охватить все политические меры и мероприятия, в данном отчете кратко изложены характер управления ИИ и особенности дискуссий в Японии. Поскольку в данном отчете в основном рассматриваются попытки управления и дискуссии до 2018²⁴¹, легко предположить, что содержащиеся в нем детали будут меняться по мере развития технологий и изменения социальных условий в будущем. Однако организация дискуссий в определенный момент времени полезна для будущих обсуждений управления ИИ и для сравнительных исследований с управлением другими передовыми технологиями. Данный отчет организован следующим образом. В главе II «Управление ИИ в Японии» сначала приводится информация о деятельности и отчетах каждого участника в промышленности, научных кругах и правительстве. Затем следует анализ документов и отчетов, опубликованных министерствами и ведомствами за 2016-2017 финансовые годы. Обзор и сравнение того, как министерства позиционируют ИИ и связанные с ним технологии (например, большие данные и IoT), представлены в дополнение к разделу о том, как они воспринимают области их использования и проблемы. В главе III, «Сравнение управления ИИ в Японии и за рубежом», мы обобщаем вопросы и перспективы, которые отсутствуют в обсуждениях управления ИИ в Японии. Наконец, в главе IV мы подводим итоги того, как следует вести дискуссии по вопросам управления ИИ в будущем.

241 Данный доклад основан на частичном переводе главы «Управление ИИ», написанной Арисой Эма и Хидеаки Сирояма в книге «Искусственный интеллект, человек и общество» (Keisho Shobo, 2020) на японском языке. (<https://www.keisoshobo.co.jp/book/b498075.html>). Некоторые японские ссылки заменены английскими версиями и дополнены информацией, полученной автором и коллегами в 2020 году.

II. Управление ИИ в Японии

I. Роль каждого участника

В этой главе обобщены обсуждения вопросов управления ИИ в Японии, проведенные различными субъектами.

В феврале 2017 года Комитет по этике Японского общества искусственного интеллекта (JSAI) опубликовал свои руководящие принципы по этике. Это не этика для технологии ИИ, а скорее этика исследователей ИИ, поскольку в них содержатся рекомендации по поведению, которому должны следовать исследователи, и они включают такие разделы, как «Подотчетность и социальная ответственность» и «Соблюдение законов и правил».²⁴²

Кроме того, Общество обработки информации Японии (IPSI) в январе 2018 года создало экспертную группу SC 42 по искусственному интеллекту для работы с SC 42, подкомитетом, созданным ISO/IEC JTC 1 (совместный технический комитет Международной организации по стандартизации и Международной электротехнической комиссии), и общество распространяет мнения Японии по этому вопросу.²⁴³ Кроме того, в июне 2018 года Совет по академическому продвижению при Японской медицинской ассоциации выпустил доклад под названием «Искусственный интеллект (ИИ) и медицина».²⁴⁴

В японских университетах существует множество исследовательских институтов, в названиях факультетов и кафедр которых присутствуют слова

²⁴² Комитет по этике заявил, что они разработали эти рекомендации, чтобы сначала завоевать доверие исследовательского сообщества, прежде чем рассматривать влияние ИИ на общество и этические перспективы. В «Этических рекомендациях Японского общества искусственного интеллекта» (<http://ai-elsi.org/archives/514>) профессор Ютака Мацуо, тогдашний председатель комитета по этике, сказал: «Профессор Шун Цучия сказал, что: «Широкая общественность обеспокоена тем, что исследователи искусственного интеллекта будут делать с технологией. Поэтому, прежде всего, важно дать понять, что исследователи стремятся создать лучшее общество, и что они не сумасшедшие ученые. Поэтому я хотел бы похвалить JSAI за выпуск таких Этических рекомендаций»

²⁴³ Information Processing Society, International Standardization on Artificial Intelligence Launched, https://www.ipsj.or.jp/release/20180110_itscjnews.html (на японском языке))

²⁴⁴ Искусственный интеллект (ИИ) и медицина», отчет о IXI заседании Совета по академическому продвижению Японской медицинской ассоциации. http://dl.med.or.jp/dl-med/teireikaiken/20180620_3.pdf (на японском языке)

«ИИ» и «интеллект». Например, в Токийском университете есть Исследовательский центр искусственного интеллекта нового поколения, который в 2017²⁴⁵, провел симпозиум «ИИ и общество», а также совместно с CFI в Великобритании организовал «Beneficial AI Tokyo», на который были приглашены представители IEEE, PAI и других соответствующих исследовательских институтов и компаний.²⁴⁶

Японская федерация бизнеса (Keidanren)

Японская федерация бизнеса (Keidanren) также инициирует дискуссии о влиянии ИИ и других передовых технологий. В мае 2018 года федерация публично опубликовала предложение под названием «Создание новых ценностей посредством реализации трех принципов бизнеса для достижения ЦУР – на пути к формулированию стратегического видения ПИС».²⁴⁷

В предложении говорится, что «Япония отстает в использовании данных в бизнесе» и что важно восстановить международную конкурентоспособность путем объединения данных с искусственным интеллектом и другими технологиями. В предложении отмечается концепция «Общества 5.0» в решении социальных проблем, а также содержится призыв развивать бизнес в соответствии с Целями устойчивого развития ООН (ЦУР). В отношении использования данных в предложении отмечается, что «необходимо создать системы, позволяющие различным организациям использовать данные с должным учетом конфиденциальности и кибербезопасности, а также стимулировать инновации в различных областях», и добавляется, что активное участие в нормотворчестве, включая международные правила и стандарты, и развитие человеческих ресурсов является сложной задачей.

245 ИИ и общество, <http://www.aiandsociety.org/>

246 Впоследствии была создана организация Beneficial AI Japan (<http://bai-japan.org/>). Отметим, что Научный центр интеллекта следующего поколения присоединился к Партнерству по ИИ (PAI) в августе 2018 года.

247 Создание новой ценности с помощью трех принципов бизнеса с поддержкой общества 5.0, 15 мая 2018 г., http://www.keidanren.or.jp/policy/2018/042_honbun.html (на японском языке).

На первом заседании Совета по социальным принципам человеко-ориентированного ИИ при кабинете министров 8 мая 2018 года был представлен документ под названием «Формулировка принципов использования ИИ для реализации «Общества 5.0»», представленный компанией²⁴⁸, которая представила информацию о формулировке промышленной версии руководства по разработке ИИ и стратегической концепции его развертывания.²⁴⁹

Национальные исследовательские институты

Исследовательский центр искусственного интеллекта, Национальный институт передовой промышленной науки и технологии

Он был создан в мае 2015 года, и по состоянию на октябрь 2018 года 12 исследовательских групп работают над различными проектами²⁵⁰. Также регулярно проводятся семинары, в том числе дискуссия на тему «ИИ и закон» в декабре 2017 года²⁵¹.

Проект RIKEN Центр передовой разведки

Основанный в 2016 году, проект RIKEN состоит из трех групп, проводящих исследования: Группа исследования общих технологий, Группа исследования технологий, ориентированных на цели, и Группа исследования искусственного интеллекта в обществе. Группа исследований искусственного интеллекта в обществе состоит из таких подгрупп, как Группа информационного законодательства и Группа этики и общества искусственного

248 Keidanren, AI Utilization Strategy For an AI-Ready Society, 19 февраля 2019 года, https://www.keidanren.or.jp/en/policy/2019/013_outline.pdf.

249 Хироаки Китано, Заявление о миссии Keidanren AI Application Principles Task Force, <http://www8.cao.go.jp/cstp/tyousakai/humanai/1kai/sanko1.pdf> (на японском языке).

250 Центр исследований искусственного интеллекта: <https://www.airc.aist.go.jp/teams/>.

251 Центр исследований искусственного интеллекта [19-й семинар по ИИ] «ИИ и закон» https://www.airc.aist.go.jp/seminar_detail/seminar_019.html (на японском языке).

интеллекта, которые проводят не только технические исследования, но и исследования в области гуманитарных и социальных наук.²⁵²

Центр содействия научным исследованиям и разработкам в области ИИ, Национальный институт информационно-коммуникационных технологий

Центр был создан в 2017 году как организация по продвижению открытых инноваций, стратегических исследований и разработок технологий ИИ в области науки об интеллекте²⁵³. Центр делится разнообразными текстовыми и речевыми данными и строит современную испытательную площадку данных ИИ, чтобы ускорить открытые инновации.

Японское агентство науки и технологий (JST), Исследовательский институт науки и технологий для общества (RISTEX)

В 2016 году RISTEX создало направление «Экосистема человек-информационные технологии» (HITE). Целью этого направления является содействие коэволюции технологий и общества путем организации вопросов, возникающих при внедрении в общество новых информационных технологий, таких как ИИ, и создания места и платформы, где такие вопросы могут быть напрямую переданы в те же области исследований и разработок.

Правительственные организации

Офис кабинета министров

В апреле 2016 года был создан Стратегический совет по стратегии ИИ в ответ на директиву премьер-министра в «Диалоге государственного и частного секторов для будущих инвестиций». Совет выполняет функции командного пункта. Помимо трех министерств (Министерство внутренних дел и коммуникаций; Министерство образования, культуры, спорта, науки и

252 Исследовательский центр интеграции инновационного интеллекта (AIP): <https://www.riken.jp/en/research/labs/aip/index.html>.

253 Центр содействия комплексным исследованиям и развитию интеллектуальной науки: <http://www2.nict.go.jp/oihq/en/>.

технологий; Министерство экономики, торговли и промышленности), в настоящее время исследования и разработки и социальное внедрение продвигаются посредством сотрудничества и взаимодействия со смежными министерствами и ведомствами, такими как Канцелярия кабинета министров, Министерство здравоохранения, труда и благосостояния, Министерство земли, инфраструктуры, транспорта и туризма, а также Министерство сельского, лесного и рыбного хозяйства. Концепция «Общество 5.0» и «сверхумное общество» были предложены в качестве общественных благ, которые должны быть реализованы.

В марте 2017 года была сформулирована стратегия развития технологий искусственного интеллекта, а также создана отраслевая дорожная карта. Согласно дорожной карте, национальные агентства должны работать над (1) исследованиями и разработками, (2) развитием человеческих ресурсов, (3) развитием среды для данных и инструментов, (4) поддержкой стартапов и (5) продвижением понимания технологии ИИ. Проект плана реализации стратегии стратегии развития технологий искусственного интеллекта был составлен в августе 2018 года.²⁵⁴

В 2016 финансовом году канцелярия кабинета министров учредила Консультативный совет по искусственному интеллекту и развитию общества, который объединил различные вопросы в единый доклад²⁵⁵, а Совет по социальным принципам человекоцентричного ИИ, который был создан в канцелярии кабинета министров в мае 2018 года, подготовил проект резюме социальных принципов человекоцентричного ИИ в конце 2018 года²⁵⁶. В марте 2019 года Социальные принципы человекоцентричного ИИ были опубликованы

254 Краткая информация о Стратегическом совете по технологиям ИИ представлена на портале ИИ NEDO: http://www.nedo.go.jp/activities/ZZJP2_100064.html (на японском языке).

255 Офис кабинета министров, «Отчет об искусственном интеллекте и человеческом обществе» https://www8.cao.go.jp/cstp/tyousakai/ai/summary/aisociety_en.pdf.

256 Офис кабинета министров, Совет по социальным принципам человеко-ориентированного ИИ, <https://www8.cao.go.jp/cstp/stmain/aisocialprinciples.pdf>.

Секретариатом кабинета министров в феврале 2019 года²⁵⁷, а в июне 2019 года, когда Япония принимала G20, эти принципы были представлены.

Министерство внутренних дел и коммуникаций (МВК)

Институт информационно-коммуникационной политики (ПСР) Министерства внутренних дел и коммуникаций — это исследовательская организация, занимающаяся изучением информационно-коммуникационной политики и проводящая исследования в области ИИ. В октябре 2016 года ПСР учредил Конференцию по сетевому обществу ИИ с целью изучения социальных, экономических, этических и правовых вопросов для продвижения сетевого общества ИИ. В этой конференции принимают участие заинтересованные стороны из промышленности и других секторов, а также эксперты в области науки и техники, гуманитарных и социальных наук. В рамках конференции были созданы Комитет по принципам НИОКР в области ИИ и Комитет по оценке воздействия и рисков.²⁵⁸ В июле 2017 года Конференция выпустила Руководство по НИОКР в области ИИ.²⁵⁹ Эти руководящие принципы были выпущены с целью внести вклад в международные дискуссии, и руководящие принципы были представлены на G7 и других конференциях. Конференция ПСР также выпустила проект Руководства по использованию ИИ, который был опубликован в качестве Руководства по использованию ИИ в 2019 году.²⁶⁰

Министерство образования, культуры, спорта, науки и технологий (МEXT)

Совещание при министре по развитию ресурсов для «Общества 5.0»²⁶¹ было создано в декабре 2017 года и заседало до июня 2018 года, его деятельность была сосредоточена на развитии человеческих ресурсов, а параллельно проводились обсуждения в целевой группе министерства по

257 Секретариат Кабинета министров, Социальные принципы человеко-ориентированного ИИ, 2019, <https://ai.bsa.org/wp-content/uploads/2019/09/humancentricai.pdf>.

258 Сирояма, Хидеаки, «Искусственный интеллект и оценка технологий: рамки, система и экспериментальные испытания», Журнал «Наука, техника и общество», 16, 2018. (на японском языке)

259 Конференция ИИ, 2017, https://www.soumu.go.jp/main_content/000507517.pdf.

260 Конференция по сетевому обществу ИИ, Руководство по использованию ИИ, 2019, https://www.soumu.go.jp/main_content/000658284.pdf.

261 МEXT, Встреча министров по развитию человеческих ресурсов для «Общества 5.0», https://www.mext.go.jp/b_menu/activity/detail/pdf2018/20180605_001.pdf.

воспитанию способности жить богато в новую эпоху. В докладе, выпущенном в июне 2018 года, в первом же разделе перечислены направления политики, которые необходимо рассмотреть в обществе, где развиваются технологии ИИ, такие как предоставление разнообразных возможностей и мест обучения для реализации «справедливого и индивидуально оптимизированного обучения», приобретение базовых академических навыков и умения пользоваться информацией, а также преодоление разрыва между гуманитарными и естественными науками.

Министерство экономики, торговли и промышленности (METI)

В январе 2016 года Министерство экономики, торговли и промышленности учредило Исследовательскую группу по идеальным подходам к политике конкуренции в условиях четвертой промышленной революции для изучения и рассмотрения будущих путей решения текущей ситуации и проблем, связанных с межсекторными системами, такими как политика конкуренции и политика интеллектуальной собственности, с учетом возможности быстрых изменений в промышленных структурах, которые произойдут в результате четвертой промышленной революции, включая развитие искусственного интеллекта. Исследование проводилось в течение июля, после чего был составлен отчет²⁶².

Комитет по новой промышленной структуре Совета по промышленной структуре Министерства экономики, торговли и промышленности²⁶³ проводит свои заседания с сентября 2015 года и сформулировал видение, которое могут разделить государственный и частный секторы. Кроме того, комитет обсудил меры, которые потребуются от государственного и частного секторов для точного реагирования на изменения, вызванные IoT, большими данными и искусственным интеллектом. В мае 2017 года был опубликован окончательный отчет о новой промышленной структуре, который был переименован в

262 METI, Исследовательская группа по изучению идеальных подходов к политике конкуренции для четвертой промышленной революции составляет отчет: https://www.meti.go.jp/english/press/2017/0628_001.html

263 METI, Совет по промышленной структуре, https://www.meti.go.jp/english/policy/economy/industrial_council/index.html.

«Будущее видение на 2030-е годы»,²⁶⁴ и были определены четыре стратегические области (мобильность, цепочки поставок, здравоохранение и жизнь) для достижения «Общества 5.0» как идеального социального видения, к которому необходимо стремиться. Кроме того, в отчете представлена дорожная карта и прорывные проекты.

Кроме того, в декабре 2017 года Отдел информационной экономики Бюро коммерческой и информационной политики начал работу исследовательской группы по разработке руководящих принципов заключения контрактов на использование ИИ и данных, а в июне 2018 года были выпущены «Руководящие принципы заключения контрактов на использование ИИ и данных».²⁶⁵ Руководство состоит из главы, посвященной данным, и главы, посвященной ИИ, в которых в качестве справочного материала для частных предприятий кратко изложены основные вопросы договоров, образцы положений договоров и факторы, которые необходимо учитывать при составлении положений, при заключении договоров, связанных с использованием данных, или договоров, связанных с разработкой и использованием программного обеспечения, использующего технологию ИИ.

Министерство здравоохранения, труда и благосостояния (МЗТБ)

В феврале 2015 года Министерство здравоохранения, труда и благосостояния начало совещания по выработке концепции политики в области здравоохранения на следующие 20 лет (2035 год) на конференции за круглым столом «Видение Японии; здравоохранение 2035».²⁶⁶ После восьми раундов обсуждений в июне 2015 года были опубликованы их предложения, в которых они изложили три основные ценности и критерия: справедливость и равенство,

264 METI, был составлен окончательный отчет о концепции новой промышленной структуры: https://www.meti.go.jp/english/press/2017/0530_003.html.

265 Министерство экономики, торговли и промышленности, Версия перевода на английский язык Руководства по заключению контрактов на использование искусственного интеллекта и данных, 2019, https://www.meti.go.jp/english/press/2019/0404_001.html.

266 Медицинское страхование 2035, краткое изложение рекомендаций: <https://www.mhlw.go.jp/seisakunitsuite/bunya/hokabunya/shakaihoshou/hokeniryou2035/future/>.

солидарность на основе автономии, а также процветание и сосуществование для Японии и всего мира.²⁶⁷

Впоследствии, в январе 2017 года началось заседание Совета по продвижению использования ИИ в страховании и медицине, и в июне 2017 года они составили отчет, в котором были выбраны шесть приоритетных областей для продвижения развития ИИ. Впоследствии, в июле 2018 года был проведен Консорциум по продвижению развития ИИ в страховании и медицине.²⁶⁸

Кроме того, с января 2016 года проводились совещания «Будущее работы: 2035», посвященные изменениям в стиле работы в связи с технологическими инновациями, в основном в технологии искусственного интеллекта, а в августе 2016 года был выпущен отчет «Будущее работы: 2035».²⁶⁹ Кроме того, в марте 2017 года компания Mitsubishi UFJ Research and Consulting по заказу Министерства здравоохранения, труда и благосостояния опубликовала отчет об исследовании влияния IoT, больших данных и искусственного интеллекта на занятость и трудовые ресурсы.²⁷⁰

Министерство земли, инфраструктуры, транспорта и туризма (MLIT)

Министерство земли, инфраструктуры, транспорта и туризма в основном проводит дискуссии по автоматизированному вождению, и в сентябре 2018 года млИТ сформулировало Руководство по безопасности автоматизированной эксплуатации сельскохозяйственных машин, в котором разъясняются требования безопасности, которым должны соответствовать автоматизированные транспортные средства уровня 3 и 4.²⁷¹ В апреле 2018 года они также объявили об исследовании «Концепции центра ИИ», который будет оценивать и сертифицировать предоставленные данные и ИИ, необходимые для

267 Медицинское страхование 2035, краткое изложение рекомендаций:

<https://www.mhlw.go.jp/seisakunitsuite/bunya/hokabunya/shakaihoshou/hokeniryou2035/future/>

268 MHLW, Консорциум по продвижению развития ИИ в страховании и медицине, https://www.mhlw.go.jp/stf/shingi2/0000148680_00002.html (на японском языке).

269 <https://www.mhlw.go.jp/file/06-Seisakujouhou-12600000-Seisakutoukatsukan/0000152705.pdf>.

270 <https://www.mhlw.go.jp/file/04-Houdouhappyou-11602000-Shokugyouanteikyoku-Koyouseisakuka/0000166533.pdf> (in Japanese).

271 MLIT, Разработка руководящих принципов технологии безопасности для самоуправляемых транспортных средств, 12 сентября 2018 года, http://www.mlit.go.jp/report/press/jidosha07_hh_000281.html.

развития мостов, тоннелей и т.д., чтобы использовать ИИ для инспекции инфраструктуры.²⁷²

Министерство сельского, лесного и рыбного хозяйства (MAFF)

В 2013 году Министерство сельского, лесного и рыбного хозяйства учредило Исследовательскую группу по продвижению «умного» сельского хозяйства в сотрудничестве с компаниями, занимающимися разработкой роботов и информационных технологий, производителями сельскохозяйственной техники, а также соответствующими министерствами и ведомствами. Эта исследовательская группа поставила перед собой задачу изучить меры и рекомендации по продвижению использования искусственного интеллекта и роботов в сельском хозяйстве (умное сельское хозяйство).²⁷³ В 2017 году исследовательская группа сформулировала Руководство по безопасности автоматизированных операций сельскохозяйственной техники, в котором определены требования к производителям и пользователям для обеспечения безопасности технологий беспилотного автоматического управления сельскохозяйственной техникой (роботизированной сельскохозяйственной техникой). В 2018 году в руководство была добавлена сводная таблица, касающаяся источников опасности, а также опасных условий для автоматизации роботов для управления чайными плантациями.²⁷⁴

II. Обсуждение ИИ в министерствах и ведомствах

Цель

Различные субъекты в Японии обсуждают эксперименты по управлению ИИ. Поэтому данный раздел посвящен сравнительному исследованию

272 Nikkan Kogyo Shimbun, AI Centralized Management of Infrastructure Inspections, Министерство земли, инфраструктуры и транспорта рассматривает «Концепцию центра AI», 26 апреля 2018 года, <https://www.nikkan.co.jp/articles/view/00471269>.

273 MAFF, Исследовательская группа по продвижению умного сельского хозяйства, https://www.maff.go.jp/e/policies/tech_res/smaagri/attach/pdf/robot-1.pdf.

274 MAFF, Руководство по обеспечению безопасности автоматического управления сельскохозяйственной техникой, март 2018 г., http://www.maff.go.jp/j/kanbo/kihyo03/gityo/g_smart_nougyo/attach/pdf/index-6.pdf (на японском языке).

дискуссий, которые в 2017 году появились в белых документах и отчетах министерств, разрабатывающих политику в области ИИ. Это делается для того, чтобы понять текущую ситуацию с многогранной точки зрения, не характерной для какого-либо одного министерства. В частности, мы рассмотрим: (1) совпадение ссылок на технологии, связанные с ИИ, в каждом министерстве и ведомстве; (2) какие области применения, преимущества, проблемы и меры упоминаются в каждом министерстве и ведомстве в связи с ИИ; и (3) какая связь существует между социальными концепциями, такими как «Общество 5.0» и четвертая промышленная революция, и технологиями, связанными с ИИ.

Цели опроса

Японские министерства и ведомства опубликовали различные белые книги и отчеты²⁷⁵. Среди них в исследование были включены следующие 19 документов, которые, как казалось, могли содержать обсуждения ИИ. В данное исследование были включены только те версии, которые были доступны по состоянию на декабрь 2017 года.

Стратегия роста 2017 года (Штаб по оживлению экономики Японии, Секретариат Кабинета министров)

Отчет о приоритетных мерах по укреплению конкурентоспособности промышленности за 2016 год (Штаб-квартира по оживлению экономики Японии, Секретариат Кабинета министров)

Ежегодный отчет об экономике и государственных финансах Японии (Канцелярия кабинета министров)

Белая книга о полиции (Национальное полицейское агентство)

Годовой отчет Агентства финансовых услуг (FSA)

Белая книга по информации и коммуникациям (Министерство внутренних дел и коммуникаций)

Основные моменты проекта бюджета на 2018 финансовый год (Министерство финансов)

²⁷⁵ На сайте офиса премьер-министра представлен список белых документов по годам (<https://www.kantei.go.jp/jp/hakusyo/index.html>).

Белая книга по науке и технологиям (Министерство образования, культуры, спорта, науки и технологий)

Белая книга по образованию, культуре, спорту, науке и технологиям (МEXT)

Ежегодный отчет по здравоохранению, труду и благосостоянию (Министерство здравоохранения, труда и благосостояния)

Белая книга по экономике труда (Министерство здравоохранения, труда и благосостояния)

Ежегодный отчет по продовольствию, сельскому хозяйству и сельским районам (Министерство сельского, лесного и рыбного хозяйства)

Белая книга по международной экономике и торговле (Министерство экономики, торговли и промышленности)

Ежегодный отчет по транспортной политике (Министерство земли, инфраструктуры, транспорта и туризма)

Белая книга по вопросам земли, инфраструктуры, транспорта и туризма (Министерство земли, инфраструктуры, транспорта и туризма)

Ежегодный отчет о текущем состоянии метеорологической службы (Японское метеорологическое агентство)

Ежегодный отчет по окружающей среде (Министерство окружающей среды)

Оборона Японии (Министерство обороны)

Белая книга по делам потребителей (Агентство по делам потребителей)

Методы исследования

Выбор слов

Следующие 13 слов (на японском языке) были найдены и извлечены в качестве возможных ключевых слов в текущих дебатах по информационным технологиям.

Искусственный интеллект, ИИ

Большие данные

IoT

Робот

Беспилотники, беспилотные летательные аппараты и малые беспилотные летательные аппараты

Автоматическое вождение и автоматизированное управление

VR, AR и MR (VR/AR/MR)

«Общество 5.0»

Четвертая промышленная революция

Подключенные отрасли (Connected Industries)

Искусственный интеллект общего назначения (AGI)

Сингулярность

Квант

Подсчет слов

Мы искали слова в 19 исследованных документах и подсчитывали их встречаемость.

Структурирование

Описания вокруг слов были извлечены и помечены под одним из следующих четырех пунктов в соответствии с их содержанием.

Позиционирование: раздел, в котором описывается определение и позиция выбранного слова, а также объяснение ситуации, которую необходимо решить.

Области применения и преимущества: раздел, в котором выбранное слово описывает область применения или преимущества

Проблемы и меры: раздел, в котором выбранное слово описывает проблему, включая политику и меры по решению этой проблемы.

Вызовы: раздел, в котором выбранное слово описывает только вызовы

В некоторых белых книгах и отчетах активно используются определенные ключевые слова. Например, «Общество 5.0» часто используется в 12-1 «Белой книге по науке и технологиям», «робот» в 14-1 «Ежегодном отчете по продовольствию, сельскому хозяйству и сельским районам», «четвертая промышленная революция» в 15-1 «Белой книге по международной экономике и торговле», а «автоматическое вождение/автоматизированное вождение», «беспилотники» и «большие данные» в 16-2 «Белой книге по земле,

инфраструктуре, транспорту и туризму». Кроме того, хотя слова, использованные в данном опросе, не были найдены в 7-1 «Годовой отчет Агентства финансовых услуг», кибербезопасность и финтех упоминаются в главе 22 «Другие вопросы», раздел 1 «Международные шаги по решению новых проблем».

Межведомственное сравнение по каждому ключевому слову

Искомые слова были разделены на следующие четыре группы: (a) позиционирование, (b) области применения и преимущества, (c) вызовы и меры, и (d) вызовы, для трех конкретных технологических областей: (I) искусственный интеллект, (II) другие технологические ключевые слова, и (III) «Общество 5.0»/четвертая промышленная революция/подключенные отрасли. В таблице 2 (Ссылки на искусственный интеллект/AI) перечислены основные метки подпунктов, использованные в (b) областях применения и преимуществах, (c) проблемах и мерах, и (d) вызовах. Обратите внимание, что графа «1» относится не к количеству упоминаний, а скорее к количеству раз, когда они упоминаются.

Искусственный интеллект/ИИ

Как показано в таблице 2, «искусственный интеллект/ИИ» активно используется в 1-3 «Стратегия роста 2017», 3-1 «Ежегодный отчет о японской экономике и государственных финансах» и 9-1 «Белая книга по информации и коммуникациям». И, наоборот, в «Белой книге о полиции» (5-1), «Ежегодном отчете Агентства финансовых услуг» (7-1), «Основных моментах проекта бюджета на 2018 финансовый год» (10-6), «Белой книге об образовании, культуре, спорте, науке и технике» (12-2), «Ежегодном отчете о транспортной политике» (16-1) и «Белой книге о земле, инфраструктуре, транспорте и туризме» (16-2) термин ИИ почти не упоминается. В этих белых книгах используются другие ключевые слова, кроме 7-1 «Ежегодный отчет Агентства финансовых услуг», в котором не использовано ни одно из ключевых слов. Например, в документе 16-2 «Белая книга по земле, инфраструктуре,

транспорту и туризму» используются такие слова, как автоматическое вождение, беспилотники и большие данные.

Позиционирование

ИИ обсуждался как предмет, который должен быть использован для реализации «Общества 5.0» и четвертой промышленной революции. Характерно, что при этом рассматривается не только ИИ, но и большие данные и другие технологии.

Используя ИИ и IoT и создавая инновации (опущено), мы должны работать в направлении будущего видения сверхумного общества («Общество 5.0»). (1-3 «Стратегия роста 2017» стр. 90-91).

Новые технологии четвертой промышленной революции, такие как IoT и AI, можно рассматривать как продолжение цифровой экономики (3-1 «Ежегодный отчет об экономике и государственных финансах Японии», стр. 149).

Использование больших данных является ключом к реализации экономического роста на основе данных и социальных преобразований. Интернет вещей (IoT) является средством сбора больших данных, а искусственный интеллект (AI) – средством анализа и использования больших данных. (9-1 «Белая книга по информации и коммуникациям», стр. 52).

Нарастание четвертой промышленной революции во всех отраслях промышленности, вызванное такими технологическими инновациями, как IoT, большие данные, искусственный интеллект и робототехника, приносит признаки перемен во все уголки общества в виде инновационных продуктов и услуг, которые выходят за пределы национальных границ. (15-1 «Белая книга по международной экономике и торговле», стр. 209).

Четвертая промышленная революция, которая использует технологические прорывы в области Интернета вещей (IoT), больших данных, искусственного интеллекта, роботов и датчиков, как ожидается, создаст новые предприятия, которые решат социальные проблемы и пробудят скрытые

потребности различных заинтересованных сторон. (17-1 «Ежегодный отчет о текущем состоянии метеорологической службы», стр. 21).

Для того чтобы японская экономика росла, важно повышать добавленную стоимость и содействовать усилиям по устранению ограничений предложения. Для этого необходимо реагировать на такие инновации, как IoT и ИИ в рамках четвертой промышленной революции, увеличивать добавленную стоимость, производимую на одного работника, другими словами, производительность труда, и реагировать на изменения в среде, окружающей стили работы, такие как участие женщин и пожилых людей в рабочей силе. (13-2 «Белая книга по экономике труда»).

Области применения и преимущества

Медицина, например, рассматривается в «Ежегодный отчет о здоровье, труде и благосостоянии». С другой стороны, «Стратегия роста 2017», «Ежегодный отчет о японской экономике и государственных финансах» и «Белая книга по информации и коммуникациям» приводят множество примеров, не останавливаясь на каком-либо конкретном секторе.

Кроме того, в этих отчетах отмечаются как преимущества изменения, которые были выявлены с помощью сценариев использования в «Бизнесе/коммерциализации», а также в «Труде/стиле работы».

Реализация «Общества 5.0» приведет к расширению способов работы, не ограниченных временем и пространством. В дополнение к возможностям людей развивать свои собственные способности и внедрять свои собственные стили работы через сотрудничество с ИИ, роботами и другими машинами, мы также можем увидеть «более умные способы работы» через продвинутую телеработу, использующую виртуальную реальность, дополненную реальность и другие формы ИКТ. (3-1 Ежегодный отчет об экономике и государственных финансах Японии).

Вопреки опасениям, что ИИ и роботы отнимут рабочие места, результаты анализа показывают, что, несмотря на возможное воздействие на некоторые

профессии и квалифицированных работников, если будет создано много новых товаров и услуг, заменяющих существующие товары, то рост производительности за счет этих продуктовых инноваций может привести к увеличению занятости. (Ежегодный отчет об экономике и государственных финансах Японии).

В части II данной Белой книги также отмечается, что использование технологических достижений путем осуществления инновационной деятельности, включая искусственный интеллект, может способствовать достижению баланса между трудовой деятельностью и личной жизнью путем создания гибких стилей работы, таких как работа не по найму, кроме того, отмечается, что достижение баланса между трудовой деятельностью и личной жизнью может привести к эффективному использованию человеческих ресурсов и создать добродетельный круг инновационной деятельности.

Четвертая промышленная революция приводит к серьезным изменениям в характере занятости. Четвертая промышленная революция (1) сделала возможным свободный обмен всеми видами бизнеса и информацией в реальном мире через данные и сети, и (2) большие объемы данных могут быть проанализированы и использованы таким образом, чтобы создать новую стоимость. Кроме того, (3) искусственный интеллект сделал возможным обучение машин и принятие ими сложных решений, превосходящих человеческие, и (4) стала возможной автоматизация разнообразных и сложных задач. Эти технологические инновации позволили реализовать общество, которое ранее считалось неосуществимым, и дали возможность кардинальных изменений в структуре промышленности и занятости.

Вызовы и меры

Проблемы, которые необходимо решить, и меры, которые необходимо принять для получения выгод, перечисленных в пункте (b), включают вопросы, связанные с развитием технологий и исследований, человеческих ресурсов и интеллектуальной собственности. Также были приведены различные примеры

создания механизмов для международной/зарубежной экспансии и сотрудничества.

Как и в пункте (б), «Стратегия роста 2017» и «Белая книга по информации и коммуникациям» касаются проблем и мер, но, кроме того, «Белая книга по науке и технологиям» уникальна тем, что охватывает различные меры.

В частности, развитие человеческих ресурсов находится в ведении Министерства образования, культуры, спорта, науки и технологий и описывается следующим образом.

Для развития человеческих ресурсов с базовыми знаниями математики и науки о данных, которые необходимы для активизации промышленной деятельности Японии в будущем, мы будем укреплять образование в области математики и науки о данных, без различия гуманитарных или естественных наук, в университетах. Целью этого является развитие человеческих ресурсов с математическими способностями, которые могут анализировать и использовать данные, решать различные социальные проблемы, разрабатывать новые идеи и создавать новые технологии. Кроме того, мы будем обогащать практическое образование, такое как обучение на основе проектов, формируя сеть практического образования посредством сотрудничества между промышленностью и академией, а также содействовать развитию систематических образовательных программ для переобучения работающих взрослых, чтобы усилить функциональность университетов по развитию человеческих ресурсов в области информационных технологий. Кроме того, исходя из растущих ожиданий в отношении инженерного образования, которое играет центральную роль в развитии человеческих ресурсов для четвертой промышленной революции и реализации «Общества 5.0», Министерство образования, культуры, спорта, науки и технологий (MEXT) создало «Комитет по обзору инженерного образования в университетах» с января 2017 года. Комитет обсуждает идеальную систему образования и учебные программы бакалавриата и магистратуры в области инженерного образования, а также

идеальный подход к совместному образованию между промышленностью и университетом, что необходимо для реформирования системы инженерного образования, чтобы она могла гибко реагировать на изменения в структуре промышленности и академических кругов.

Хотя в «Стратегии роста на 2017 год» и «Белой книге по информации и коммуникациям» и «Ежегодном отчете о здоровье, труде и благосостоянии» упоминаются «законы», «учреждения» и «правила», чтобы представить инициативы по социальной реализации, только в «Белой книге по науке и технологиям» упоминаются «этические, социальные и правовые вопросы».

В недавно созданном центре RIKEN Центре передовой разведки при Министерстве образования, культуры, спорта, науки и технологий ведется сотрудничество со смежными министерствами, компаниями, университетами и исследовательскими институтами для создания инновационных базовых технологий искусственного интеллекта на ближайшие 10 лет. Кроме того, они стремятся к дальнейшему развитию областей, в которых Япония имеет сильные стороны, таких как iPS-клетки и производство, а также содействуют развитию секторов здравоохранения и предотвращения стихийных бедствий и разработке новых технологий. Одновременно с проведением прикладных исследований для решения социальных проблем Японии, таких как инфраструктура, мы также проводим исследования этических, правовых и социальных вопросов, возникающих при распространении технологии искусственного интеллекта.

Кроме того, в документе «Ежегодный отчет о здравоохранении, труде и благосостоянии» говорится в основном о советах и учреждениях.

В январе 2017 года Министерство здравоохранения, труда и благосостояния учредило Целевую группу по продвижению здоровья данных под руководством министра здравоохранения, труда и благосостояния для изучения внедрения инфраструктуры ИКТ, которая органично связывает сектора здравоохранения, медицины и сестринского ухода, с целью ее полномасштабного функционирования в 2020 финансовом году. («Ежегодный отчет по здравоохранению, труду и благосостоянию», с. 153).

Ожидается, что искусственный интеллект (ИИ) позволит создать новые методы диагностики и лечения, создать среду, в которой люди смогут получать самую современную медицинскую помощь в любой точке Японии, а также снизить нагрузку на медицинских и сестринских работников. Поэтому Министерство здравоохранения, труда и благосостояния (MHLW) провело штаб по продвижению реформы здравоохранения для обсуждения необходимых мер по использованию ИИ в области здравоохранения и медицины («Ежегодный отчет о здравоохранении, труде и благосостоянии», с. 163).

В «Ежегодном отчете о продовольствии, сельском хозяйстве и сельской местности» также упоминается, что проект создания будущего сельского хозяйства на основе искусственного интеллекта реализуется с целью ускорения исследований и разработок технологии ИИ в сельскохозяйственном секторе и скорейшего внедрения ее в практическое использование.

Вызовы

Если (с) показывает проблемы и меры, то (d) — это список пунктов, в котором представлены только проблемы. Как видно из таблицы 2, пункт (с) чаще встречается в Стратегии роста 2017 года, а пункт (d) – в 3-1 Ежегодном отчете об экономике и государственных финансах Японии. Это можно объяснить различным характером отчетов.

Например, в «Стратегии роста на 2017 год» для «Человеческих ресурсов» в пункте (с) указано следующее, что указывает на меры по решению проблемы:

Мы рассмотрим конкретные изменения в реформировании системы инженерного образования для развития необходимых человеческих ресурсов на основе информационных технологий для продвижения структурной реформы промышленности, таких как ИИ, IoT и Big Data для создания новых отраслей, а также путем коренного пересмотра вертикальной сегментированной структуры кафедр и факультетов, например, сделав образовательные периоды, такие как 6-летняя интегрированная система для бакалавриата и магистратуры, более

гибкими и позволяющими студентам приобретать суб-специальности в дополнение к их основной дисциплине, к концу этого финансового года. Цель – поэтапное внедрение изменений со следующего финансового года, а полное внедрение – с 2019 финансового года. («Стратегия будущих инвестиций на 2017 год», стр. 92-93).

С другой стороны, «Белая книга по экономической и фискальной политике» фокусируется только на (d) проблемах, как описано ниже. Хотя в ней говорится о необходимости принятия мер, в ней не указывается, какие конкретно.

В США и других развитых странах отмечается явление относительно высокой заработной платы для высокообразованных людей, спрос на которых растет, в связи с изменениями в экономических структурах и другими факторами, такими как переход к информационным технологиям. В Японии процент учащихся, поступающих в высшие учебные заведения, такие как университеты, младшие колледжи и специальные учебные заведения (профессиональные школы), растет, и до сих пор разрыв в заработной плате не увеличился, поскольку спрос на них избыточен. Однако по мере развития технологических инноваций нам необходимо обеспечить себя высокообразованными кадрами, освоившими самые передовые технологии и способными принести пользу. При этом экономическая эффективность образования будет повышена за счет создания системы, которая определяет приоритетность развития необходимых человеческих ресурсов. Это происходит в свете экономических и социальных структурных изменений в Японии, таких как рост спроса на услуги здравоохранения в стране и за рубежом, рост спроса на въездной туризм, а также создание и углубление ИИ и IoT в экономике и обществе. Инвестиции в человеческие ресурсы важны независимо от возраста сотрудников. Важно укреплять потенциал роста экономики в целом путем содействия обучению работающих людей (повторное образование) и поощрения миграции рабочей силы в отрасли с высокой добавленной стоимостью, одновременно улучшая возможности трудоустройства работников

и повышая заработную плату. («Белая книга 3-1 по экономической и фискальной политике» стр. 83).

Что касается «работы/занятости», которая была перечислена в качестве преимущества в пункте (b), «Белая книга по экономике труда» 13-2 также демонстрирует некоторое признание проблем.

Другие ключевые слова, связанные с технологиями

В данном опросе обсуждаются шесть ключевых слов, связанных с ИИ: большие данные, IoT, робототехника, беспилотники, автоматизированное вождение/автоматизированное вождение и VR/AR/MR.

Интересно отметить, что в отчетах о работе с психологической и эмоциональной стороной пользователей и широкой общественности можно найти упоминания о таких понятиях, как чувство безопасности, которые являются общими для «роботов», «автоматизированного вождения» и «беспилотников».

Было отмечено, что сопротивление некоторых людей роботам-ухаживателям является одной из причин, почему учреждения по уходу не смогли внедрить их, и важно добиться понимания общественностью новых систем ухода с использованием роботов в будущем. Если роботы по уходу станут более распространенными, это может изменить отношение людей (см. колонку). В вышеупомянутой «Стратегии новых роботов» правительство поставило цель увеличить процент людей, которые «хотят использовать» и «хотят, чтобы пожилые люди использовали» роботов по уходу до 80%. (Ежегодный отчет о здравоохранении, труде и благосостоянии).

Как показывают опросы общественного мнения и вышеупомянутые усилия по внедрению автоматизированного вождения в обществе для того, чтобы новые технологии и услуги были приняты обществом, необходимо обеспечить техническую безопасность, а также дать пользователям и обществу чувство безопасности с помощью правил и социальных экспериментов. Например, использование небольших беспилотных дронов и других малогабаритных беспилотных транспортных средств в логистическом бизнесе,

а также улучшение характеристик и совершенствование систем, работает на улучшение условий для коммерциализации, одновременно снимая беспокойство общества по поводу безопасности. (Белая книга по земле, инфраструктуре, транспорту и туризму).

Нет упоминания об искусственном интеллекте / сингулярности / квантах.

Искусственный интеллект и сингулярность не были упомянуты ни в одном из докладов и отчетов. Квант был упомянут в «Тенденции в военной науке и технике» в «Обороне Японии».

В последние годы были сделаны новые разработки в области ИКТ. Например, в августе 2016 года Китай запустил спутник под названием Mozi, который является первым в мире спутником для тестирования квантово-кодированной связи. В дополнение к квантово-кодированной связи, новые технологии, такие как искусственный интеллект (ИИ) и анализ больших данных, могут быть применены в военном секторе различными странами в будущем («Оборона Японии»).

Оборона Японии в основном говорит о пользе информационно-коммуникационных технологий (ИКТ) для обороны, а такие технологии, как ИИ и анализ больших данных, также позиционируются как «новые технологии», как и квантовая связь.

«Общество 5.0» / Четвертая промышленная революция / Подключенные отрасли

Определение «Общества 5.0» следующее: «Новое общество, которое является пятым обществом в истории человечества после (1) охотничьего общества, (2) аграрного общества, (3) индустриального общества и (4) информационного общества. Новые ценности и услуги будут создаваться одна за другой, принося изобилие людям, которые являются главными действующими лицами в обществе. 1-3 Стратегия роста 2017 используется как видение будущего японским правительством. Разница между этой стратегией и

четвертой промышленной революцией также указана в том же докладе «Стратегия роста 2017».

Индустрия 4.0 в Германии и Промышленный интернет в США – это в основном попытки использовать IoT для оптимизации производства и управления запасами в обрабатывающей промышленности за рамками отдельных заводов и компаний. В связи с этим Япония должна реализовать «подключенные отрасли», которые выходят за рамки производства и соединяют товары с товарами, вещи с вещами, людей с машинами и системами, людей с технологиями, компании с компаниями в различных отраслях, людей с людьми разных поколений и производителей с потребителями. «Общество 5.0», к которому стремится Япония, — это попытка решить целый ряд социальных проблем путем внедрения передовых технологий во все отрасли промышленности и социальной жизни, а также предоставления необходимых товаров и услуг людям, которые в них нуждаются, когда они в них нуждаются и в том количестве, которое им необходимо.

Определяющей характеристикой «Общества 5.0» является то, что его «еще предстоит увидеть». По этой причине описание (b) областей применения и преимуществ «Общества 5.0» представляет собой обсуждение процесса, «когда «Общество 5.0» станет реальностью».

Реализация «Общества 5.0» расширит способы работы, не ограниченные временем и пространством. В дополнение к возможностям людей развивать свои собственные способности и внедрять свои собственные стили работы через сотрудничество с ИИ, роботами и другими машинами, мы также можем увидеть «более умные способы работы» через продвинутую телеработу, использующую виртуальную реальность, дополненную реальность и другие формы ИКТ. (Ежегодный отчет об экономике и государственных финансах Японии).

По этой причине (в) задачи и меры и (г) проблемы представлены с точки зрения того, как достичь «Общества 5.0», а конкретного обсуждения (б) областей применения и преимуществ практически нет, и это создает

абстрактное обсуждение. В результате, неизбежно, большинство обсуждений «Общества 5.0» касаются (с) задач и мер и (d) проблем. В частности, мы можем отметить следующие ярлыки «исследования и разработки / основные технологии», «инновации», «человеческие ресурсы», «интеллектуальная собственность», «зарубежная экспансия», «институты», «нормативные акты», «сотрудничество», «внедрение», «стандарты», «оценка» и «демонстрация», перечисленные в (1) ярлыках подраздела «Искусственный интеллект / ИИ».

Четвертая промышленная революция часто используется параллельно с «Обществом 5.0», но, как показывает таблица 1, четвертая промышленная революция используется в белых книгах и отчетах чаще, чем «Общество 5.0». В частности, она чаще используется в «Белой книге по информации и коммуникациям» и «Белой книге по международной экономике и торговле». Это объясняется тем, что во многих отчетах «Общество 5.0» описывается как следующее после Четвертой промышленной революции, и поэтому описания Четвертой промышленной революции более распространены.²⁷⁶

В отличие от «Общества 5.0», четвертая промышленная революция рассматривается как событие, которое уже произошло.

Таким образом, фирмы, реально использующие новые технологии четвертой промышленной революции, чувствуют себя более успешными в инновациях продукции и увеличении возможностей продаж, чем в аспектах эффективности производства. («Ежегодный отчет о японской экономике и государственных финансах»).

²⁷⁶ С другой стороны, в документе 12-1 "Белая книга по науке и технологиям" широко используется только термин "Общество 5.0". Кроме того, в документе 10-6 "Основные моменты проекта бюджета на 2018 финансовый год" также используется термин "Общество 5.0". Кроме того, в документе 10-6 "Основные положения проекта бюджета на 2018 финансовый год" также используется термин "Общество 5.0" вместо термина "Четвертая промышленная революция". Термин "Общество 5.0" также часто используется параллельно с термином "сверхумное общество".

III. Сравнение управления ИИ в Японии и за рубежом

I. Сравнение ролей различных субъектов

В этой главе рассматривается, какие отчеты публикуются субъектами в Японии по сравнению с другими странами.

Роль научных обществ и университетов

В Японии Японское общество искусственного интеллекта (JSAI) начало обсуждение вопросов этики на относительно ранней стадии, в 2014 году, и опубликовало этическое руководство. Оно также сотрудничает с IEEE и TFS²⁷⁷, однако деятельность направлена в основном на внутренних членов, и общество не активно взаимодействует с внешним миром. Например, в декабре 2019 года три научных общества, включая JSAI, которые занимаются машинным обучением, выпустили «Заявление о машинном обучении и справедливости» и заявили, что (1) машинное обучение — это не более чем инструмент, помогающий человеку принимать решения, и (2) исследователи машинного обучения стремятся улучшить справедливость в обществе путем изучения возможных применений машинного обучения.²⁷⁸

Кроме того, по сравнению с зарубежными странами, университетские инициативы, исследовательские отчеты и предложения, проводимые в основном университетскими центрами и факультетами, и особенно дискуссии, в том числе с точки зрения гуманитарных и социальных наук, публикуются на английском языке не так часто.²⁷⁹ В настоящее время исследования оставлены на усмотрение отдельных исследователей.

²⁷⁷ Однако следует отметить, что именно из-за "проблемы обложки" начались дебаты, и что осведомленность об этом вопросе несколько отличается от других стран.

²⁷⁸ Это заявление было выпущено в ответ на разжигание ненависти со стороны исследователя ИИ из Токийского университета. <https://arisaema0.wordpress.com/2020/01/08/ai-governance-in-2019/>.

²⁷⁹ «Перспективы искусственного интеллекта/робототехники и работы/занятости», «Научно-технический исследовательский проект» Исследовательского и законодательного справочного бюро (RLRB) Национальной библиотеки диет, включает в себя исследования и технологические тенденции в первой части, вторая часть охватывает влияние на занятость в восьми областях, включая медицинское обслуживание, уход за больными, образование и сельское хозяйство, а третья часть охватывает зарубежные тенденции в области искусственного

Роль отраслевых ассоциаций

Сотрудничество между отраслями имеет важное значение для создания передового опыта, и многие международные отраслевые ассоциации, работающие над передовым опытом, такие как Партнерство по ИИ и Совет индустрии информационных технологий (ITI), базируются в США. Хотя некоторые компании, которым особенно необходимо принять меры, находятся в США, многие из них создают свои собственные правила управления. Исследование Accenture 2018 года показало, что среди компаний, внедривших ИИ (72% опрошенных компаний), 70% провели этическое обучение технических специалистов, а 63% создали комитет по этике для оценки использования ИИ,²⁸⁰ и это свидетельствует о добровольных усилиях, которые требуются от компаний.²⁸¹

В настоящее время рекомендации по ИИ обсуждаются в основном по инициативе крупных технологических компаний в США и Китае, и не так много голосов от начинающих компаний, у которых меньше ресурсов для участия в подобных обсуждениях. Структура отрасли имеет значение и с точки зрения типов бизнеса. В Японии много компаний, работающих по принципу «бизнес для бизнеса» (B2B), по сравнению с компаниями, работающими по принципу «бизнес для потребителя» (B2C). Вся структура выглядит как B2B2B2B2B2...B2C, а цепочка поставок в Японии очень длинная. Поэтому японским компаниям необходимо учитывать подотчетность, надежность и ответственность наряду с этой очень длинной отраслевой цепочкой поставок.

В международные организации, такие как ITI, входят многие крупные японские компании, поэтому ожидается, что в будущем дискуссии в международных рамках будут активизироваться. Кроме того, Японская

интеллекта и занятости, а также развитие, использование и управление человеческими ресурсами. Данное исследование было проведено по заказу университета, а основная часть была написана преподавателями университета. Перевод исследования на английский язык доступен на сайте AIR (<http://sig-air.org/publications/perspectives-on-ai>).

280 Accenture считает, что компании наращивают усилия по этичному и ответственному использованию искусственного интеллекта – SAS, Accenture, Intel и Forbes Insights Latest Survey, 23 октября 2018 года, <https://www.accenture.com/jp-en/company-news-releases-20181023>.

281 В документе IEEE «Ethically Aligned Design, Second Edition» также предлагается необходимость создания комитета по этике или сотрудников, которые могут охватить все с точки зрения ценности, например, главного специалиста по ценности (CVO).

ассоциация глубокого обучения, основанная в 2017 году, является организацией, занимающейся продвижением бизнес-приложений глубокого обучения, а также предлагающей сертификационные экзамены в рамках развития своих кадровых ресурсов. G-тест, предназначенный для кадров для использования в бизнесе (дженералистов), включает вопросы по промышленному применению, праву, этике и текущим социальным проблемам. Кроме того, в июле 2020 года начала работу исследовательская группа по управлению ИИ и его оценке. Исследовательская группа рассматривает рамки и экосистемную сеть, включающую страховые компании, аудиторские компании, а также системы информирования о нарушениях и сторонние комитеты по расследованию несчастных случаев. Исследовательская группа также рассматривает экосистемы снижения рисков ИИ, в которые входят не только крупные технологические компании, но и стартапы.

Есть надежда, что нынешние обсуждения послужат вкладом в этические и социальные перспективы японских венчурных компаний, особенно в плане самоуправления в отраслях.

Роль правительственных организаций

Во многих странах развитие человеческих ресурсов и научных исследований являются важными национальными стратегиями в области ИИ, поэтому задача состоит в том, как создать центр сотрудничества между промышленностью, научными кругами и частным сектором и как стимулировать дискуссии, демонстрируя лидерство.

Следует отметить, что, хотя Япония поощряет исследования, которые не ограничиваются только наукой и технологиями, но также включают социальные и гуманитарные науки, инвестиции в исследовательские институты и развитие человеческих ресурсов со стороны правительства через бюджет Министерства образования, культуры, спорта, науки и технологий ниже, чем в

других странах.²⁸² Вместо этого Институт информационно-коммуникационной политики (ИСПР), который активизировал проведение конференций с 2016 года, собрал вместе широкий круг заинтересованных сторон и сформировал сеть для дискуссий. Сеть состоит из десятков людей из разных областей и отраслей.

Однако в случае Японии существует сильная тенденция напрямую связывать оценку социального воздействия технологий с разработкой стратегии и принятием политических решений в области технологий, а это означает, что дистанция и независимость между ними не соблюдаются. Конференция по сетевому обществу ИИ, учрежденная ИСПР Министерства внутренних дел и коммуникаций (МИС), создала два отдельных подкомитета — Комитет по принципам НИОКР в области ИИ и Комитет по оценке воздействия и рисков, но в обоих комитетах акцент был сделан на подготовке принципов и руководств.

Также было высказано предположение, что обсуждения, направленные на неапонскую аудиторию, и обсуждения, направленные на японскую аудиторию, не связаны должным образом. Что касается международного присутствия, Япония активно участвует во встречах ОЭСР по вопросам ИИ, начиная со встречи G7 в Такамацу в 2016 году, а Министерство внутренних дел и коммуникаций принимало участие в последующих встречах ОЭСР. Принципы ИИ были одним из центральных вопросов на встрече министров цифровой экономики G20 в Японии в 2019 году.

С другой стороны, обсуждения принципов ИИ в Канцелярии кабинета министров и Министерстве внутренних дел и коммуникаций, которые стали основой для дискуссий, направленных на неапонскую аудиторию, не всегда были связаны с обсуждениями в правительственных министерствах и ведомствах, которые проходили в П-1(5), а анализ Белой книги не затрагивает многие этические или социальные вопросы.

282 Общий бюджет правительства на ИИ в общем бюджете на 2018 финансовый год составляет 77 млрд иен, что, как сообщается, меньше 20% от бюджета США (500 млрд иен) и Китая (450 млрд иен), <https://www.japantimes.co.jp/news/2018/02/25/business/tech/japanese-government-spending-ai-less-20-u-s-china/>.

II. Тенденции и проблемы в конкретных технологиях и областях ИИ

Интенсивность обсуждения конкретных технологий ИИ в каждом министерстве

В разделе II данного отчета одновременно рассматривались не только технологии ИИ, но и периферийные технологии и концепции. Таблица 1 показывает, что «искусственный интеллект/ИИ» упоминается во многих документах и отчетах наряду с «IoT» и «Big Data». В частности, эти ключевые слова упоминаются в документах 1-3 «Стратегия роста 2017» и 3-1 «Ежегодный отчет об экономике и государственных финансах Японии» Секретариата Кабинета министров, а также в документе 9-1 «Белая книга по информации и коммуникациям» Министерства внутренних дел и коммуникаций. Институт информационно-коммуникационной политики (ПСР) Министерства внутренних дел и коммуникаций проводит конференцию «Сетевое общество искусственного интеллекта», и выводы и обсуждения этой конференции также попали в «Белую книгу по информации и коммуникациям» Министерства внутренних дел и коммуникаций.

В Стратегическом совете по технологиям искусственного интеллекта, помимо основных трех министерств (Министерство внутренних дел и коммуникаций, Министерство образования, культуры, спорта, науки и технологий и Министерство экономики, торговли и промышленности), исследования и разработки, а также социальное внедрение в настоящее время продвигаются в сотрудничестве и взаимодействии со смежными министерствами и ведомствами, такими как Офис кабинета министров, Министерство здравоохранения, труда и социального обеспечения, Министерство земли, инфраструктуры, транспорта и туризма, а также Министерство сельского, лесного и рыбного хозяйства. Сравнение отчетов и белых книг показывает, что в Белой книге по информации и коммуникациям широко обсуждаются области использования, преимущества, проблемы и меры. В отличие от этого, Белая книга по науке и технологиям Министерства образования, культуры, спорта, науки и технологий сосредоточена на вопросах

и мерах, связанных с технологическими исследованиями и разработками, человеческими ресурсами и сотрудничеством. В то же время в «Белой книге по международной экономике и торговле» Министерства экономики, торговли и промышленности практически не упоминается ИИ, а лишь затрагиваются вопросы и меры, связанные с занятостью, стилем работы и инновациями.

В свою очередь, Белая книга по науке и технологиям характеризуется частыми ссылками на «Общество 5.0», а Белая книга по международной экономике и торговле – частыми ссылками на «Четвертую промышленную революцию». «Общество 5.0» и «четвертая промышленная революция» являются ключевыми словами в описании видения общества, к которому мы должны стремиться, а технология ИИ — это всего лишь одна из технологий, способствующих реализации этого общества. В Белой книге по земле, инфраструктуре, транспорту и туризму активно используются и, следовательно, особо подчеркиваются «автоматическое вождение/автоматизированное вождение», «беспилотники» и «большие данные», а не «искусственный интеллект/ИИ». В Ежегодном отчете по здравоохранению, труду и благосостоянию, и Белой книге по продовольствию, сельскому хозяйству и сельской местности действительно упоминается «искусственный интеллект/ИИ», но чаще всего речь идет о «роботах».

Технология ИИ иногда используется в программном обеспечении некоторых аппаратных средств, таких как самодвижущиеся автомобили, беспилотники и роботы, поэтому нельзя сказать, что они не связаны между собой, но существуют некоторые различия в типах технологий, на которых сосредоточены различные министерства.

Периферийные технологии и области применения ИИ

Таблица 3 показывает, что областями применения и преимуществами периферийных технологий ИИ являются медицина/здоровье/медицинский уход, транспорт, логистика, предотвращение стихийных бедствий и сельское хозяйство. Это те области «Общества 5.0», которые требуют интеграции

кибернетического и физического пространства, и они также считаются сильными сторонами Японии.

С другой стороны, финансовые услуги (финтех), жилищное строительство (умные дома и помощники) и туризм, с точки зрения накопления данных и развития инфраструктуры, будут развиваться менее активно, чем в других странах. Кроме того, технологические разработки, связанные с национальной обороной и безопасностью, обсуждаются во многих дискуссиях между промышленностью, академией и правительством за рубежом, в то время как в Японии они обсуждаются только Министерством обороны. Кроме того, система смертоносного автономного оружия (LAWS), которая обсуждается в Организации Объединенных Наций, редко обсуждается за пределами Министерства иностранных дел в Японии, хотя в 2018 году она упоминалась в парламенте. Отчасти это может быть связано с тем, что Министерство обороны, Министерство иностранных дел и Агентство финансовых услуг не очень вовлечены в соответствующие органы Стратегического совета по технологиям искусственного интеллекта.

И хотя такие слова, как IoT и большие данные, используются наряду с ИИ, практически не упоминаются VR и квантовые технологии, которые, как было сказано, являются развивающимися технологиями. Также не упоминаются такие слова, как сингулярность и искусственный общий интеллект.

Недостаточное понимание текущей ситуации

Для развития управления ИИ необходимы дискуссии, основанные на данных. Поэтому правительства, университеты и исследовательские институты во многих странах оценивают социальное воздействие технологий и создают для этого индикаторы. Все зависит от того, какие именно показатели должны быть созданы, но, по крайней мере, в Японии проблема заключается в том, что не хватает данных для понимания и оценки текущей ситуации, хотя цели уже поставлены.

Однако, как показано в данном отчете, хотя почти все министерства и ведомства обсуждают развитие человеческих ресурсов, например, «Развивать человеческие ресурсы с базовыми знаниями математики и науки о данных», нет данных для понимания текущей ситуации, хотя этот вопрос упоминается как важная цель. Хотя есть оценки, что ИТ-кадры, вероятно, будут в дефиците в будущем, а в Индексе предпринимательской активности ТЕА (2014) указано соотношение предпринимательских кадров, нет собранного списка данных для понимания текущей ситуации.

В «Проекте стратегии ИИ (общий обзор)» Офиса Кабинета министров есть график «Глобальные инвестиции в ИИ, НИОКР и человеческие ресурсы», но данные по «Числу выпускников университетов, обученных анализу данных» относятся к 2008 году, более 10 лет назад.²⁸³ Хотя необходимо изучить, кто создает какие показатели, на базовом информационном уровне создание показателей необходимо для оценки влияния технологии и выявления потенциальных потребностей. Это, безусловно, вопрос будущего. Разработка такой базовой информации также послужит основой для того, чтобы сделать функцию оценки технологий в некоторой степени независимой от разработки стратегии и формирования политики.

III. Как создать форум для обсуждения и его проблемы

Место и роль сотрудничества каждого участника

Примечательно, что промышленность, научные круги и частный сектор в США активно участвуют в проведении дискуссий, и фактически промышленность и научные круги играют ведущую роль в формировании дискуссий, а не под руководством правительства. Кроме того, поскольку IEEE и ISO, которые являются международными организациями, ориентированы на

283 Оригинальный документ из офиса кабинета министров (<https://www.kantei.go.jp/jp/singi/tougou-innovation/dai2/siryo1.pdf>) – Белая книга по информации и коммуникациям Министерства внутренних дел и коммуникаций 2014 года (<http://www.soumu.go.jp/>) (johotsusintokei/whitepaper/en/h26/html/nc134020.html) (на японском языке).

Запад, необходимо рассмотреть, как Япония участвует в международных платформах, и может ли Япония сформировать свою собственную платформу.

В Японии правительство поощряет создание принципов по собственной инициативе. Однако принципы имеют лишь абстрактное существование, и их внедрение в практику требует взаимодействия и сотрудничества между всеми сферами и отраслями. Характер управления технологиями различается в зависимости от сектора²⁸⁴. В Японии министерства и ведомства начали обсуждения с прицелом на реализацию, и, например, Японская медицинская ассоциация подвела итоги собственных обсуждений. Однако существует необходимость в создании форума, на котором отрасли могли бы сотрудничать для формулирования структур и лучших практик управления технологиями, а также для создания и распространения стандартов. В настоящее время консорциумы формируются по отраслям, но необходимо рассмотреть будущее управление, и будет ли оно подобно тому, что существует в Европе или США, где участвуют различные заинтересованные стороны, или в других формах.

Кроме того, управление через сотрудничество между субъектами необходимо не только для каждой области, но и важно для реализации того, какое общество будет построено с помощью технологий. В Японии «Общество 5.0» упоминается в общих чертах, но видение этого общества еще предстоит воплотить в жизнь. Также необходимо создать форум для сотрудничества между участниками как механизм для гибкого рассмотрения вопроса о том, к какому типу общества следует стремиться.

Необходимость дискуссий с участием широкой общественности

Некоммерческие и неправительственные организации играют важную роль в формировании пространства для дискуссий, в которых участвует широкий круг заинтересованных сторон. Однако в Японии не так много некоммерческих организаций по сравнению с другими странами. Японское общество искусственного интеллекта (JSAI) внесло свой вклад в онлайн-дискуссии Общества будущего. Хотя организация онлайн-диалога может быть

284 Сирояма Хидеаки, Наука, техника и политика, глава 7.

сложной в некоторых отношениях, это интересная инициатива как способ охватить различные заинтересованные стороны. В Японии подобные мероприятия, связывающие общественные обсуждения с политическими рекомендациями, все еще рассматриваются как процесс публичных комментариев.

В других областях в Японии в прошлом проводились мероприятия по диалогу с участием граждан для выработки политических рекомендаций по таким вопросам, как генетически модифицированные растения, нанотехнологии и биоразнообразие.²⁸⁵ Что касается робототехники, то одно мероприятие было проведено как «интерактивный общественный комментарий».²⁸⁶ Также необходимо рассмотреть идеальную форму управления, которая предполагает обсуждение между многими заинтересованными сторонами.

IV. Выводы

В данном отчете было указано, какие социальные вопросы решаются различными заинтересованными сторонами в отношении управления ИИ, а также рассмотрены особенности и проблемы в Японии.

Лидеры в области управления и связанные с этим проблемы также меняются в зависимости от страны и сектора. В областях, где институты и правила играют важную роль, таких как здравоохранение и транспорт, как правило, лидирует национальное управление. Поэтому необходимо разрабатывать испытания с использованием регулятивной песочницы. Кроме того, поскольку услуги, предоставляемые GAFA, развиваются в глобальном масштабе, а технологии быстро прогрессируют, необходимо сотрудничать с различными организациями снизу вверх, а не сверху вниз. Поэтому для

285 Тадаши Кобаяши, «Кто думает о науке и технологии?». (2004), Nagoya University Press, Nobuko Kori, et al.: 'From Participating Observations of WWViews Debate Process on Biodiversity' Science Communication, 13: 31-46, 2013.

286 Роботы наступают! Симпозиум «Робот x будущее x видение мечты» – мост между «потребностями в роботах» и «внедрением в общество», http://interactive.pesti.jp/robot/wp/wp-content/uploads/F01_Introduction.pdf.

продвижения дискуссий по управлению ИИ важно создать механизм и форум для постоянного обсуждения того, насколько одни и те же принципы и правила могут применяться на глобальном и локальном уровне (в каждой стране и регионе), какие заинтересованные стороны должны быть вовлечены и кто должен взять на себя инициативу.

Благодарности

Данный отчет основан на переводе главы «Управление ИИ», написанной Арисой Эма и Хидеаки Сирояма в книге «Искусственный интеллект, человечество и общество» (Keisho Shobo, 2020).²⁸⁷ Мы благодарим г-жу Чихиро Саито и г-на Хидеаки Кодзима, г-на Йоичи Иида, г-на Коичи Такаги и д-ра Такаши Эгава за поддержку нашей работы, чтобы она вошла в отчет САНАИ. Мы благодарим г-жу Харуку Мацумото, аспирантку Токийского университета (в то время), за помощь в проведении исследования данного отчета. Данное исследование является частью результатов исследовательского проекта по науке и технологиям Бюро исследований и законодательных справочников (RLRB) за 2017 финансовый год: «Перспективы искусственного интеллекта/робототехники и работы/занятости» и гранта на помощь в научных исследованиях (А): Международное управление новыми ИТ и биотехнологиями – роль обмена информацией и частных субъектов (номер сертификата 18H03620).

Глава III. ИИ-приложения в Мексике. Взгляд изнутри

I. Введение

Цель данной статьи — представить некоторые из наиболее распространенных и расширенных видов использования приложений

²⁸⁷ Искусственный интеллект, человечество и общество (Keisho Shobo, 2020), <https://www.keisoshobo.co.jp/book/b498075.html> (на японском языке).

искусственного интеллекта (ПИИ) в Мексике, а также нормативно-правовую базу, применимую к ПИИ. Мы стремимся с высокой степенью точности представить контекст, в котором работает каждая система ИИ. Контекст, а также краткое описание системы, как мы надеемся, предоставят читателю достаточно информации, чтобы провести сравнение с другими юрисдикциями и пролить свет на сложности, связанные с потенциальным регулированием ИИ-приложений в Мексике и других странах. Основной проблемой в представлении мексиканского случая является отсутствие систематического источника информации об ИИ-приложениях. Существуют три основных источника информации, которые не равны ни по качеству содержащихся данных, ни по актуальности для целей данной статьи. Первый источник информации — это академические работы различных национальных исследовательских центров, ученых в области ИИ, которые работают в различных местных юрисдикциях и учреждениях. Трудность здесь заключается в том, чтобы выбрать адекватную литературу, чтобы указать на уже зрелые ИИ-приложения и отбросить ранние исследовательские проекты или еще не проверенные решения. Вторым источником информации является экосистема инноваторов и стартапов, разбросанных по разным городам на территории Мексики; частные лица и небольшие корпорации, которые появляются с предложениями по повышению деловой ценности оперативных решений, основанных на ИИ-приложениях. Наконец, третий источник информации — это официальный источник, то есть источник федеральных и местных органов власти Мексики, которые публикуют отчеты и объявляют о государственной политике в различных СМИ (включая официальные веб-сайты). Публичная информация, которая равнозначна публичным действиям в пользу инициатив в области искусственного интеллекта. Состояние публичных данных в Мексике далеко от состояния страны открытых данных, что делает сбор соответствующей публичной информации трудоемкой задачей для любого исследователя. Но помимо различных уровней качества информации из этих источников и, соответственно, дискриминации примеров применения ИИ, мы

будем стремиться представить возникающую головоломку, части которой складываются в соответствии с двумя основными стимулами использования приложений ИИ в Мексике. То есть, вместо того чтобы представлять группы приложений по семейству технологий, мы будем исходить из того, что большинство ИИ-приложений являются результатом необходимости решения конкретной социальной и технической проблемы. Это имеет гораздо больше смысла в такой стране, как Мексика, где внедрять и продвигать инновационные идеи вдвое сложнее, чем в других странах, где больше ресурсов и экосистем, объединенных в сеть для принятия новых решений сложных проблем. В то же время Мексика является 11-й по величине экономикой в мире и 2-м торговым партнером (по объему экспорта) на рынке США. Даже перед лицом глубокого неравенства в мексиканском обществе экономические силы продолжают стимулировать и поощрять создание приложений ИИ в частном и социальном секторах. Мы достигли той стадии в Мексике, когда мы можем зрело спросить себя, как мы хотим регулировать ИИ-приложения и в каком направлении. Вы увидите, что на мексиканской арене существует коллективный разум для разработки национальной повестки дня для ИИ; эта повестка дня требует структурирования и руководства со стороны государственных учреждений. В то же время с экономической точки зрения может возникнуть соблазн сказать, что нерегулируемые рынки приложений ИИ лучше оставить в покое. Однако именно в этих формирующихся экосистемах инноваторов в Мексике мы фиксируем общее мнение о необходимости регулирования ИИ-приложений, всегда помня о принципах верховенства закона, прав человека и демократии. Первая часть данной работы будет посвящена представлению ярких примеров ИИ-приложений Мексики (I), после чего мы рассмотрим скудную нормативную базу, включая государственную политику (II). Последний раздел мы оставим для кратких выводов.

Наблюдение за государственными и частными ИИ-приложениями в Мексике

В Мексике использование искусственного интеллекта (ИИ)²⁸⁸ и связанных с ним технологий набирает обороты в некоторых секторах мексиканской экономики и государственных программах. Динамика, вызванная применением искусственного интеллекта в некоторых секторах мексиканской экономики, происходит на фоне недостаточного развития страны в плане расходов на исследования и разработки. Мексика ежегодно тратит 0,3% своего ВВП на инновации и технологии²⁸⁹. В сочетании с тем фактом, что 20% населения Мексики живет менее чем на 5,50 долларов США в день по международным ценам, объяснение того, почему у нас повсеместно распространены инициативы и приложения ИИ, связано с конкретной политикой правительства страны в сочетании с необходимостью поиска инноваций в частном секторе. В связи с этим в Мексике не наблюдается нормального, повсеместного внедрения технологий искусственного интеллекта. Мы наблюдаем, как отдельные отрасли быстро внедряют технологии искусственного интеллекта для конкретных целей и, в то же время, перспективные приложения, которые получают широкое распространение в ближайшие годы. Действительно, существует новое поколение инноваторов, которые применяют уже зрелые технологии ИИ, создавая одну нескоординированную волну продуктов и услуг, которые, тем не менее, способствуют популяризации ИИ в Мексике.²⁹⁰ Вместо того, чтобы

288 На протяжении всего этого документа мы будем исходить из определения ИИ, данного AI HLEG – независимой экспертной группой, которая была создана Европейской комиссией в июне 2018 года. Соответственно, вместо ИИ мы взаимозаменяемо используем понятия ИИ-система или ИИ-применение. См. A definition of AI: Main capabilities and scientific disciplines High-Level Expert Group on Artificial Intelligence, European Commission, April 8, 2019.

289 <https://data.oecd.org/mexico.htm#profile-innovationandtechnology> (accessed on October 31, 2020).

290 Разумеется, мы не намерены перечислять все проекты, связанные с ИИ, которые разрабатываются или внедряются в этой главе. На мексиканской сцене ИИ происходят захватывающие инновации. Для более широкого обзора инициатив в области ИИ см. раздел «Примеры использования ИИ» в разделе «Направление стратегии ИИ в Мексике» ниже; см. «Приложение II» к Отчету об ИИ и экономическом росте.

<https://news.microsoft.com/uploads/prod/sites/41/2018/11/IA-y-Crecimiento-MEXICO.pdf> (доступ получен 31 октября 2020 года). См. также «Доклад о сервисных роботах», подготовленный Мексиканской академией вычислений.

<https://www.researchgate.net/profile/Yasmin-Hernandez2/publication/340634786-Robotica-de-Servicio/links/5e96334a4585150839de623a/Robotica-de-Servicio.pdf> (дата обращения: 31 октября 2020 года). Наконец, существует обширный реестр задокументированных проектов ИИ в журнале *Komputer Sapiens*,

группировать технологии ИИ по типу архитектуры, стоящей за ними, полезнее описать наиболее распространенные и заметные инициативы, используя критерий источника инновации. С одной стороны, мы видим, что федеральные и местные органы власти в Мексике поддерживают внедрение технологий ИИ для решения конкретных государственных задач таким образом, что они оказывают глубокое повседневное влияние на жизнь мексиканских граждан (А). С другой стороны, частные субъекты, будь то хорошо финансируемые (например, банковская система) или микроинноваторы, в настоящее время являются движущей силой для огромного количества услуг Brotherton, основанных на технологии ИИ; продукты и услуги, которые теперь являются повседневными для мексиканцев (В).

ИИ-приложения были признаны полезным инструментом для достижения общественного блага и достижения политических целей во многих секторах мексиканского правительства, как на федеральном, так и на местном уровне. Некоторые ИИ-приложения, которые мы обнаружили в государственном секторе Мексики, сопоставимы с аналогичными усилиями, предпринимаемыми в других юрисдикциях, например, в Службе внутренних доходов, правительственных агентствах, которым поручено предотвращать отмывание денег и деятельность по финансированию терроризма, или в Службе общественного здравоохранения. Существует глобальная тенденция к использованию ИИ-приложений для более эффективного выполнения задач, связанных с обработкой больших массивов данных из различных источников, которые к тому же увеличиваются в объеме за очень короткое время и в различных форматах (Большие данные). Мы также видим, как приложения ИИ пытаются решить социальные проблемы, которые могут возникнуть в результате глобальной пандемии COVID-19. В этом отношении мы можем похвалить усилия мексиканского правительства, которое справляется с последствиями пандемии без помощи других стран. Есть и другие AI-

приложения, которые отвечают специфике мексиканской страны: огромный исторический поток граждан, мигрирующих в США, зарождающаяся отрасль электроэнергетики и повторяющаяся разрушительная сейсмическая активность на больших территориях Мексики.

В течение последних семи лет Служба внутренних доходов Мексики (SAT, сокращенное название на испанском языке) разрабатывает приложение искусственного интеллекта для выявления уклонения от уплаты налогов.²⁹¹ Эта система пытается бороться с одной из форм организованной преступности. Лица, вовлеченные в подставные компании, используют имитацию законных сделок по продаже услуг и товаров. Эти подставные компании производят миллионы квитанций налогоплательщиков, чтобы уклониться от уплаты подоходного налога и налога на добавленную стоимость мексиканскому правительству. В июне 2020 года Министерство финансов сделало публичное заявление, в котором сообщило о предъявлении уголовных обвинений 8 212 налогоплательщикам (как физическим, так и юридическим лицам).²⁹² Сумма мошеннических операций составила примерно 55 083 957 551 086 мексиканских песо (приблизительный эквивалент 2 223 125 505 255 в евро). Приложение ИИ SAT проанализировало более 22 миллионов квитанций и операций налогоплательщиков. Это очень сложное применение ИИ было задокументировано ведущими учеными страны в этой области за несколько лет до этого в исследовательской работе²⁹³: В докладе «White Paper», финансируемом британским посольством, позже будет показано, что приложение ИИ SAT работало путем «выявления нарушений шаблонов в данных, проанализированных с помощью R Studio, языка Python и БД in-memory Redis».²⁹⁴ В последние годы налоговое законодательство Мексики перешло на электронный режим налоговых квитанций. Каждая налоговая

291 <https://www.gob.mx/innovamx/articulos/inteligencia-artificial-131287> (accessed on October 31, 2020).

292 <https://www.gob.mx/shcp/prensa/comunicado-no-054-shcp-mediante-sat-y-pff-anuncia-acciones-en-contra-de-defraudacion-fiscal-a-traves-de-empresas-factureras> (accessed on October 31, 2020).

293 http://omawww.sat.gob.mx/gobmxtransparencia/Paginas/documentos/estudio_opiniones/Evasion_en_IVA_Analisis_de_Rede_s.pdf (accessed on October 31, 2020).

294 TOWARDS AN AI STRATEGY IN MEXICO: Harnessing the AI Revolution, White Paper Report, British Embassy in Mexico through the Prosperity Fund, Oxford Insights and C Minds, June 2018, p. 23. Электронная консультация на сайте <https://www.oxfordinsights.com/mexico> (доступ получен 31 октября 2020 года).

квитанция служит узлом для идентификации многочисленных данных операции. Выпустив почти 12 миллиардов налоговых квитанций только в 2019 году, SAT надеется увеличить внутренний доход ВВП на 3 пункта.²⁹⁵

ИИ для борьбы с «отмыванием» денег

В соответствии с глобальной тенденцией в борьбе с «отмыванием» денег появились скоординированные усилия в разных юрисдикциях. Отмывание денег — особенно чувствительная тема в Мексике, учитывая борьбу с наркокартелями. Но, возможно, в более национальном масштабе мы сегодня также наблюдаем использование ИИ-приложений в борьбе с коррупцией (то есть коррупцией на государственном уровне, организованной из высших правительственных сфер, целью которой является размытие государственного бюджета посредством сложных транзакций). В Министерстве финансов существует подразделение, специализирующееся на финансовой разведке (UIF, сокращенно на испанском языке). Основной задачей UIF является, среди прочего, реализация международных компромиссов мексиканского правительства как постоянного члена Группы разработки финансовых мер борьбы с отмыванием денег (FATF), также известной под своим французским названием Groupe d'action financiers (GAFI).²⁹⁶ Нет более динамичного примера постоянно меняющихся схем, с которыми можно справиться с помощью методов предиктивного ИИ, чем схема отмывания денег. Когда к сложным схемам отмывания денег добавляются операции, придуманные чиновниками для имитации законных государственных закупок, сложность выявления и преследования этих организованных преступлений потребует огромного количества человеческих ресурсов без помощи систем искусственного интеллекта. Разнообразие направлений деятельности, курируемых UIF, в настоящее время включает борьбу с хищением бензина, торговлей людьми, схемами подставных компаний, незаконным оборотом

²⁹⁵ <https://www.eluniversal.com.mx/cartera/sat-va-por-mas-recaudacion-con-inteligencia-artificial>.

²⁹⁶ <https://www.uif.gob.mx/> (accessed on October 31, 2020).

наркотиков и политической коррупции.²⁹⁷ Примечательным в случае этого специализированного подразделения является не только широкое использование методов машинного обучения и больших данных, но и то, что результаты их расследований воспринимаются генеральной прокуратурой как неопровержимые доказательства.²⁹⁸ То есть, Мексика — это случай, когда результатом применения ИИ в уголовном расследовании является признание в качестве доказательства выводов, сделанных системой ИИ.²⁹⁹ Пока что это печально известный прецедент в Латинской Америке.

ИИ для расширения услуг общественного здравоохранения

К концу 2019 года результаты Национального опроса о здоровье и питании показали, что примерно 75,2% взрослых мексиканцев страдают избыточным весом или ожирением, а 10,3% из них также страдают диабетом.³⁰⁰ Эти данные свидетельствуют о серьезной проблеме общественного здравоохранения, затрагивающей не только взрослых, но и детей. Результаты исследования показали, что дети женщин, страдавших ожирением во время беременности, имеют на 1,4 больше шансов иметь избыточный вес в возрасте 3 лет, чем дети женщин с адекватным весом. У женщин, страдавших ожирением до беременности, дети имеют на 4,5 больше шансов получить избыточный вес к 7 годам, чем дети матерей, имевших адекватный вес до беременности.³⁰¹ Эта информация имеет ключевое значение для предотвращения и обращения вспять

297 [https://www.gob.mx/cms/uploads/attachment/file/520516/Comunicado UIF 011 .pdf](https://www.gob.mx/cms/uploads/attachment/file/520516/Comunicado_UIF_011.pdf).

298 В отдельном прецеденте судья постановил, что АИФ будет допущен в качестве истца вместе с Генеральной прокуратурой к судебному преследованию по делу о политической коррупции. См. номер дела FED/FECC/UNAI-CDMX/0000002/2020 на сайте <https://datos.cdmx.gob.mx/explore/dataset/carpetas-de-investigacion-pgj-cdmx/table/?q=FED%2FFECC%2FUNAI-CDMX%2F0000002%2F2020> (доступ получен 31 октября 2020 года).

299 В этом отношении есть точки сопоставления между мексиканским и австралийским случаем. Австралийский центр отчетов и анализа транзакций (Austrac) по данным Ассоциации сертифицированных специалистов по финансовым преступлениям на сайте.

<https://www.acfcs.org/austrac-using-machine-learning-to-better-uncover-interconnected-criminal-groups-improve-aml-alerts/> (доступ получен 31 октября 2020 года).

300 https://ensanut.insp.mx/encuestas/ensanut2018/doctos/informes/ensanut_2018_presentacion_resultados.pdf.

301 Yu Z, Han S, Zhu J, Sun X, Ji C, Guo X. Индекс массы тела до беременности в связи с весом новорожденных и избыточным весом/ожирением потомства: систематический обзор и мета-анализ. PloS one. 2013;8:e61627. DOI.

<https://doi.org/10.1371/journal.pone.0061627> and Weng SF, Redsell SA, Swift JA, Yang M, Glazebrook CP. Systematic review and meta-analyses of risk factors for childhood overweight identifiable during infancy. Arch Dis Child. 2012;97:1019-26. DOI. <https://doi.org/10.1136/archdischild-2012-302263>.

последствий ожирения; известно, что сосредоточенность государственной политики на информации оказывает положительное влияние на выбор в пользу здоровья, который люди делают со временем; и это было особенно верно в случае Мексики.³⁰² Стоимость распространения актуальной информации в режиме реального времени среди всего населения высока, и ИИ-приложения могут помочь решить проблему распространения такого рода информации, что поможет населению принимать более правильные решения о здоровье. MiSalud — это ИИ-приложение, которое доставляет интерактивные сообщения, используя системы обмена сообщениями Facebook и Twitter. Приложение не только дает советы по здоровью, но и взаимодействует с пользователями, создавая напоминания о визитах к врачу и контроле, а также предоставляет в реальном времени информацию о пандемиях или угрозах здоровью, и обладает способностью создавать подталкивания в пользу лучшего выбора здоровья. Пилотная программа стартовала в 2017 году с акцентом на материнское здоровье. MiSalud разослал SMS-сообщения 5 000 женщин с советами по улучшению их здоровья и здоровья их детей. Отмечалось увеличение частоты медицинских консультаций и веса новорожденных, а также снижение числа матерей и детей с избыточным весом. В настоящее время программа распространяется на диабет, ожирение, вакцинацию детей и оценку риска наркомании.³⁰³

Существует несколько возможностей для использования ИИ в Мексике и других странах. Перед лицом глобальной пандемии COVID-19 различные государства в обществе ускорили внедрение нескольких технологий, чтобы сдержать распространение вирусного заболевания и справиться с ним.

В сценарии COVID-19 раннее обнаружение скоплений потенциально инфицированных людей является жизненно важным для сдерживания болезни. В мегаполисе Мехико проживает около 9 миллионов человек. На ранних

302 См. исследование «Ожирение в Мексике», проведенное Национальным институтом общественного здравоохранения, доступное по адресу: <https://www.insp.mx/resources/images/stories/2019/Docs/190213 LaObesidadenMexico.pdf>, стр. 28 (дата обращения: 31 октября 2020 г.).

303 <https://www.gob.mx/misalud/>.

стадиях пандемии COVID-19 местные власти Мехико внедрили две системы помощи с использованием методов искусственного интеллекта. Первая система — это интерактивная система SMS-диагностики. Она работает следующим образом: отправив SMS-сообщение со словом «COVID-19», пользователь получает серию вопросов. Эти вопросы оценивают риск и вероятность заражения человека вирусом. Если риск заражения высок, система подает сигнал тревоги, который используется службами здравоохранения для определения местонахождения человека или предоставления ему дополнительной информации для обращения в Центр здоровья для лечения. Когда заражение определено, эта же система служит для доставки набора для экспресс-диагностики для использования кластером людей, которые взаимодействовали с диагностированным носителем вируса; а также для получения дополнительной экономической и социальной помощи для этого человека и его семьи. Вторая система работает параллельно с первой и предоставляет в режиме реального времени обновленную информацию о наличии мест в больницах и их переполненности в соответствии с GPS-информацией пользователя. Определяя ближайшую и наименее переполненную больницу или медицинское учреждение, эта система максимизирует время реагирования для оказания помощи людям, потенциально инфицированным вирусом, и минимизирует риск дальнейшего заражения. На сегодняшний день первая система — система быстрой диагностики — обслуживает около 20 миллионов пользователей; она была разработана Цифровым агентством общественных инноваций Мехико (DAPI). Вторая система доступна в виде приложения как для Android, так и для iOS. Информация, полученная с помощью обеих систем, обрабатывается в центре Больших Данных для создания ежедневной стратегии для команд эпидемиологов, которые работают над критическими кластерами по всему городу, а также для переоценки поведения эпидемии на территории мегаполиса площадью 1495 км². Параллельно с системами DAPI мы находим COVIDBot, ИИ-агента для WhatsApp, обученного отвечать на вопросы в интерактивном мультимедийном

контенте, создаваемом организациями здравоохранения. ИИ-бот является филантропическим творением мексиканской компании Intevolution. Основная цель COVIDBot – борьба с фальшивыми новостями, связанными с пандемией, которые распространяются в приложении WhatsApp. COVIDBot — это бесплатный сервис, который в режиме реального времени получает информацию из реестров и данных Всемирной организации здравоохранения, Центра по контролю за заболеваниями в США и Министерства здравоохранения Мексики. Ключевая особенность COVIDBot заключается в том, что он может найти ближайшую сертифицированную лабораторию для проведения тестирования на COVID-19.³⁰⁴

Если отбросить контекст глобальной пандемии, использование ИИ в здравоохранении может быть применено к более широкой политике Мексики: политике, направленной на удовлетворение потребностей обездоленных людей. Обслуживание бесправных групп населения в стране с высоким уровнем неравенства представляет собой огромную проблему для любого правительства, находящегося у власти, и необычное поле для использования приложений ИИ для помощи в заботе о тех, кто больше всего нуждается.

ИИ для помощи мексиканским иммигрантам в США

Исторически сложилось так, что мексиканцы мигрируют из Мексики в США в поисках лучших возможностей, будь то сезонная деятельность, связанная с сельским хозяйством, или постоянная иммиграция с целью более длительного пребывания. К 2018 году в США насчитывалось около 11,17 миллиона иммигрантов мексиканского происхождения.³⁰⁵ Мексиканские иммигранты в США сталкиваются с рядом проблем и потребностей. Удовлетворение потребностей мексиканского населения, проживающего в

304 <https://www.infobae.com/tecno/2020/04/01/asi-es-el-chatbot-mexicano-que-lucha-contra-la-desinformacion-sobre-coronavirus/>.

(accessed on October 31, 2020).

305 Даффин Эрин, Число мексиканских иммигрантов в Соединенных Штатах 1850-2018 гг. на сайте <https://www.statista.com/statistics/673350/mexican-immigrants-in-the-united-states/> (дата обращения: 31 октября 2020 г.).

США, также является приоритетом для мексиканского государства. Мексиканское консульство может сделать чрезвычайно мало для разрозненных (и часто скрытых) групп граждан, которые нуждаются в доступе к жизненно важной информации об их правах и юридической помощи. Регистрация новорожденного как мексиканца, возвращение родственника, умершего на чужбине, и выполнение судебных соглашений по алиментам — вот некоторые ситуации, требующие помощи мексиканской дипломатической службы. Наиболее часто запрашиваемым заявлением является восстановление паспорта, поскольку он является основным средством идентификации личности среди мексиканцев в США. Если человек задержан, он имеет право на юридическую помощь мексиканского консульства, и воспользоваться правом на идентификацию по паспорту крайне важно. Использование ботов для взаимодействия с мексиканскими иммигрантами на данный момент является перспективным решением для удовлетворения потребностей миллионов мексиканцев в США. Приложение ИИ под названием «Asistente Virtual SRE-UNAM» [Виртуальный помощник SRE-UNAM] распознает и идентифицирует тип административного запроса иммигранта. Это приложение ИИ генерирует автоматическую запись на прием в соответствии с ответами, полученными во время взаимодействия с пользователем. Виртуальный помощник является совместной разработкой Министерства иностранных дел и Национального автономного университета Мексики (UNAM, сокращенно на испанском языке).³⁰⁶

Инновации, обусловленные общественными интересами, иногда могут возникать в конкретном контексте, который дает возможность решить сложную проблему, и таков случай с миграцией людей. В других случаях сложность идет рука об руку с необходимостью создания новых способов содействия

306 https://www.youtube.com/watch?reload=9&v=1dv-gVKBRBA&ab_channel=UNAMGlobal (accessed on October 31, 2020). Технический отчет о приложении можно найти здесь: <https://cscwcivicttechnologies.files.wordpress.com/2020/10/asistente-virtual-saul-esparza.pdf> (доступ получен 31 октября 2020 года). За командой инноваторов, разработавших приложение, стоит профессор Саип Саваж, она была названа одним из инноваторов в возрасте до 35 лет в Латинской Америке 2018 года по версии MIT Technology Review: <https://www.innovatorsunder35.com/the-list/saiph-savage/> (доступ получен 31 октября 2020 года).

улучшению окружающей среды. Таков случай применения ИИ для выработки электроэнергии в Мексике.

ИИ для получения электроэнергии от ветра

К концу 2019 года ветряные электростанции вырабатывали 15 000 мегаватт энергии на сумму 826 миллионов долларов США. Существует множество проблем, связанных с реализацией Eolic Electricity; среди них – прогнозирование силы ветра для оценки энергии, которая будет поставляться в электрическую систему общественной сети. Проблема изменчивости обычно решается с помощью имитационных моделей и прогнозных симуляций, которые требуют больших вычислительных ресурсов (из-за необходимости большого количества итераций). Использование методов искусственного интеллекта для прогнозирования того, сколько энергии будет доступно для подачи в сеть от ветропарка, значительно минимизировало эту проблему. Техника ИИ для прогнозирования ветровой энергии использует информацию о погоде, собранную в течение нескольких лет, с помощью динамических байесовских сетей (DBN). Результаты применения этого ИИ удовлетворительно сравнивались с результатами прогнозирования с помощью предыдущих методов временных рядов, что указывает на то, что DBN является перспективным инструментом для прогнозирования ветровой энергии.³⁰⁷ Тенденция, которую мы наблюдаем здесь, заключается в использовании подхода машинного обучения, который может быть использован для замены строгой имитационной модели на суррогатную модель (например, с использованием алгоритмов векторной регрессии).³⁰⁸ Суррогатная модель может быть получена за короткое время и с гораздо меньшими вычислительными ресурсами.³⁰⁹ Исследования и применение ИИ

307 Ибаргуэнгойтия-Гонсалес П. Х., Борунда-Пачеко М., Рейес-Бальестерос А., Гарсия-Лопес Уриэль Алехандро, Прогнозирование ветровой энергии с использованием инструментов искусственного интеллекта, *Ingenieria, Investigacion y Tecnologia*, 2018, 19-4: 1-11. DOI. <http://dx.doi.org/10.22201/ii.25940732e.2018.19n4.033>

308 Santamaria-Bonfil G, Reyes-Ballesteros A, Gershenson C, Прогнозирование скорости ветра для ветропарков: A method based on support vector regression, *Renewable Energy*, 85, 2016: 790-809, ISSN 0960-1481, DOI. <https://doi.org/10.1016/j.renene.2015.07.004>.

309 Rosado-Tamariz E, Zuniga-Garcia M A, Santamaria-Bonfil G, Batres R, A machine-learning approach to speed-up simulation towards the design of optimum operating profiles of power plants. В материалах 8-й Международной конференции по информатике, окружающей среде, энергетике и приложениям (IEEA '19).

соответствуют быстрому росту ветряных электростанций на территории Мексики. Ключевым вопросом для повышения конкурентоспособности сбора энергии Eolic Energy является смягчение изменений спроса и изменчивости ветровых турбин.

Экономический рост, связанный с природными условиями, является ключевым фактором для индустрии «зеленой» энергии. Другие природные условия, которые представляют собой бдительный вызов из-за своей разрушительной силы, — это явления, порождающие разрушения и хаос, такие как извержения вулканов или землетрясения, которые пока непредсказуемы. Мексика имеет давнюю историю землетрясений. Большая часть ее территории подвержена сейсмическим движениям. За последние 365 дней произошло 2 215 землетрясений магнитудой 1,5 или выше. ИИ и IoT в сочетании с методами машинного обучения в облачных вычислениях являются частью инструментария, которым располагают мексиканцы для борьбы с последствиями землетрясений и жертвами.

ИИ для предотвращения человеческих жертв от землетрясений

В Мексике с 1991 года действует одна из немногих в мире систем раннего предупреждения землетрясений (СРПЗ) под названием SASMEX, однако активные разработки в этой области начались раньше, в частности, после землетрясения 1985 года.³¹⁰ SASMEX работает с сенсорными станциями вдоль южного тихоокеанского побережья для обнаружения сейсмической активности. При обнаружении сейсмического события включается система оповещения по длинноволновому радио. В Мехико каждая государственная школа имеет приемник и звуковую систему для озвучивания сигнала тревоги. С учетом государственных школ и других объектов, в Мехико и столичном регионе

Association for Computing Machinery, New York, NY, USA: 194–199. DOI. <https://doi.org/10.1145/3323716.3323735>. Как отмечает NB, проект DeepMind компании Alphabet также проводит исследования для решения проблем стабильности и колебаний на ветряных электростанциях, так что Мексика находится в хорошей компании, чтобы мчаться к решению глобальной энергетической проблемы. См. <https://deepmind.com/blog/article/machine-learning-can-boost-value-wind-energy> (доступ получен 31 октября 2020 года).

310 <http://www.cires.org.mx/sasmex.es.php>.

насчитывается около 90 000 приемников. В среднем, системы оповещения обеспечивают 100 секунд предупреждения до прихода сейсмической волны. Эти 100 секунд, в дополнение к интенсивной гражданской подготовке и репетициям по безопасной эвакуации из зданий, являются ключом к предотвращению гибели людей в случае сильного землетрясения. SASMEX также имеет приложение для Android и iOS, а также аккаунт в Twitter для распространения уведомлений с сенсорных станций. Спрос на точные и широко распространенные системы раннего оповещения привел к появлению на мексиканском рынке ИИ-приложений. Есть две основные компании в этой области: Skyalert и Grillo. Обе они используют IoT и облачные вычисления для создания сейсмических предупреждений в приложениях на мобильных устройствах; а в случае Skyalert – для систем голосового оповещения. Skyalert в большей степени ориентирована на корпоративный рынок, предлагая обучение и информирование, а также системы оповещения на месте. Эта компания также обеспечивает оповещение за 120 секунд до прихода разрушительной волны с помощью персонализированной системы GPS-сигнализации, которая фильтрует сейсмические события для пользователя. Grillo имеет нулевой показатель ложных срабатываний и постоянно выигрывает в сравнении с Skyalert и SASMEX. Обе компании используют облачные вычисления для обработки данных с собственных датчиков. Grillo работает с облачными сервисами AWS, время доставки данных от датчика до облака составляет 100 мс, а Skyalert обрабатывает свои данные на платформе Azure от Microsoft. Grillo имеет линию датчиков для обнаружения структурных повреждений зданий, чтобы предотвратить дальнейшие человеческие потери после землетрясения. В целом, компании Grillo и Skyalert занимают 97% рынка услуг раннего сейсмического оповещения. Очень характерной инициативой Grillo является партнерство с IBM и Linux Foundation. Grillo открыла свои данные и код в рамках проекта OpenEEW, который стимулирует глобальное сотрудничество вокруг технологии Grillo, в надежде разжечь подобные инициативы по всему миру.

Побудительным мотивом применения ИИ в государственном секторе Мексики является достижение общего блага, достижение и защита общественных интересов на благо всех граждан. Многие проекты ИИ, как на уровне исследований, так и на уровне приложений, направлены на решение социальных проблем, даже если эти проекты являются частной инициативой. Иногда грань между частными и общественными интересами не является достаточно четкой, чтобы классифицировать конкретное исследование или приложение ИИ. Однако есть и явные случаи, когда применение ИИ возникает как ответ на деловую возможность, рыночную нишу или в результате целенаправленной инновационной деятельности корпораций.

II. Применение ИИ в Мексике, обусловленное частными интересами

В отсутствие общего открытого реестра ИИ-приложений в Мексике мы можем рассказать лишь о некоторых наиболее заметных частных предприятиях, основанных на ИИ-приложениях. Почти во всех основных отраслях промышленности уже существуют формы применения ИИ-приложений, начиная с автомобильной промышленности, где используются роботы и мехатронные инструменты для сборки автомобилей, и заканчивая армией ботов, которые политические партии использовали в прошлом для дезинформации или раскола общества. В то же время, близость мексиканской экономики к рынку США способствовала росту предложения услуг, базирующихся в США (или имеющих корпоративную матрицу в США) на мексиканской территории. В этом отношении мексиканцы взаимодействуют с глобальным предложением услуг, связанных с цифровыми товарами: Netflix, Uber, DiDi, Facebook, Twitter, Instagram, Amazon и Google, например. В этом широком контексте есть некоторые области местных инноваций, которые начались от ныне консолидированных стартап-компаний или от крупных корпораций, которые инвестировали в использование ИИ в рамках своих программ исследований и разработок, чтобы сохранить конкурентное преимущество. Со стороны стартапов мы имеем широкий спектр инноваторов,

которые в настоящее время предоставляют ИИ-приложения на мексиканский рынок цифровых услуг. В число инноваторов входят FinTech, CRM-чат-боты, аналитика социальных сетей, недвижимость и медицинские роботы как услуга. Поскольку рассматриваемые нами примеры — это лишь несколько участников более широкого рынка США-Мексики, неизбежно будет видно, что в большинстве случаев инициативы являются местным ответом (или версией) других существующих решений для англоязычных потребителей и услуг в США.

По данным Мексиканского агентства по финансовому регулированию (CNBV, сокращенно на испанском языке), к 2019 году в Мексике действовало 500 FinTech-компаний; однако другая независимая фирма насчитывает 640. Согласно данным Statista, к 2022 году мексиканские FinTech-компании будут осуществлять операции, которые будут оцениваться примерно в 69 000 миллионов долларов США, с ежегодным темпом роста 17,3%. С более чем 158 стартапами, Мексика имеет самый большой рынок FinTech в Латинской Америке, даже больше, чем в Бразилии и Колумбии. При этом только 4,5% FinTech-компаний в Мексике прекратили свою деятельность с июня 2019 года по июнь 2020 года. Основные операции в FinTech интенсивно проводятся в области машинного обучения для оценки рисков, использования Больших данных и искусственного интеллекта. Почти весь бизнес-цикл компаний FinTech вращается вокруг использования технологий, связанных с ИИ: от платежей и денежных переводов, управления личными финансами до краудфандинга и кредитования. В мексиканском сценарии существует огромная вселенная FinTech и смежных «технологий», таких как InsurTech. Акцент на применении ИИ, который использует каждая компания FinTech, тесно связан с тем, в какой сфере использование ИИ принесет наибольшую конкурентоспособность.

Для Conekta, FinTech, предлагающей свои услуги в качестве сертифицированного агрегатора платежей (будь то наличные, «точка-точка» или электронная коммерция), безопасность — это все. Эта компания

разработала систему Conekta Shield, которая использует машинное обучение и байесовские сети для моделирования поведения потребителей и пользователей с целью создания динамических моделей для минимизации мошенничества и транзакционных рисков для клиентов Conekta. Как и Conekta, Dapp имеет схожую философию бизнеса, которая заключается в автоматизации платежей в различных экосистемах. В отличие от Conekta, Dapp предоставляет банкам технологию для генерации QR. Пользователи сканируют QR со своего мобильного телефона и оплачивают с помощью выбранного ими приложения-кошелька. Как только кошелек переводит средства за транзакцию, коммерсант получает платеж.

Другие финтех-компании фокусируются на использовании больших данных и машинного обучения для составления профиля клиента и «подталкивания» его на протяжении всего процесса повторной оплаты.

Это случай Kubo Financiero, первого сообщества онлайн-кредитования «равный – равному» в Мексике и первого в Латинской Америке, уполномоченного финансовым органом. Kubo Financiero предоставляет онлайн-кредиты с мобильного телефона в размере от 400 до 4 100 долларов США, а также предлагает привлекательную процентную ставку для микроинвесторов, которая начинается всего с 22 долларов США. Процесс профилирования включает в себя заполнение анкеты, загрузку нескольких документов и видеоконференцию. Соискатели кредитов с хорошей кредитной историей получают более выгодные процентные ставки и условия, а инвесторы, которые кредитуют заемщиков, получают более высокую прибыль на свои вложения. Kubo использует супервизорные модели из науки о данных для обогащения процессов сбора данных от клиентов, а также машинное обучение для калибровки механизма кредитного скоринга. На аналогичном сегменте рынка компания Kueski использует AI-приложения для определения профиля оценки риска при выдаче денег в долг, используя информацию из нетрадиционных точек данных, включая присутствие в социальных сетях. Заявки на кредит принимаются и оплачиваются в течение нескольких минут. Заемщики могут

постепенно запрашивать более высокие суммы. Специализацией Kueski являются микрозаймы, которые составляют от 100 до 200 долларов США на срок до 30 дней. Оценка рисков с использованием ИИ может применяться для улучшения прогноза успешности малого бизнеса. В этой области ИИ может быть использован для прогнозирования вероятности того, удастся ли идее или предпринимательству стать высокоэффективной компанией в мексиканской экономике. Konfio использует запатентованные алгоритмы и анализ данных для расширения доступных кредитных операций и ускорения процесса финансирования малых и средних предприятий (МСП). МСП в Мексике составляют 95% рынка. Успех Konfio основан на интенсивном использовании точек данных по балансу своих потенциальных клиентов для прогнозирования их шансов на успех и последующей адаптации оптимальной ставки кредитования для достижения этого успеха в МСП. Используя алгоритмы в циклах совершенствования, Konfio удалось снизить ставку суровости к стоимости кредита в 10 раз с 2016 по 2017 год.

Хотя профилирование рисков особенно чувствительно для платформ денежного кредитования, оно является ключевым моментом для компаний, использующих цифровые валюты, такие как биткойн. В Мексике Bitso — компания, которая считает, что ключ к глобальной финансовой доступности будет достигнут за счет использования Биткойна и других криптовалют (в настоящее время компания использует 10 различных видов криптовалют). Они инвестировали значительные средства в технологии искусственного интеллекта для автоматизированной проверки клиентов с целью предотвращения отмывания денег и финансирования терроризма.

В то же время они упростили фронт-энд своего приложения для инвестирования, использования и транзакций с криптовалютами, которые опираются на технологию Blockchain.

Наконец-то появился сегмент рынка, который также полагается на интенсивные приложения машинного обучения и больших данных для взаимодействия с клиентами, чтобы предлагать услуги и обеспечивать

конкурентоспособные транзакции. FinTech-компании, называемые «нео-банками», сами по себе являются процветающим примером использования приложений, связанных с искусственным интеллектом. Подробное описание каждого участника мексиканского рынка не входит в рамки данного исследования. Достаточно признать, что все нео-банки в той или иной форме используют технологии, связанные с ИИ: будь то автономное взаимодействие с клиентами, обработка данных для оценки рисков, предотвращение киберпреступлений или персонализированные микро-целевые услуги на основе структурных моделей Больших Данных.

Оценка рисков с использованием нетрадиционных точек для создания профиля является довольно распространенной техникой ИИ, применяемой в секторе FinTech. Менее распространенное применение методов предиктивного моделирования рисков – в сфере аренды домохозяйств. Мексиканский рынок аренды городской недвижимости для домашнего использования является доминирующим. Некоторые опросы показывают, что примерно 42% населения Мексики живет в арендованных домах. Nomie использует прогностическую модель с сотнями переменных для оценки вероятности невыплаты арендатором арендных платежей (коэффициент просрочки арендаторов). Модель машинного обучения анализирует не только доход и кредитную историю заявителя, но и степень соблюдения им финансовых обязательств. Nomie выбирает лучшего кандидата для владельцев. После выбора лучшего кандидата Nomie заботится о сборе арендной платы и даже имеет страховой полис, по которому в случае невыполнения обязательств он выплатит владельцу деньги и вернет имущество в течение четырех месяцев. Служба поддержки клиентов Nomie опирается на четкие алгоритмы прогнозирования, которые являются одним из преимуществ, заставляющих владельцев доверять свои активы Nomie. В этом отношении первый отклик между клиентами и службами является ключевым фактором для поддержания прочной и вовлеченной клиентской базы.

ИИ-приложения для виртуальных помощников

AI-приложения в виде ChatBots с различными уровнями сложности сегодня являются глобальной тенденцией для решения вопросов, комментариев и претензий клиентов, а также для привлечения новых клиентов. Согласно отчету Oracle о развивающихся технологиях, ведущие компании будут инвестировать в чат-боты, в том числе в интеллектуальные голосовые приложения. 51% респондентов заявили, что выгода от инвестиций в интеллектуальные голосовые технологии заключается в «Ускорении времени решения проблем клиентов», а 50% ответили, что их основная выгода заключается в «Повышении операционной эффективности». По прогнозам Juniper Research, к 2022 году компании сэкономят 8 миллиардов долларов США за счет использования ChatBots для обслуживания клиентов. Аналогично отчету Oracle, Gartner оценивает, что к 2022 году 70% белых воротничков будут ежедневно взаимодействовать с разговорными платформами.

В Мексике существует более нескольких компаний на базе ИИ, которые предлагают услуги ChatBot с различной степенью автономности и диапазоном взаимодействия. По данным Nanalyze, YaloChat объединил семь других глобальных компаний, которые легко внедряют технологию ChatBot – то есть круг компаний, которые могут «предоставить нам ChatBot быстро и с минимальной суетой». YaloChat работает с компаниями, чтобы понять отношения с клиентами, а затем разработать индивидуального Бота для улучшения качества обслуживания. YaloChat отправляет и управляет уведомлениями клиентов, включая демонстрацию продуктов. ИИ, стоящий за YaloChat, автоматизирует часто задаваемые вопросы, предсказывая 90% потенциальных разговоров с клиентами. Наконец, YaloChat детализирует профили клиентов, чтобы предоставить им особый интерактивный опыт для повышения вовлеченности в процесс обслуживания. YaloChat создал ботов, ориентированных на ИИ, для многих крупных компаний, работающих в Мексике, включая Amazon, Pepsi, Volkswagen, Aeromexico и Walmart. По стопам YaloChat идет Gus Chat, еще одна мексиканская компания,

занимающаяся проектированием и разработкой ИИ-ботов для автоматизации обслуживания клиентов, специализирующаяся на электронной коммерции, финтех и страховании. Gus Chat разрабатывает различные алгоритмы для каждой задачи, поставленной перед ботом: генерация лидов продаж, транзакционные боты, боты для обслуживания клиентов, боты для B2B, боты для маркетинга и рекламы. Такое разнообразие алгоритмов делает Gus Chat инновационной компанией в области обработки естественного языка в регионе.

Использование алгоритмов естественного языка для поддержки клиентов и услуг в виде интерактивных ботов – одно из многих применений этой технологии ИИ. Распознавание символов и преобразование голоса в текст также являются двумя ключевыми технологиями, которые улучшают бизнес-процессы в Мексике.

Nowports³¹¹ это платформа, которая соединяет клиентов, поставщиков, таможенные службы, порты и перевозчиков. В процессе используется блокчейн, чтобы сделать каждую транзакцию безопасной и прозрачной для каждого участника процесса. Nowports обрабатывает каждый запрос на перевозку груза автоматически по электронной почте, независимо от формата или документации. Система искусственного интеллекта обрабатывает запрос, а затем размещает бюджет с фиксированной ценой без вариаций. Агент искусственного интеллекта соединяет все остальные конечные точки для завершения транзакции; после одобрения клиентом он размещает заказ на отгрузку товара. Клиент может в режиме реального времени отслеживать путь товара вплоть до пункта доставки. Благодаря технологии искусственного интеллекта Nowports может обрабатывать логистические запросы на 70% быстрее, чем другие компании, занимающие конкурирующую долю рынка.

При использовании ИИ-ботов сочетание когнитивных вычислений, облачного компьютера и общей обучающей архитектуры ИИ образует мощную комбинацию. Nearshore Delivery Solutions (NDS)³¹² – мексиканская компания,

311 <https://nowports.com>.

312 <https://nearshoremx.com/cognitive>.

которая сотрудничает с IBM для обучения и создания когнитивных вычислительных ботов. NDS смогла создать виртуального помощника для крупного мексиканского банка всего за 12 недель. NDS обучила IBM Watson Assistant с помощью анализа электронной почты, стенограмм телефонных звонков и сообщений чата, которыми в течение года обменивались клиенты и служба поддержки банка. Благодаря знаниям, полученным в результате такого глубокого и всестороннего анализа, IBM Watson Assistant смог ответить почти на 1 000 вопросов. Бот также определял, когда вопрос требовал вмешательства человека, чтобы перенаправить звонок к эксперту-человеку. Кроме того, NDS использовала Watson Tone Analyzer для определения тревоги или гнева клиента для быстрого вмешательства человека, чтобы предотвратить дальнейшее разочарование клиента. 75% всех звонков клиентов теперь обрабатываются когнитивным ботом, который решает вопросы в среднем за две минуты, что на 80% быстрее, чем обычный помощник в центре обработки вызовов.³¹³

Использование такой общей архитектуры ИИ, как Watson, обеспечивает надежную основу для инноваций в области когнитивных вычислений. В частности, анализ тональности и эмоциональных реакций является важнейшей информацией для подготовки точных ответов клиентам и заказчикам.

Существуют две мексиканские компании, которые разрабатывают и применяют технологии ИИ для управления сообщениями. Metric – мексиканская компания, которая начала создавать автоматизированный контент для социальных сетей. Затем они перешли к программированию ботов для определения эмоций и отношения, и, таким образом, для реагирования с помощью специфического контента, чтобы повлиять на эти эмоции. В настоящее время их бизнес заключается в моделировании и прогнозировании поведения в социальных сетях. Их продукты генерируют оценочные прогнозы влияния кампаний, позиций заинтересованных сторон и тенденций до того, как они произойдут.³¹⁴ Они могут оценить сценарии рисков (и возможностей) для

313 <https://www.ibm.com/downloads/cas/JOEXYB08>.

314 <https://buy.metricser.com/revelio/>.

стоимости торговых марок, включая обнаружение ботов, троллей и фальшивых новостей. Аналогичным образом, но ориентируясь на рекламодателей и рекламные агентства, мексиканские.

Adext³¹⁵ использует возможности искусственного интеллекта для проведения кампаний в социальных сетях. Adext создал модели машинного обучения, которые запускают симуляцию для определения наилучшего времени, целевой аудитории и цифрового пространства (размещения) для рекламы в социальных сетях. Моделирование определяет оптимальное маркетинговое воздействие с учетом бюджетных ограничений из тысяч потенциальных сценариев, чтобы получить невероятную детализацию для аудитории. Adext утверждает, что с самого начала кампании дают не менее 25% прироста. Эта компания также предлагает демонстрацию, позволяющую в режиме реального времени увидеть количество конверсий, полученных кампанией с использованием их моделей.³¹⁶

Боты (как и их физические собратья – роботы) работают с использованием моделей и информации для выполнения конкретных задач. Понятие цифрового пространства делает ботов и их присутствие менее очевидным, чем присутствие роботов-гуманоидов. Однако в области искусственного интеллекта в Мексике наблюдаются успехи в проектировании и разработке сервисных роботов (большинство из них производятся в других странах).³¹⁷

III-приложения для служб здравоохранения

Обнаружение и обеспечение быстрой изоляции инфекционных кластеров — одна из многих составляющих борьбы с пандемией COVID-19. Другой проблемой является уход за пациентами, зараженными вирусом, во

315 <https://www.adext.ai>.

316 <https://www.latamdigitalmarketing.com/blog/software-campanas-digitales/>.

317 Есть много примеров, когда роботы уже используются в качестве сервиса, см. например, <https://integritas.mx/robots-de-servicio/>, <http://robotsmexico.mx>, а также исследовательскую работу Сукара для анализа зрелости этой области в Мексике в <https://ccc.inaoep.mx/~emorales/Papers/2009/eduardo.pdf> (все сайты посещены 31 октября 2020 года).

время госпитализации. В настоящее время существует огромное количество информации, описывающей риски и заражение медицинских работников, которые ухаживают за инфицированными пациентами. Мексиканская компания GESEDIG предлагает три модели сервисных роботов, которые могут помочь медицинскому персоналу в уходе за больными Covid-19. Роботы-гуманоиды могут опрашивать пациентов, когда те прибывают в медицинский центр. Они могут совершать обходы в приемных и залах ожидания, демонстрируя информационные видеоролики и обеспечивая профилактику заболеваний. Роботы также могут направлять пациентов в различные медотсеки, используя датчики движения, а также доставлять еду и выдавать лекарства. Роботы-гуманоиды являются средством предотвращения дальнейшего распространения вируса среди медицинских работников. ИИ робота был разработан и запрограммирован мексиканскими инженерами. Ядро ИИ включает модуль обработки естественного языка, распознавание лиц, считывание штрих-кодов, процессы управления взаимоотношениями с клиентами и возможности Click to Call.³¹⁸

В Мексике мы наблюдаем многочисленные случаи использования приложений ИИ в различных сферах. Начиная с общественных интересов и заканчивая частным сектором, можно увидеть, как приложения ИИ влияют на отрасли, услуги и государственные задачи. Жизненный цикл ИИ-приложения включает в себя финансовые, технологические и человеческие ресурсы. В развивающейся экономике, такой как мексиканская, переменные, которые будут негативно влиять на процветание исследований и приложений ИИ, обширны, но не безграничны. Аналогичным образом, существует неизвестная степень взаимодействия между индивидуальным мышлением, коллективной осведомленностью, институциональной инфраструктурой, политикой и законодательной базой для стимулирования и поощрения развития ИИ. Анализ государственной политики, нормативно-правовой базы и частных

318 <https://gesedig.com>.

координационных инициатив может дать частичное представление о динамике развития ИИ в Мексике.

III. Учет регулирования ИИ в Мексике

В Мексике не существует правовой, законодательной базы, которая регулировала бы вопросы ИИ. Отсутствие законодательного органа по ИИ — это отсутствие как на федеральном уровне, так и на уровне местного штата (т.е. местного законодательного органа). Однако существует законодательство и свод правил различной иерархии и источников в конкретных областях, где появляются ИИ-приложения. Полезно признать, что в Мексике существует косвенное регулирование систем ИИ. Отсутствие законодательного инструмента для регулирования систем ИИ восполняется государственной политикой, которая реализуется по крайней мере с 2019 года, хотя существует государственная политика в области цифровой трансформации и цифровой инклюзии, которая реализуется по крайней мере с одного десятилетия. Мексика является федеративной республикой, что означает, что в вопросах ИИ есть два лица, определяющих политику. Мы видим, что федеральное правительство прилагает усилия для стимулирования инноваций и развития технологий ИИ. В то же время, некоторые местные органы власти также приняли различные правительственные меры в пользу внедрения технологий ИИ. Помимо двух уровней регулирования и выработки политики — федерального и местного, мы обнаружили, что существуют государственные агентства с конституционными полномочиями, которые также формулируют правила, относящиеся к системам ИИ. В Мексике есть государственные и частные университеты, а также государственные и частные исследовательские центры, и все они, так или иначе, заинтересованы в государственной политике, которую продвигают федеральные и местные органы власти. Чтобы оценить конкретные механизмы проведения государственной политики в области ИИ, нам нужно взглянуть на федеральное правительственное агентство, отвечающее за технологии и развитие. Если мы посмотрим на примеры, которые мы представили в первой

части этой главы, то увидим, что большинство инициатив возникли еще до того, как в Мексике начала действовать официальная федеральная государственная политика. В то же время, проанализированные нами ИИ-приложения, основанные на частных интересах, не являются прямым результатом какой-либо конкретной политики в отношении развивающихся ИИ-инициатив в Мексике. Большинство ИИ-приложений в частном секторе являются результатом сочетания рыночных сил и особого индивидуального стремления к инновациям. Возможно, отсутствие сильной законодательной базы и некоторые косвенные благоприятные условия регулирования привели к тому импульсу развития ИИ в Мексике, который мы наблюдаем в последние годы. В то же время организованные коллективы с членами из различных областей (таких как промышленность, наука, предпринимательство и активизм) предлагают национальную повестку дня для развития ИИ в Мексике. Демократические, спонтанные действия коллектива в пользу ИИ не имеют прецедентов в стране. Мы представим и проанализируем правовую базу, косвенно относящуюся к ИИ-приложениям, и государственную политику как на федеральном, так и на местном уровне, которая была предложена и реализована в Мексике (А); прежде чем перейти к анализу коллективного движения в поддержку ИИ, которое выдвигает повестку дня тем для регулирования мексиканским государством (Б).

Чтобы представить нормативно-правовую базу, которая косвенно затрагивает ИИ-приложения, мы проведем различие между национальными и местными нормативно-правовыми актами. В Мексике определенные правовые области могут регулироваться только и исключительно федеральным Конгрессом, исключая, таким образом, полномочия местных законодателей. Правовая область, которая является исключительной властью федерального Конгресса, называется национальным регулированием. Существуют и другие правовые области, в которых местные и федеральные законодатели могут вырабатывать нормы, действующие одновременно. По такому же принципу мы

будем различать политику федерального правительства и политику местного правительства в отношении ИИ.

Защита персональных данных

Права на защиту персональных данных регулируются на конституционном уровне. Конституция Мексики устанавливает национальное агентство с полной автономией для надзора за соблюдением прав, связанных с защитой данных. Но в то же время она повторяет структуру автономного надзора за соблюдением прав на защиту данных в каждом местном штате Мексиканской Федерации. Национальное агентство по надзору за защитой прав, связанных с персональными данными (называемое на испанском языке аббревиатурой INAI), обладает достаточными полномочиями для проведения аудита, наложения санкций и даже отзыва операций в случае нарушения принципов и стандартов, касающихся персональных данных. Интересно отметить, что Мексика является участником 108 Конвенции, а также Дополнительного протокола к Конвенции о защите физических лиц при автоматизированной обработке персональных данных, касающегося надзорных органов и трансграничных потоков данных. В дополнение к национальным и международным правовым документам INAI является частью Иbero-американской сети по защите данных (NDP). NDP выпустила Специальные руководства по соблюдению принципов и прав, регулирующих защиту персональных данных в проектах искусственного интеллекта, а также Общие рекомендации по обработке персональных данных.³¹⁹ Оба документа считаются «мягким правом» с точки зрения их институциональной силы в мексиканской правовой системе, и, тем не менее, они обеспечивают очень подробные руководящие рамки принципов и стандартов для INAI.

Любой человек может подать иск в INAI против любого правительственного агентства, частной организации или физического лица,

³¹⁹<http://inicio.inai.org.mx/nuevo/GeneralRecommendationsfortheProcessingofPersonalDatainArti%EF%AC%81cialIntelligence.pdf>

которые могли нарушить конституционные и юридические обязанности по защите персональных данных. В Мексике не было ни одного случая обращения в национальный орган по правам персональных данных, связанного со сбором, обработкой или использованием данных в приложениях ИИ.

Финансовое регулирование

В отличие от защиты данных, финансовое регулирование в Мексике относится исключительно к федеральному регулированию, общенациональному. Если какая-либо компания или стартап FinTech в Мексике принимает деньги от населения в целом, или если они предоставляют какие-либо банковские услуги, эти компании и стартапы обязаны соблюдать Федеральный закон о FinTech наряду со всеми подробными правилами, изданными Федеральным управлением по финансовому регулированию и правилами, применимыми Центральным банком Мексики. С точки зрения нормативной системы, AI-приложения не являются главной темой нормативной системы, регулирующей FinTech. Нет ни одной ссылки, которая бы указывала на какой-либо стандарт для разработки или использования систем ИИ при предоставлении финансовых услуг с помощью технологий. Оценка финансовых рисков на основе доказательств, различные методы профилирования, а также информация, которую любой субъект рынка финансовых услуг обязан предоставлять регулятору, являются одинаковыми по обоснованию регулятора. Существует обоснование для поиска гарантий при получении денег и адекватного управления рисками при предоставлении денег в долг, независимо от технических и технологических средств для определения клиентской базы. Среди сообщества FinTech в Мексике широко признано, что федеральное регулирование способствует развитию бизнеса и инноваций. В этом отношении статут FinTech косвенно поддерживает легкий подход к технологическим инновациям на рынке финансовых услуг. Ожидается, что благодаря законодательной политике, которая не слишком навязчиво относится к

инновациям, системы и приложения искусственного интеллекта на рынке FinTech также будут процветать.

Хотя регулирование финансовых услуг не направлено конкретно на ИИ-приложения, федеральные и местные органы власти проводят активную пульсирующую политику по продвижению и поддержке ИИ-приложений в Мексике.

21 марта 2018 года состоялась презентация исследования, проведенного по заказу посольства Великобритании в Мексике при поддержке офиса президента Мексики компанией Oxford Insights. Исследование представляло собой проект национальной стратегии в области ИИ для правительства Мексики. В конце презентации мексиканское правительство представило Стратегию искусственного интеллекта в Мексике на 2018 год (Стратегия MX-AI2018). Стратегия состояла из пяти мер государственной политики: 1) Разработать адекватную структуру управления для открытого диалога с участием многих заинтересованных сторон. Правительство создаст комиссию министерств по развитию электронного правительства. 2) Определить потребности отрасли и лучшие практики в правительстве, связанные с ИИ. 3) Выступать за международные усилия в области ИИ, уделяя особое внимание роли Мексики в ОЭСР и G-7. 4) Открыть общественные консультации по рекомендациям исследования Oxford Insights. 5) Работать с экспертами и гражданами в подкомитете по искусственному интеллекту и глубокому обучению, созданном при Министерстве гражданской службы. В то время Мексика стала одной из первых десяти стран мира, инициировавших конкретные действия в области государственной политики, направленные на развитие, внедрение и использование искусственного интеллекта.

В 2019 году Мексика приняла Принципы ОЭСР по искусственному интеллекту «для продвижения ИИ, который является инновационным и заслуживающим доверия, а также уважает права человека и демократические ценности». Приняв Принципы ОЭСР по ИИ, Мексика была вынуждена следовать пяти взаимодополняющим принципам, основанным на ценностях,

для ответственного управления заслуживающим доверия ИИ, а также пяти рекомендациям ОЭСР для правительств. К тому времени федеральное правительство все же создало комиссию министерств по развитию электронного правительства для координации различных усилий по внедрению стратегии MX-AI2018.

Одним из важных результатов национальной государственной политики стала инициатива, организованная Национальным советом по науке и технологиям (CONACYT – сокращенное название на испанском языке). CONACYT отвечает за контроль и финансирование государственных исследований через Национальные государственные исследовательские центры (NPRC) по всей стране. Каждый центр имеет независимую и автономную программу исследований отчасти потому, что они создаются в государственных университетах, а затем аккредитуются CONACYT как NPRC. Под руководством CONACYT восемь NPRC объединили усилия для создания Инициативы по искусственному интеллекту. Эта инициатива позволяет проводить совместные исследования в междисциплинарных областях между исследователями, аспирантами и студентами из разных областей знаний и дисциплин. Инициатива направила свои усилия на моделирование природных и социальных явлений в области медицины, общественной безопасности, мобильности людей и транспорта, предотвращения стихийных бедствий и подготовки кадров в области искусственного интеллекта. С тех пор в рамках инициативы было проведено огромное количество исследований, представленных в различных публикациях в престижных журналах по различным дисциплинам. В рамках инициативы также были созданы исследовательские группы из ученых и аспирантов из различных NPRC; на данный момент их всего 7, но темы постоянно пересматриваются, чтобы идти в ногу с передовыми исследованиями.³²⁰ Альянс доказал, что в Мексике самые богатые людские ресурсы для междисциплинарной работы над ИИ-приложениями.

320 <https://www.consorticioia.mx/grupos-investigacion>.

Параллельно с государственной политикой федерального правительства мы видим, что местные власти штатов также активно рассматривают ИИ-приложения. В более широком контексте следует всегда помнить об огромной территории Мексики, а также о разнообразии между различными штатами. В этом отношении некоторые штаты имеют большее развитие в плане человеческих ресурсов, промышленности и дохода на душу населения, чем другие. Поэтому неудивительно, что некоторые инициативы в области государственной политики, выдвигаемые местными органами власти совместно с частным сектором, исходят из конкретных регионов страны. Штаты Пуэбла и Керетаро начали местные программы по преобразованию своих столиц в «умные города» с использованием приложений ИИ. Правительства штатов Юкатан и Гуанахуато также начали реализацию агрессивных планов государственной политики по привлечению технологической промышленности, они создали альянсы с соответствующими технологическими компаниями или исследовательскими центрами, а также поддерживали местные государственные университеты, чтобы они играли важную роль в государственной политике. В столице Юкатана, городе Мерида, реализуется проект по созданию исследовательского центра искусственного интеллекта. В Веракрусе, южном штате, исследовательский центр по ИИ работает с 1994 года. Другие местные штаты с исследовательскими центрами по ИИ – Чиуауа, Нуэво Леон и Халиско.

Халиско – пожалуй, самый заметный из вышеперечисленных штатов Мексики. Халиско называют мексиканской Силиконовой долиной из-за количества стартапов, крупных технологических игроков, которые разместили в штате свои лаборатории исследований и разработок, а также из-за высокой плотности специалистов технологического профиля.³²¹

В прошлом году местное правительство Халиско заключило соглашение о сотрудничестве с Tec de Monterrey (одним из самых престижных частных технологических университетов Мексики) и Межамериканским банком развития для создания первого в Мексике центра искусственного интеллекта.

321 <https://inmobiliare.com/por-que-guadalajara-es-el-silicon-valley-mexicano/>.

Центр предполагает инвестиции в размере 900 000 долларов США от Tec de Monterrey, участие Intel, IBM и Amdocs, Sparkcognition, Tata Consultancy Services и Wizeline, Калифорнийского университета Беркли, Пекинского технологического института и Института исследований по информатике и автоматизации из Франции. Одним из новшеств этого консорциума является то, что он будет сосредоточен на решении социально ориентированных сложных проблем государственной политики в Халиско в области здравоохранения, образования и общественной безопасности. Консорциум будет работать над созданием ИИ-приложений для лечения и выявления диабетической ретинопатии, для предотвращения отсева из школ и для предотвращения преступности среди несовершеннолетних.³²²

Частный сектор был довольно активен в плане создания приложений ИИ и развития инициатив по всей стране. Возможно, наиболее значимым мероприятием в пользу разработки комплексной дорожной карты для национальной повестки дня по искусственному интеллекту является коллективный проект IA2030.

II. Коллективная основа для регулирования ИИ и о регулировании ИИ

IA2030 — это коллектив организаций и ключевых участников организаций из гражданского общества, научных кругов, частного сектора, промышленности в целом, независимых консультантов, государственных чиновников и государственных исследовательских центров. Руководство коллективом осуществляет стартап под названием C Minds. Результатом интенсивной координации усилий стали два продукта: национальный опрос и национальная программа развития искусственного интеллекта в Мексике.

Опрос был разработан почти 50 независимыми организациями и профессионалами на общественных началах. Опрос был доступен онлайн в течение одного месяца в сфере ИИ с 15 августа по 18 сентября 2018 года. Хотя результаты опроса нельзя обобщить (то есть он не имеет статистической

322 <https://tec.mx/es/noticias/guadalaiaara/investigacion/asi-sera-el-innovador-hub-de-inteligencia-artificial-del-tec>.

достоверности), он служит хорошим социальным термометром для заинтересованных сторон в области ИИ. Из 1585 респондентов почти 90% были из Мехико, штата Мехико и Халиско. Во всех трех местах сосредоточено большое количество технологических усилий и человеческих ресурсов, связанных с ИИ в Мексике. 43% респондентов работали на правительственных должностях, 31% – в частном секторе и 18% – в академических кругах. Среди основных выводов исследования можно отметить, что 80% респондентов считают, что ИИ окажет положительное влияние на их жизнь. Однако 53% считают, что с появлением ИИ возрастет безработица, а 45% обеспокоены тем, что их частная жизнь и личные данные могут быть нарушены ИИ-приложениями. Почти 45% заявили о существовании этических проблем и проблем неравенства в связи с широким использованием ИИ без адекватного регулирования. Что касается государственной политики, то здесь мнения разделились на активную роль мексиканского государства в стимулировании исследований, разработок и внедрения ИИ-технологий, улучшении государственных услуг и формировании большего количества человеческих ресурсов, специализирующихся на ИИ-приложениях.

Национальная повестка дня по ИИ является результатом совместных усилий 400 человек из различных секторов и профессий, разделенных на шесть рабочих групп в соответствии с опытом каждого участника: Данные, цифровая инфраструктура и кибербезопасность; Этика, управление, правительство и общественные услуги; Исследования и разработки; Образование, возможности и навыки; Охват мексиканцев-иммигрантов. Результатом работы является представление конкретных проблем в рамках каждой темы рабочей группы, предложение направлений действий и ключевых показателей для оценки предложенных направлений действий. Хотя конкретное содержание требует более детального анализа, чем тот, который представлен в этой главе, есть некоторые общие темы, возникающие в рабочих группах, которые стоит упомянуть. С одной стороны, идея стратегических целей в каждой теме. Затем, использование всеобъемлющих принципов. И, наконец, пересекающийся

анализ между темами во всем документе. С другой стороны, в отчете представлены показатели ключевых точек для каждого предложенного направления действий с указанием того, кто несет ответственность за реализацию курса действий (т.е. законодательная власть, регулирующий орган или гражданское общество). С этой точки зрения доклад является не только «дорожной картой» для регулирования ИИ в Мексике, но и GPS, показывающим, как возможные нормативные рамки пересекаются друг с другом как по содержанию, так и по охвату и нормативным аспектам.

IV. Выводы

Смотреть на движущееся изображение — это не то же самое, что смотреть фильм. ИИ-приложения в Мексике — это движущееся изображение, композиция действующих лиц и инициатив, предприятий и технологий в быстрой фазе. Даже если может показаться, что расположение проектов в этой главе имеет последовательный порядок, в действительности между экосистемой ИИ в Мексике существует сложное взаимодействие, которое трудно систематизировать. Сложность обусловлена фрагментарным представлением о продукте эпизодических инициатив, процветающих при наличии конкретного основания в институциональной программе. Большинство представленных здесь примеров использования являются скорее продуктом индивидуального мышления и таланта, чем результатом создания среды, подобной инкубатору для ИИ. Тем не менее, движение инициатив в области ИИ, представленных здесь, и тех, которые были оставлены в ходе исследования для данной статьи, указывает на то, что мексиканский рынок ИИ-приложений является динамичным и многообещающим. В стране с глубоким неравенством и все еще не решенным вопросом о верховенстве закона и правах человека всегда есть место для инвестиций в институты, и даже больше, если они способствуют повышению благосостояния и инклюзии с помощью технологических достижений. Существует достаточно заинтересованных сторон как в частном, так и в государственном секторе, а также в научных

кругах, чтобы инициировать плодотворную институционализацию ИИ в Мексике. Однако инертность огромного нерегулируемого рынка может стать главной чертой мексиканского случая.

Ссылки

Определение ИИ: основные возможности и научные дисциплины Группа экспертов высокого уровня по искусственному интеллекту, Европейская комиссия, 8 апреля 2019 года.

Статья 4 Общего положения о прозрачности и доступе к публичной информации на сайте http://www.diputados.gob.mx/LeyesBiblio/pdf/LGTAIP_130820.pdf (доступ получен 31 октября 2020 года).

Статья 6 Конституции Мексики на http://www.diputados.gob.mx/LeyesBiblio/pdf_mov/Constitucion_Politica.pdf (дата обращения: 31 октября 2020 г.).

Duffin Erin, Number of Mexican Immigrants in the United States 1850-2018 на <https://www.statista.com/statistics/673350/mexican-immigrants-in-the-united-states/> (дата обращения: 31 октября 2020 г.).

<http://inicio.inai.org.mx/nuevo/GeneralRecommendationsfortheProcessingofPersonalDatainArtificialIntelligence.pdf> (дата обращения: 31 октября 2020 г.).

<http://inicio.inai.org.mx/nuevo/SPECIFICGUIDELINESARTIFICIALINTELLIGENCEPROJECTIONS2019.pdf> (дата обращения: 31 октября 2020 г.).

http://omawww.sat.gob.mx/gobmxtransparencia/Paginas/documentos/estudio_opiniones/Evasion_en_IVA_Analisis_de_Redes.pdf (дата обращения: 31 октября 2020 г.).

<http://smia.mx/komputersapiens/> (дата обращения: 31 октября 2020 г.).

http://www.cires.org.mx/sasmex_es.php (дата обращения: 31 октября 2020 г.).

http://www.diputados.gob.mx/LeyesBiblio/pdf/LRITF_090318.pdf (accessed on October 31, 2020).

<http://www.ebizlatam.com/seguridad-e-inclusion-soluciones-conekta/> (доступ получен 31 октября 2020 г.).

<http://www.ift.org.mx/sites/default/files/contenidogeneral/transparencia/1vision19-23.pdf> (доступ получен 31 октября 2020 г.).

<http://www.oecd.org/going-digital/ai/principles/> (доступ получен 31 октября 2020 г.).

<https://adip.cdmx.gob.mx/proyectos>

<https://amqueretaro.com/queretaro/2020/10/22/convertiran-a-queretaro-en-ciudad-inteligente/>

<https://bitso.com> (дата обращения: 31 октября 2020 г.).

<https://blog.grillo.io/the-grillo-journey-e60b2dcee224>

<https://buy.metricser.com/revelio/>

<https://conekta.com>

<https://dapp.mx>

<https://data.oecd.org/mexico.htm#profile-innovationandtechnology>

<https://datos.cdmx.gob.mx/explore/dataset/carpetas-de-investigacion-pgj-cdmx/table/?q=FED%2FFECC%2FUNAI-CDMX%2F0000002%2F2020>

<https://datos.gob.mx/blog/estrategia-de-inteligencia-artificial-mx-2018>

https://earthquaketrack.com/p/mexico/recent?mag_filter=7

<https://elceo.com/tecnologia/cnbv-alista-autorizaciones-para-80-empresas-fintech-entre-octubre-y-noviembre/>

https://english.elpais.com/elpais/2017/03/13/inenglish/1489403756_441981.html

https://ensanut.insp.mx/encuestas/ensanut2018/doctos/informes/ensanut_2018_presentacion_resultados.pdf

<https://expansion.mx/tecnologia/2017/08/15/konfio-un-ejemplo-de-como-las-fintech-dejan-atras-a-los-bancos>

<https://gesedig.com>

<https://grillo.io/linux-foundation-hosting-openeew/> and <https://grillo.io/open-source/>

<https://gus.chat>

<https://home.inai.org.mx>

<https://homie.mx/h/>
<https://inmobiliare.com/homie-la-plataforma-que-facilita-la-renta-de-departamento>
<https://inmobiliare.com/por-que-guadalajara-es-el-silicon-valley-mexicano/>
<https://konfio.mx/>
<https://kueski.com>
<https://kueski.com/blog/tecnologia/machine-learning-aprendizaje-supervisado/>
https://nearshoremx.com/cognitive_computing
<https://news.microsoft.com/uploads/prod/sites/41/2018/11/IA-y-Crecimiento-MEXICO.pdf>
<https://nowports.com>
<https://panamericanworld.com/en/magazine/startups/meet-the-five-most-innovative-mexican-fintechs/>
<https://publications.iadb.org/publications/english/document/FINTECH--Innovations-You-May-Not-Know-were-from-Latin-America-and-the-Caribbean.pdf>
<https://skyalert.mx/>
<https://tec.mx/es/noticias/guadalajara/investigacion/asi-sera-el-innovador-hub-de-inteligencia-artificial-del-tec>
<https://www.adext.ai>
<https://www.banxico.org.mx/marco-normativo/normativa-agrupada-por-sujeto.html>
<https://www.cnbv.gob.mx/SECTORES-SUPERVISADOS/Fintech/Paginas/NORMATIVIDAD-FINTECH.aspx>
<https://www.consortioia.mx/grupos-investigacion>
<https://www.eleconomista.com.mx/estados/Centro-de-Inteligencia-Artificial-en-NL-alista-inicio-de-operaciones-20180820-0020.html>
<https://www.eleconomista.com.mx/estados/Guanajuato-epicentro-de-la-cuarta-Revolucion-Industrial-20201029-0104.html>
<https://www.eleconomista.com.mx/finanzaspersonales/En-el-2019-70-de-las-nuevas-familias-compraria-una-casa-20181216-0053.html>
<https://www.eleconomista.com.mx/sectorfinanciero/Tenemos-mapeadas-mas-de-500-fintech-en-Mexico-CNBV-20190526-0072.html>

<https://www.eluniversal.com.mx/cartera/sat-va-por-mas-recaudacion-con-inteligencia-artificial>
<https://www.eluniversal.com.mx/estados/proyectan-puebla-como-ciudad-de-inteligencia-artificial>
<https://www.facebook.com/InteligenciaArtificial.Center/>
<https://www.finnovista.com/radar/el-numero-de-startups-fintech-en-mexico-crecio-un-14-en-un-ano-hasta-las-441/>
<https://www.gartner.com/smarterwithgartner/chatbots-will-appeal-to-modern-workers/>
<https://www.gob.mx/cidge/articulos/crea-sfp-subcomision-de-inteligencia-artificial-y-deep-learning-de-la-cidge-161421?tab>
[https://www.gob.mx/cms/uploads/attachment/file/520516/Comunicado_UIF_011 .pdf](https://www.gob.mx/cms/uploads/attachment/file/520516/Comunicado_UIF_011.pdf)
<https://www.gob.mx/innovamx/articulos/inteligencia-artificial-131287>
<https://www.gob.mx/misalud/>
<https://www.gob.mx/shcp/prensa/comunicado-no-054-shcp-mediante-sat-y-pff-anuncia-acciones-en-contra-de-defraudacion-fiscal-a-traves-de-empresas-factureras>
<https://www.ibm.com/downloads/cas/JOEXYB08>
<https://www.infobae.com/tecno/2020/04/01/asi-es-el-chatbot-mexicano-que-lucha-contra-la-desinformacion-sobre-coronavirus/>
<https://www.innovatorsunder35.com/the-list/saiph-savage/>
[https://www.insp.mx/resources/images/stories/2019/Docs/190213_LaObesidadenMexico .pdf](https://www.insp.mx/resources/images/stories/2019/Docs/190213_LaObesidadenMexico.pdf), c. 28
<https://www.juniperresearch.com/new-trending/analystxpress/july-2017/chatbot-conversations-to-deliver-8bn-cost-saving>
<https://www.kubofinanciero.com/Kubo/Portal/productos/productos.xhtml>
<https://www.latamdigitalmarketing.com/blog/software-campanas-digitales/>
<https://www.legalparadox.com/post/neo-banks-who-will-control-the-mexican-market>
<https://www.nanalyze.com/2017/07/7-chatbot-platforms-chatbots-easy/>
<https://www.oracle.com/a/ocom/docs/dc/em/lpd100807811-impact-of-emerging-technology-on-cx->

excellence.pdf?elqTrackId=d368a11a304041c8a7bf8e7a2f2a71e2&elqaid=82669&elqat=2

<https://www.oxfordinsights.com/mexico>

<https://www.redipd.org/es/la-red/historia-de-la-red-iberoamericana-de-proteccion-de-datos-ripd>

https://www.researchgate.net/profile/Yasmin_Hernandez2/publication/340634786_Robotica_de_Servicio/links/5e96334a4585150839de623a/Robotica-de-Servicio.pdf

<https://www.statista.com/study/45600/statista-report-fintech/>

<https://www.uif.gob.mx/>

<https://www.uv.mx/ciia/>

<https://www.yalochat.com>

[https://www.youtube.com/watch?reload=9&v=1dv-](https://www.youtube.com/watch?reload=9&v=1dv-gVKBRBA&ab_channel=UNAMGlobal)

[gVKBRBA&ab_channel=UNAMGlobal](https://www.youtube.com/watch?reload=9&v=1dv-gVKBRBA&ab_channel=UNAMGlobal)

<https://yucatanahora.mx/yucatan-busca-desarrollar-proyectos-de-inteligencia-artificial/>

Rosado-Tamariz E, Zuniga-Garcia M A, Santamaria-Bonfil G, Batres R, A machine-learning approach to speed-up simulation towards the design of optimum operating profiles of power plants. В материалах 8-й Международной конференции по информатике, окружающей среде, энергетике и приложениям (IEEEA '19). Association for Computing Machinery, New York, NY, USA: 194-199. DOI. <https://doi.org/10.1145/3323716.3323735>.

Santamaria-Bonfil G, Reyes-Ballesteros A, Gershenson C, Прогнозирование скорости ветра для ветропарков: A method based on support vector regression, Renewable Energy, 85, 2016: 790-809, ISSN 0960-1481, DOI. <https://doi.org/10.1016/j.renene.2015.07.004>.

TOWARDS AN AI STRATEGY IN MEXICO: Harnessing the AI Revolution, White Paper Report, British Embassy in Mexico through the Prosperity Fund, Oxford Insights and C Minds, June 2018, p. 23. Электронная консультация на сайте <https://www.oxfordinsights.com/mexico> (доступ получен 31 октября 2020 года). Wind Power Forecasting using Artificial Intelligence Tools, Ingenieria,

Investigación y Tecnología, 2018, 19-4: 1-11. DOI. <http://dx.doi.org/10.22201/fi.25940732e.2018.19n4.033>

Yu Z, Han S, Zhu J, Sun X, Ji C, Guo X. Индекс массы тела до беременности в связи с весом новорожденных и избыточным весом/ожирением потомства: систематический обзор и мета-анализ. PloS one. 2013;8:e61627. DOI. <https://doi.org/10.1371/journal.pone.0061627> and Weng SF, Redsell SA, Swift JA, Yang M, Glazebrook CP. Systematic review and meta-analyses of risk factors for childhood overweight identifiable during infancy. Arch Dis Child. 2012;97:1019-26. DOI. <https://doi.org/10.1136/archdischild-2012-30>

Совет Европы – ведущая правозащитная организация континента. В него входят 47 государств-членов, включая всех членов Европейского Союза. Все государства-члены Совета Европы подписали Европейскую конвенцию по правам человека – договор, призванный защищать права человека, демократию и верховенство закона. Европейский суд по правам человека следит за выполнением Конвенции в странах-членах.

Недобросовестное использование и злоупотребление искусственным интеллектом



Информация, представленная здесь, предназначена только для общих информационных и образовательных целей. Это не предназначено и не должно толковаться как юридическая консультация. Информация, содержащаяся в настоящем документе, может быть применима не ко всем ситуациям и может не отражать самую текущую ситуацию. Ни на что, содержащееся в настоящем документе, не следует полагаться или действовать без юридической консультации, основанной на конкретных представленных фактах и обстоятельствах, и ничто в настоящем документе не должно толковаться иначе. Trend Micro оставляет за собой право изменять содержание данного документа в любое время без предварительного уведомления.

Переводы любых материалов на другие языки предназначены исключительно для удобства. Точность перевода не гарантируется и не подразумевается. Если возникают какие-либо вопросы, связанные с точностью перевода, пожалуйста, обратитесь к официальной версии документа на языке оригинала. Любые расхождения или различия, возникшие в переводе, не

являются обязательными и не имеют юридической силы для целей соблюдения или обеспечения соблюдения.

Хотя Trend Micro прилагает разумные усилия для включения в настоящий документ точной и актуальной информации, Trend Micro не дает никаких гарантий или заверений любого рода в отношении ее точности, актуальности или полноты. Вы соглашаетесь с тем, что доступ к настоящему документу и его содержанию, а также использование и доверие к нему, осуществляется на ваш собственный риск. Trend Micro отказывается от всех гарантий любого рода, явных или подразумеваемых. Ни Trend Micro, ни любая сторона, участвующая в создании, производстве или доставке этого документа, не несет ответственности за любые последствия, убытки или ущерб, включая прямые, косвенные, специальные, упущенную выгоду или специальные убытки, возникшие в результате доступа, использования или невозможности использования или в связи с использованием этого документа, а также за любые ошибки или упущения в его содержании. Использование этой информации означает согласие на использование в условиях «как есть».

ОТКАЗ ОТ ОТВЕТСТВЕННОСТИ ЮНИКРИ

Мнения, выводы, заключения и рекомендации, изложенные в настоящем документе, не обязательно отражают мнения ЮНИКРИ или любого другого вовлеченного национального, регионального или международного органа.

Используемые обозначения и материалы, представленные в настоящей публикации, не подразумевают выражения какого-либо мнения со стороны Секретариата Организации Объединенных Наций относительно правового статуса какой-либо страны, территории, города или района ее властей или относительно делимитации ее границ или границ.

Содержание этой публикации может цитироваться или воспроизводиться при условии, что источник информации признан. ЮНИКРИ хотел бы получить копию документа, в котором используется или цитируется эта публикация.

Появление новых технологий формирует мир во все большем числе секторов. Хотя существует множество способов определения этих технологий и их различных функций, их можно кратко суммировать.^{1, 2, 3} Искусственный интеллект (ИИ) широко распространен в своих приложениях и имеет большие перспективы для решения ряда сложных глобальных проблем. Используя беспрецедентное количество данных, ИИ продемонстрировал потенциал от рекомендаций по навигации и контенту до более точного выявления рака у радиолога-человека.

ИИ – это область компьютерных наук (со многими перекрестными связями с другими дисциплинами), посвященная теории и разработке компьютерных систем, которые могут выполнять задачи, обычно требующие человеческого интеллекта, такие как визуальное восприятие, распознавание речи, перевод между языками, принятие решений и решение проблем. Машинное обучение (МО), само по себе подобласть ИИ, состоит из алгоритмов, которые используют статистические методы, чтобы дать компьютерным системам возможность «учиться» на данных, то есть постепенно повышать производительность конкретной задачи. По сравнению с другими компьютерными программами алгоритмы МО не требуют явных инструкций от людей, а скорее извлекают шаблоны и изучают неявные правила из значительного числа примеров.

В то время как алгоритмы ИИ и МО могут принести огромную пользу обществу, эти технологии также могут обеспечить целый ряд цифровых, физических и политических угроз. Точно так же, как Всемирная паутина выдвинула на первый план множество новых видов преступлений и способствовала ряду более нетрадиционных, ИИ готов сделать то же самое.^{4, 5} В непрерывном переходе от аналогового к цифровому также раскрывается потенциал для злонамеренного использования новых технологий.

Таким образом, хотя в настоящем докладе рассматривается нынешнее состояние технологий ИИ и МО, в нем также делается попытка предсказать возможные способы, которыми преступники будут использовать эти

технологии в будущем — задача, которая, хотя и кажется сложной, имеет первостепенное значение для индустрии кибербезопасности и правоохранительных органов, чтобы всегда оставаться на шаг впереди преступников.

Вступление

Системы искусственного интеллекта (ИИ) – это программные (и, возможно, аппаратные) системы, разработанные людьми, которые, задавшись сложной целью, действуют в физическом или цифровом измерении, воспринимая окружающую среду посредством сбора данных, интерпретируя собранные структурированные или неструктурированные данные, рассуждая на основе знаний или обрабатывая информацию, полученную из этих данных, и принимая решение о наилучшем действии (действиях) для достижения поставленной цели. Системы ИИ могут либо использовать символические правила, либо изучать числовую модель, а также адаптировать свое поведение, анализируя, как на окружающую среду влияют их предыдущие действия. (В ссылке для этого определения используется британское написание «analysing»).

В динамичном мире технологий и компьютерных наук ИИ продолжает предлагать широкий спектр возможных применений для предприятий и частных лиц. К сожалению, обещание более эффективной автоматизации и автономии неотделимо от различных схем, на которые способны злоумышленники.

Например, преступники могут использовать ИИ для облегчения и улучшения своих атак, максимизируя возможности получения прибыли за более короткое время, эксплуатируя новых жертв и создавая более инновационные модели преступного бизнеса, одновременно снижая шансы быть пойманными. Кроме того, по мере того как ИИ как услуга становится все более распространенной, это снизит барьер для входа за счет сокращения навыков и технических знаний, необходимых для использования ИИ.

Преступники и организованные преступные группы (ОПГ) быстро интегрируют новые технологии в свои методы работы, тем самым создавая не только постоянные изменения в криминальном ландшафте во всем мире, но и создавая значительные проблемы для правоохранительных органов и кибербезопасности в целом. Бизнес-модель «Преступление как услуга» (CaaS), которая позволяет нетехнологически подкованным преступникам приобретать технические инструменты и услуги в цифровом подполье, которые позволяют им расширять свои возможности и изощренность атак, еще больше увеличивает потенциал для того, чтобы новые технологии, такие как ИИ, могли быть использованы преступниками и стать движущей силой преступности.

Накопление знаний о потенциальном использовании ИИ преступниками повысит способность индустрии кибербезопасности в целом и правоохранительных органов в частности предвидеть возможные вредоносные и преступные действия, а также предотвращать, реагировать или смягчать последствия таких атак упреждающим образом. Понимание возможностей, сценариев и векторов атаки является ключом к повышению готовности и повышению устойчивости.

В соответствии с целью внести свой вклад в накопление знаний об ИИ, настоящий доклад, подготовленный совместными усилиями Trend Micro, Межрегионального научно-исследовательского института Организации Объединенных Наций по вопросам преступности и правосудия (ЮНИКРИ) и Европола, направлен на тщательное и углубленное изучение нынешнего и возможного будущего злонамеренного использования и злоупотреблений ИИ и связанных с ним технологий.

Целевая аудитория этого доклада включает экспертов и практиков в области кибербезопасности, правоохранительные органы, инновационные центры, политиков и исследователей.

Доклад также может быть использован членами этой аудитории в качестве пространства для размышлений и призыва к большему вниманию к возможному злонамеренному использованию или злоупотреблению ИИ.

Выводы, содержащиеся в докладе, основаны на материалах, представленных тремя организациями. Они были объединены с материалами, собранными в ходе целенаправленного семинара в марте 2020 года, который был организован Европолом, Trend Micro и ЮНИКРИ. Среди участников семинара были члены Совместной целевой группы по борьбе с киберпреступностью (J-CAT), Международного уголовного суда и несколько членов консультативных групп Европейского центра по борьбе с киберпреступностью Европола (ЕСЗ).

В настоящем докладе используется важное различие, которое следует отметить с самого начала, а именно различие между злонамеренным использованием и злоупотреблениями ИИ. «Использование» во «вредоносном использовании ИИ» относится к случаям, когда преступники могут использовать системы ИИ для достижения своих целей атаки — например, используя мл для автоматизации кибератак и улучшения вредоносных программ. С другой стороны, «злоупотребления» в «злонамеренных злоупотреблениях ИИ» относятся к случаям, когда преступники могут попытаться атаковать и использовать существующие системы ИИ, чтобы взломать или обойти их — например, путем взлома помощников по умному дому.

Машинное обучение (мл) – это подмножество ИИ, в котором алгоритмы обучаются выводить определенные закономерности на основе набора данных, чтобы определить действия, необходимые для достижения заданной цели. Имея в виду это различие, мы составили отчет вокруг двух основных компонентов: настоящее вредоносное использование или злоупотребление ИИ, в отношении которого имеются задокументированные случаи и исследования, и будущее вредоносное использование или злоупотребление, в отношении которого пока нет доказательств или литературы. Тем не менее, мы считаем, что на основе технологических тенденций и разработок будущие виды использования или злоупотребления могут стать реальностью в недалеком будущем. Чтобы подготовиться к такому использованию или злоупотреблениям, спекулятивные

сценарии, которые, вероятно, произойдут, должны быть концептуализированы. Чтобы обосновать доклад, в связи с этим, мы изучаем текущие тенденции на подпольных форумах, поскольку они могут дать представление о том, как может выглядеть злонамеренное злоупотребление угроз ИИ в ближайшем будущем.

Что касается нынешнего злонамеренного использования или злоупотреблений, то здесь также определены возможные контрмеры для борьбы со злонамеренным использованием ИИ. Однако следует подчеркнуть, что эти контрмеры не являются исчерпывающими и лишь указывают на один или несколько возможных путей борьбы с выявленным конкретным использованием или злоупотреблением.

Наконец, в рамках тематического исследования в конце этого доклада мы более подробно рассмотрим deepfakes, конкретное использование ИИ, которое получило широкое внимание средств массовой информации и представляет значительный потенциал для целого ряда вредоносных и преступных целей. Мы также включаем обзор технологических аспектов некоторых конкретных случаев использования, которые определены в отчете, наряду с дальнейшим техническим содержанием в этом тематическом исследовании.

Нынешнее состояние злонамеренного использования и злоупотребления ИИ

В последние годы наблюдается значительный скачок в развитии ИИ. В то время как в этой статье описываются возможные будущие события злонамеренного использования и злоупотребления технологией ИИ, подобное предсказание имеет ценность только тогда, когда оно должным образом обосновано в настоящем.

С этой целью в этом первом разделе представлены текущие разработки ИИ в области возможного злонамеренного использования и злоупотреблений, охватывающие результаты исследований, доказательства концепции или обсуждения среди преступников, при условии, что они задокументированы и поддаются проверке. Этот раздел включает в себя как случаи использования

преступниками технологий ИИ для собственной выгоды, так и преступные попытки злоупотребления законными системами ИИ, которые легко доступны для злонамеренных злоупотреблений ИИ.

Вредоносное ПО ИИ

Использование ИИ для повышения эффективности вредоносных программ все еще находится в зачаточном состоянии. Исследования по-прежнему проводятся на академическом уровне и атаки в основном носят теоретический характер и разрабатываются в качестве доказательств концепции исследователями безопасности. Тем не менее, методы кибератак с поддержкой ИИ или с улучшением ИИ, которые были изучены, являются доказательством того, что преступники уже предпринимают шаги по расширению использования ИИ. Поэтому такие попытки требуют дальнейшего наблюдения, чтобы остановить эти текущие попытки и подготовиться к будущим атакам как можно раньше, прежде чем они станут основными.

В настоящее время разработчики вредоносных программ могут использовать ИИ более запутанными способами, не будучи обнаруженными исследователями и аналитиками. Как следствие, можно только искать наблюдаемые признаки, которые можно было бы ожидать от вредоносной активности ИИ. На самом деле, один из типов, связанных с вредоносными программами эксплойтов ИИ включает в себя методы на основе ИИ, направленные на повышение эффективности «традиционных» кибератак.

В 2017 году во время конференции по информационной безопасности Black Hat USA исследователи продемонстрировали, как использовать методы ИИ для анализа многолетних данных, связанных с атаками на компрометацию деловой электронной почты (BEC), формой киберпреступности, которая использует мошенничество с электронной почтой для мошенничества организаций с целью выявления потенциальных целей атаки. Эта система использует как утечки данных, так и общедоступную информацию в социальных сетях. Примечательно, что, основываясь на своей истории, система может точно предсказать, будет ли атака успешной или нет.

На той же конференции по безопасности исследователи представили AVPASS – инструмент, предназначенный для определения для любого конкретного антивирусного движка его функций обнаружения и цепочки правил обнаружения. Затем инструмент использует этот вывод, чтобы замаскировать вредоносное ПО Android как доброкачественное приложение. Следует подчеркнуть, что AVPASS достигла 0%-ного уровня обнаружения в онлайн-сервисе анализа вредоносных программ VirusTotal с более чем 5000 образцами вредоносных программ для Android. Другими словами, AVPASS создал оперативно необнаруживаемое вредоносное ПО.

В настоящее время поставщики антивирусов также рассматривают мл как свой инструмент для улучшения методов обнаружения вредоносных программ, благодаря его способности обобщать новые типы вредоносных программ, которые никогда раньше не встречались. Однако было доказано, что системы обнаружения на основе мл могут быть обмануты агентом ИИ, предназначенным для зондирования и поиска слабых мест. Например, исследователей смогли создать вредоносное ПО с функциями, которые позволяют ему оставаться незамеченным даже антивирусными системами на основе мл. Система использует обучение с подкреплением для разработки соревновательной игровой техники между собой и антивирусным детектором. Он также выбирает функции, сохраняющие функциональность вредоносного файла Windows, и вводит варианты, которые увеличивают вероятность того, что образец вредоносного ПО пройдет незамеченным.

Обучение с подкреплением относится к обучению, которое проходит агент, чтобы научиться определенному поведению в динамической среде. Поскольку агент не ориентирован заранее на различные результаты для каждого действия, которое он может предпринять, он должен полагаться как на знания, которые он приобретает через опыт, так и на вознаграждения, которые он получает за каждое действие, которое он предпринимает, чтобы научиться поведению.

Наконец, ИИ также может улучшить традиционные методы взлома, вводя новые способы выполнения атак, которые людям было бы трудно предсказать. На DEF CON 2017, одной из крупнейших подпольных хакерских конвенций, участники Дэн Петро и Бен Моррис представили DeepHack, инструмент ИИ с открытым исходным кодом, предназначенный для проведения тестирования на проникновение в Интернет без необходимости полагаться на какие-либо предварительные знания целевой системы. DeepHack реализует нейронную сеть, способную создавать строки SQL-инъекций без какой-либо информации, кроме ответов целевого сервера, тем самым автоматизируя процесс взлома веб-баз данных.

Следуя аналогичному подходу, DeepExploit – это система, которая способна полностью автоматизировать тестирование на проникновение с помощью мл. Система напрямую взаимодействует с Metasploit, платформой тестирования на проникновение, для всех обычных задач сбора информации, разработки и тестирования эксплойта. Тем не менее, он использует алгоритм обучения с подкреплением, называемый асинхронными агентами акторов-критиков (АСЗ), чтобы сначала узнать (из открыто эксплуатируемых служб, таких как Metasploitable), какой эксплойт следует использовать в определенных условиях, прежде чем тестировать такие условия на целевом сервере.

В ранее описанных атаках ИИ помогает улучшить и оптимизировать методы, традиционно используемые преступниками. И наоборот, второй вид атак фокусируется на подрыве существующих систем ИИ с целью изменения их возможностей или поведения.

Например, исследователи IBM недавно представили новый подход к использованию ИИ во вредоносных целях. Их система DeepLocker встраивает возможности ИИ в саму вредоносную программу, чтобы улучшить ее методы уклонения. Это достигается за счет использования общей слабости большинства алгоритмов ИИ – то есть их недостаточной объяснимости – в интересах системы.

По своей природе алгоритмы мл оказываются, с точки зрения реализации, не более чем умножением факторов и матриц, в то время как массив факторов, составляющих входные данные, умножается на одну или несколько матриц. В свою очередь, значение этих матриц определяется на этапе обучения алгоритма, что приводит к выходному значению указанного алгоритма. Это относится почти ко всем контролируемым алгоритмам машинного обучения, за очень редкими исключениями. Из-за этого, как правило, очень трудно объяснить простым языком, почему алгоритм машинного обучения, обученный с помощью данного набора данных, примет конкретное решение. Для иллюстрации, если мы возьмем пример спам-фильтра, основанного на глубоком обучении для идентификации спам-контента, будет очень трудно объяснить, почему он помечает некоторый контент как спам. Аналогичным образом, было бы сложно отладить и исправить ложное срабатывание, в котором законное сообщение ошибочно помечено как спам.

Вместо этого обычно предполагается, что алгоритм является «черным ящиком» и что проектирование системы вращается вокруг выбора правильного черного ящика, настройки параметров его функционирования и поиска подходящего обучающего набора для данной задачи.

DeepLocker использует эту внутреннюю непостижимость в своих собственных интересах: вредоносное ПО обычно использует несколько методов, чтобы избежать обнаружения, от запутывания полезной нагрузки и шифрования до специальных проверок, чтобы сорвать анализ. Например, при целевых атаках вредоносное ПО часто выполняет определенные проверки, чтобы активироваться только при запуске на целевой машине. Эти проверки, однако, являются самой последней подсказкой, которая может выдать, что часть программного обеспечения является целевой вредоносной программой и может служить в качестве подписи для самой вредоносной программы.

DeepLocker удастся запутать эту последнюю часть вредоносного поведения, реализуя указанные проверки через глубокую нейронную сеть (DNN), таким образом, используя вышеупомянутую «необъяснимую» природу

алгоритма, чтобы замаскировать эти проверки как простые математические операции между матрицами.

В результате становится сложнее перепроектировать поведение вредоносного ПО или даже запустить его в тестовой среде, поскольку условия запуска никогда явно не кодируются во вредоносном ПО. Следовательно, анализ также становится намного сложнее. Например, DeepLocker использует множество атрибутов для идентификации целевой машины, включая программную среду, геолокацию и даже аудио и видео, что усложняет прогнозирование и идентификацию условий запуска вредоносного ПО.

DeepLocker делает запутывание еще на один шаг дальше: Его нейронная сеть не выводит явное значение активации. Он также не отвечает на вопрос: «Нахожусь ли я на целевой машине?» вместо «Да» или «Нет» он выводит ключ расшифровки, который может быть использован, когда условия ввода совпадают, для расшифровки оставшейся полезной нагрузки вредоносного ПО.

Таким образом, любая явная информация о цели вредоносного ПО запутывается в виде матричных факторов, что делает практически невозможным реинжиниринг и идентификацию цели. До того момента, пока вредоносная программа не окажется на цели, остальная часть ее кода полностью зашифрована и непостижима.

Стоит отметить, что такие атаки, как DeepLocker, трудно остановить, так как защитник не будет знать, что искать, пока вредоносное ПО не будет полностью расшифровано. Аналогично на целевой машине, DeepLocker можно остановить, обнаружив атаку только после ее расшифровки и запуска. Таким образом, сигнатура антивируса в памяти, поведенческое правило или правило брандмауэра смогут увидеть атаку по мере ее возникновения. Тем не менее, проблема может быть намного хуже, если вредоносное ПО будет проанализировано на тестовой машине или если инженер решит, является ли набор данных вредоносным. В таких случаях DeepLocker не будет правильно расшифровываться, и в результате инженер не сможет увидеть вредоносные функции для надлежащего анализа. Если вредоносная программа никогда не

будет правильно проанализирована, эти правила в памяти или поведения, таким образом, никогда не будут созданы.

Между тем, такие атаки, как DeepHack и DeepExploit, также трудно обнаружить. Несмотря на это, хорошо настроенный брандмауэр приложений может помочь защитить сеть от этих атак и даже запретить дальнейшие подключения от атаки с поддержкой ИИ.

Вредоносное ПО ИИ в целом

Чтобы проверить, используют ли авторы вредоносных программ уже доступные в настоящее время алгоритмы мл, мы использовали следующий подход:

1. Мы типизировали популярные двигатели мл и создали правила обнаружения, используя YARA,²⁵ отраслевой стандарт для поведенческого обнаружения файлов, таких как вредоносные программы, на основе правил.

2. Мы провели ретроохотливые поиски с этими наборами правил на VirusTotal, чтобы узнать, не были ли какие-либо из этих двигателей каким-либо образом изменены или подделаны.

3. Мы изучили результаты для тех файлов со следами вредоносного кода. Чтобы избежать ложных срабатываний, нам понадобилось по крайней мере два обнаружения от разных антивирусных систем.

Наиболее популярными и легкодоступными двигателями мл являются следующие:

- TensorFlow + Keras
- RapidMiner
- PyTorch

И TensorFlow, и PyTorch – это проекты с открытым исходным кодом, написанные на Python. Поскольку они так легко изменяемы, определенные правила YARA очень универсальны и специально ищут узнаваемые строки в текстовых файлах или исполняемых файлах. RapidMiner, с другой стороны,

является двигателем на основе Java; поэтому соответствующее правило YARA ищет определенные соответствующие строки в файлах архива Java (JAR).

Наш поиск вредоносных файлов, связанных с TensorFlow, RapidMiner или PyTorch, не дал никаких результатов. Первые попытки привели к нескольким файлам, которые позже оказались ложными срабатываниями. После уточнения исходных правил и поиска определенных строк, связанных с каждым из этих фреймворков, мы не получили никаких дополнительных результатов.

В случае TensorFlow и PyTorch правила очень похожи. Мы искали файлы Python (с расширением «.py») или портативные исполняемые файлы «PE» для учета скомпилированных двоичных файлов Python на платформах Windows. Правила, более подробно описанные в Приложении А, разработаны таким образом, чтобы соответствовать определенным строкам, которые были определены в открытом исходном коде соответствующих проектов.

Хотя невозможно с уверенностью доказать, что вредоносного ПО, использующего мл, не существует, поиск вредоносных файлов, похожих на популярные фреймворки мл, на момент написания статьи не дал никаких результатов.

Злоупотребление облачными сервисами ИИ

Важно отметить, что алгоритмы мл не обязательно должны выполняться на той же хост-машине, на которой работает вредоносное ПО. Другой подход к обнаружению того, что злоумышленники уже начали использовать МО при разработке вредоносных программ, заключается в том, чтобы проверить, подключается ли конкретная вредоносная программа к облачным сервисам ИИ.

Чтобы следовать этому подходу, мы составили список всех сервисов ИИ, предлагаемых крупными облачными провайдерами, включая Amazon AWS, Microsoft Azure, Google Cloud, Alibaba Cloud, Яндекс и другие. Все они начали предлагать услуги ИИ в последние годы: например, Amazon через свою платформу AWS предлагает услуги распознавания изображений (Amazon Rekognition), неструктурированную текстовую аналитику (Amazon

Comprehend) или извлечение именованных сущностей (Amazon Textract), а конкуренты предлагают аналогичные собственные услуги. К каждой из этих служб можно получить доступ через HTTPS-соединение с определенным URL-адресом конечной точки, который использует определенный шаблон. Например, все сервисы платформы Amazon AWS содержат «aws.amazon.com» в имени хоста и имени службы в пути.

Используя эти шаблоны, мы провели обширный поиск в базах данных Trend Micro, сосредоточив внимание, в частности, на двух наборах данных:

- Первый набор данных содержит поведенческую информацию об образцах вредоносных программ, которые были проанализированы в прошлом.
- Второй набор данных содержит анонимизированные отчеты о подключениях, выполненных вредоносным программным обеспечением, обнаруженным антивирусным программным обеспечением Trend Micro.

Используя эти шаблоны и сосредоточившись на двух наборах данных, можно было проверить, подключаются ли какие-либо новые образцы вредоносных программ, найденные в Интернете, к облачным сервисам ИИ и почему, или если ранее было обнаружено какое-либо вредоносное подключение к указанным сервисам, исходящее от вредоносного ПО на компьютере жертвы.

На момент написания этого отчета наш поиск не выявил никаких следов вредоносного программного обеспечения, использующего облачные сервисы. Существует две возможные причины такого отсутствия идентификации. Во-первых, масштабирование (с точки зрения злоумышленника) может привести к более высоким затратам. Во-вторых, без надлежащих мер переход на использование облачных сервисов может увеличить вероятность раскрытия злоумышленника.

Тем не менее, облачные сервисы должны отслеживаться на предмет такого рода соединений, поскольку следующая итерация вредоносных программ ИИ, нацеленных на них, все еще может появиться.

Злоупотребление умными помощниками

Альтернативный подход к атаке алгоритмов ИИ заключается в том, чтобы нацелить помощников ИИ либо на использование их присутствия в домашних хозяйствах, либо на злоупотребление их моделью развития. В частности, помощник по ИИ может быть нацелен на разработку конкурирующих систем или загрязнение наборов данных.

Пример состязательной системы подробно описан в статье Trend Micro «Звук целевой атаки», где злоумышленник может использовать открытые интеллектуальные динамики для выдачи звуковых команд ближайшему интеллектуальному помощнику, такому как Amazon Alexa или Google Home. Если динамик подключен к Интернету, его уязвимости могут быть использованы, что приведет к воспроизведению аудиофайла, размещенного на произвольном веб-адресе, заданном злоумышленником. Кроме того, файл может содержать речь, которая выдает определенную команду ближайшим интеллектуальным помощникам. Кроме того, скрытая атака может использовать выданную команду, которая не воспринимается человеческим ухом. Открытые устройства можно легко найти с помощью таких сервисов, как Shodan, поисковая система для подключенных к Интернету устройств и систем.

Возможность таких атак еще более усугубляется тем фактом, что умные помощники часто контролируют системы домашней автоматизации. В статье «Риски кибербезопасности в сложных средах Интернета вещей: угрозы умным домам, зданиям и другим структурам» Trend Micro доказывает, что захват умного помощника с помощью открытых аудиоустройств является лишь одной из многих атак, которые могут быть осуществлены злоумышленником, заинтересованным во взломе умного дома.

Угадывание пароля с помощью ИИ

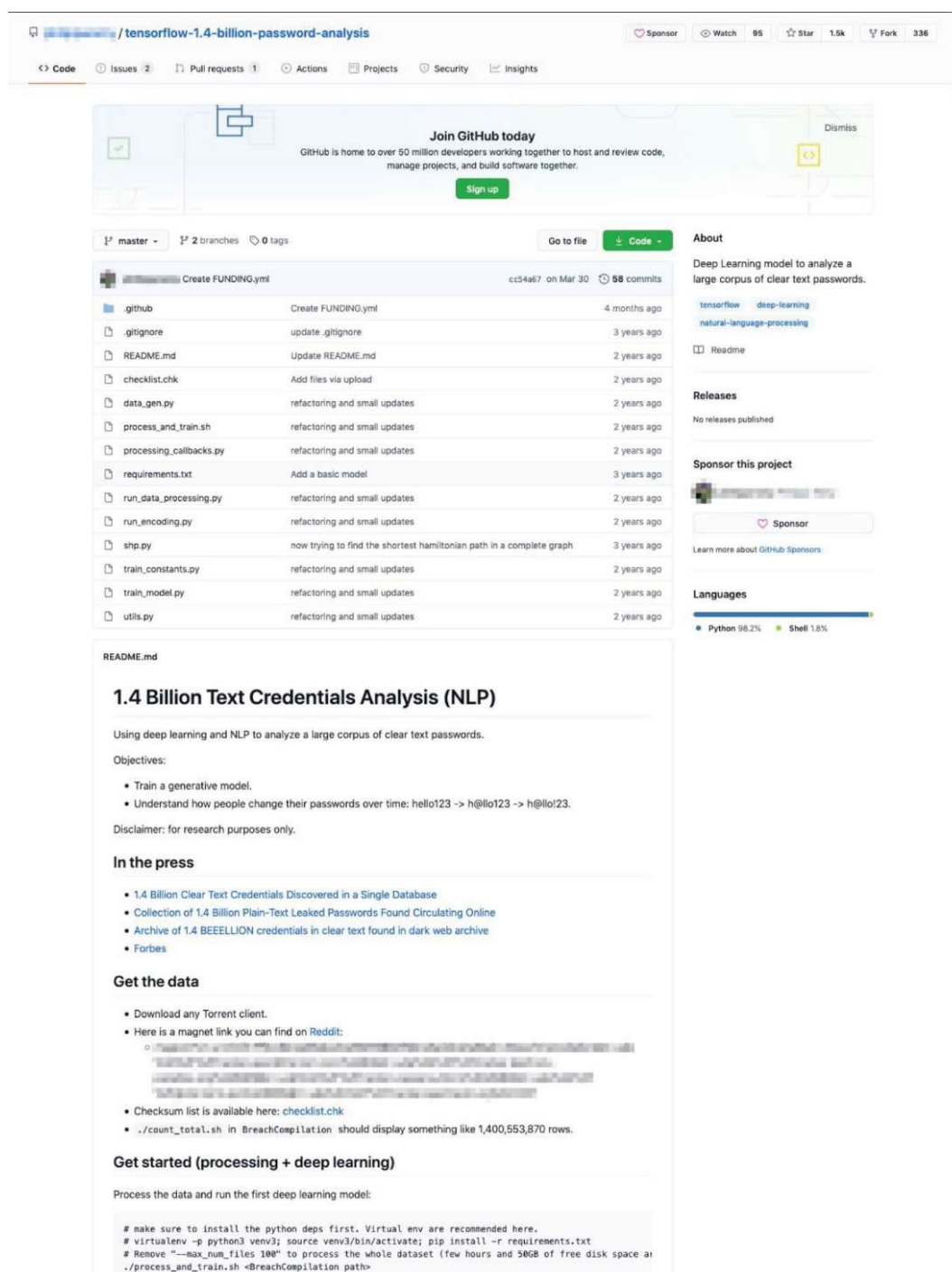
Другое приложение включает в себя использование мл для улучшения алгоритмов подбора паролей. Несмотря на то, что это приложение исследуется с умеренным успехом, важно отметить, что это же приложение уже доказало свою эффективность по сравнению с более традиционными подходами. Его

также довольно трудно обнаружить самостоятельно. В результате нет ничего надуманного в предположении, что алгоритмы и инструменты, поддерживаемые ИИ, постоянно концептуализируются и разрабатываются отдельными лицами или группами, которые могут злоупотреблять ими.

Традиционные инструменты для подбора паролей, такие как HashCat и John the Ripper, обычно работают, сравнивая множество различных вариантов хэша пароля, чтобы определить пароль, соответствующий хэшу. Попытки генерируются из словаря часто используемых паролей; после этого вносятся изменения в зависимости от состава пароля. Например, варианты «пароля» могут быть «password12345» или «p4ssw0rd.»

Благодаря использованию нейронных сетей и, в частности, генеративных состязательных сетей (GANS), можно анализировать большой набор паролей и генерировать вариации, соответствующие статистическому распределению, например, для утечек паролей. Это приводит к более целенаправленным и эффективным угадываниям паролей.

Генеративная агрессивная сеть – это тип нейронной сети, состоящий из двух нейросетевых моделей: генератора и дискриминатора, которые GAN использует для различения реальных образцов и образцов, которые были сгенерированы. Ранняя попытка этого уже очевидна в посте на подпольном форуме от февраля 2020 года. В сообщении упоминается репозиторий GitHub того же года, в котором программное обеспечение способно анализировать 1,4 миллиарда учетных данных и генерировать правила изменения паролей на основе своих результатов.



Система прохождения, опубликованная в 2019 году, представляет логичный подход. Он использует GAN для изучения статистического распределения паролей по утечкам паролей и генерирует высококачественные подборки паролей. Эта система смогла сопоставить на 51-73% больше паролей сравнению только с HashCat.

Взлом капчи с поддержкой ИИ

Применение ИИ для взлома систем безопасности капчи часто рассматривается на криминальных форумах. Изображения капчи обычно используются на веб-сайтах, чтобы помешать преступникам, когда они пытаются злоупотреблять веб-сервисами, особенно когда они пытаются автоматизировать атаки (некоторые попытки включают создание новых учетных записей или добавление новых комментариев или ответов на форумах, среди прочего). Таким образом, разработка систем, которые пытаются взломать изображения капчи, чтобы автоматизировать злоупотребление этими услугами, была бы естественным шагом для киберпреступников.

Программное обеспечение, реализующее нейронные сети для решения капчи, такое как XEvil 4.0, в настоящее время тестируется на криминальных форумах. Более того, утверждалось, что эта нейронная сеть может взламывать системы распознавания людей с помощью капчи на страницах Яндекса. Инструмент использует от 80 до 100 потоков процессора (с процессорами, работающими параллельно), чтобы ускорить решение капчи. Он также рекламируется на российских подпольных форумах и сдается в аренду пользователям за 4000 рублей в неделю (примерно 54 доллара США на момент написания) или 10 000 рублей в месяц (примерно 136 долларов США на момент написания).

Чтобы затруднить взлом капчи и снизить эффективность алгоритмов ИИ, разработчики капчи должны увеличить разнообразие цветовых узоров и форм. Однако, учитывая то, как работают механизмы защиты от капчи, вполне возможно, что это приведет лишь к очень незначительной задержке во времени, которое потребуется преступникам для взлома капчи. За исключением этого, возможно, стоило бы разработать какой-то программный фильтр для обнаружения сценариев или программ, которые постоянно пытаются решить капчу слишком быстро. Этот программный фильтр также может быть разработан для отправки полностью поддельных, неразрешимых изображений в эти сценарии или программы, чтобы попытаться отравить их наборы данных.

BotmasterLabs
Dear users!

A very important update to the beta version of XEvil 4.0 [Beta-26] took place today, bringing the program closer to the official release. All beta test participants can download the update in their account at [BotmasterRu.Com](#).

Changes and improvements in XEvil 4.0 [Beta-26]

1. The recognition of two-condition Yandex-captcha has been improved several times

The neural network processing module has been updated, the dictionary and the set of letters / symbols have been expanded, a new teaching method has been used. Yandex CAPTCHA is case sensitive and also requires entering punctuation marks and special characters in addition to letters and numbers - so this task turned out to be very nontrivial, but we managed to get a very good result:

IMAGE 

At the same time, even where XEvil makes small mistakes - Yandex accepts answers, tk. allows you to make a mistake in one word / letter.

Also, XEvil is trained to recognize large Yandex-captcha of standard size 600x180 (authorization), so the new version of XEvil will be indispensable when working with all Yandex services:

- yandex.ru (search results)
- mail.yandex.ru
- wordstat.yandex.ru
- market.yandex.ru
- webmaster.yandex.ru
- and others

Including, recognition of Yandex captcha is very important in the work with behavioral cheating (through search results).

2. Improved processing of ReCaptcha-2, expanded settings

A number of options from hidden settings have been moved to the interface of the RModule window, in addition, the statistics block now displays the average and median time of a successful solution of ReCaptcha-2.

We also did some work on optimization of Google Chromium parameters, which further reduced the time for a successful solution of Recaptcha-2 "to a tick": **Recaptcha-2 recognition became much faster**

3. Added the option of proxy rotation in RModule

IMAGE 

With this option, each request (SiteKey + SiteURL) will assign its own proxy, each time a new one. Without this option, one best proxy is saved within one stream.

ATTENTION! Turn on the rotation option only if you are using very high-speed, high-quality proxies with extremely low connection times. Otherwise, **switching on the result will not improve, but worsen** - because of the need to re-create the connection each time.

Technical details:
code:

- + At times improved recognizability two-vocable Yandex-CAPTCHA (Yandeks.Poisk, Yandeks.Wc)
- + By optimizing parameters decreased mean time solutions reCAPTCHA-2
- + obtaining improved low-level function proxy list for a URL RModule
- + proxy improved extraction mechanism from the proxy list by URL, if it is in HTML format
- + improved support for the DeathByCaptcha API (processing requests in the "www-urlencoded" improvements to the DeathByCaptcha API continue)
- + the RotateProxies option has been added to RecapModule.ini, when enabled, for each request within the API, for setting them when processing Google ReCaptcha-2:
- + is assigned each time new proxy (relevant for those who use elite ultra-fast proxies)
- + these parameters in XEvil.ini are now enabled by default, they affect the reception of
 - AcceptAPI_Cookie
 - AcceptAPI_Proxy
 - AcceptAPI_UA

In the near future, we plan to work on Google ReCaptcha-3, as well as to improve the accuracy of the general neural network used for regular captchas.

For all questions, please contact:

[REGISTER]  and BUY XRUMER + XEVIL 

03 OCT 2019 15:57:00

Подпольный инструмент для взлома CAPTCHA под названием XEvil

Шифрование с помощью ИИ

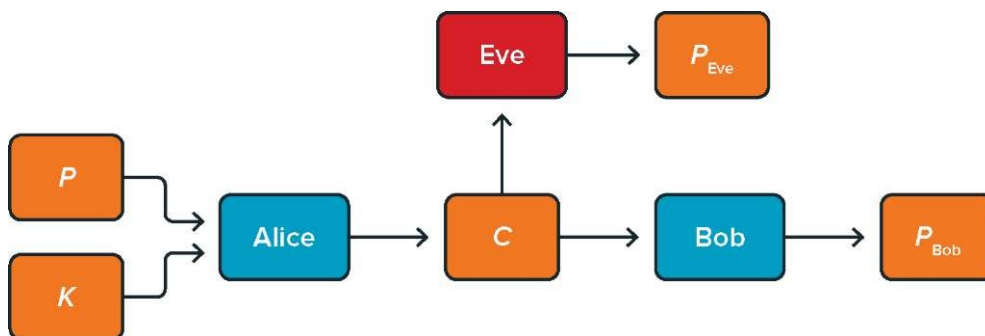
Согласно отчету Европола за 2019 год, названному «Первый отчет о функции обсерватории по шифрованию»,³⁴ приложения ИИ, связанные с улучшением или взломом шифрования, все еще находятся в зачаточном состоянии. Однако, крайне важно узнать больше о шифровании с помощью ИИ, поскольку в этой области уже существует значительный потенциал. Эксперимент, проведенный Google в 2016 году,³⁵ является убедительным доказательством этого потенциала.

В этом эксперименте две нейронные сети по имени Алиса и Боб были обучены общаться друг с другом, не позволяя третьей нейронной сети по имени Ева подслушивать их общение.

Нейронная сеть Алиса получила открытый текст P в качестве входных данных, которые она должна была зашифровать в зашифрованный текст C , который затем был получен нейронной сетью Боб и подделан нейронной сетью Ева. У Алисы и Боба был общий ключ расшифровки K , который Боб использовал для расшифровки зашифрованного текста в открытый текст P_{Bob} , в то время как Еве пришлось реконструировать открытый текст P_{Eve} без помощи ключа расшифровки K . Конкретный алгоритм шифрования не был указан; вместо этого сети должны были выяснить, как безопасно общаться самостоятельно. Результаты показали, что Алиса и Боб смогли научиться выполнять формы шифрования и дешифрования автономно. Кроме того, они научились избирательно применять шифрование только к тем частям данных, которые были необходимы для достижения поставленных целей конфиденциальности.

Этот эксперимент Google демонстрирует еще одно использование ИИ, которое трудно не только обнаружить, но и защитить от него. С внешней точки зрения не очевидно, что используется ИИ, и поэтому Еве (или любому подслушивающему) будет трудно — если не невозможно — собрать воедино исходное сообщение.

Таким образом, дешифрование с помощью ИИ имеет высокий потенциал для включения в арсенал злоумышленников для защиты их коммуникаций и инфраструктуры С2.



Модель коммуникации в исследовательской работе Google.

В 2017 году исследователи из Стэнфорда и Google продемонстрировали, как нейронные сети могут использовать стеганографию. В своем исследовании исследователи обнаружили, что нейронная сеть под названием CycleGAN научилась обманывать при переводе изображения в изображение. В процессе перевода спутникового изображения на карту исследователи предположили, что аэрофотосъемка не будет слишком близко напоминать исходное изображение, поскольку они ожидали, что некоторые особенности, вероятно, будут потеряны при переводе CycleGAN. Однако исследователи все же обнаружили эти особенности в воздушной реконструкции, произведенной нейронной сетью. В этот момент они поняли, что CycleGAN научился копировать аспекты первичного изображения в шумовые паттерны карты. При этом нейронная сеть могла бы скрыть некоторую информацию для себя — ту же информацию, которая ей понадобится для преобразования карты в восстановленное изображение со спутника, — чтобы она могла произвести реконструкцию, которая была бы удивительно похожа на оригинал.

Стоит отметить, как это сделали исследователи в данном случае, что способность нейронных сетей внедрять информацию является уязвимостью, которая может быть полезна для злоумышленников. Эта возможность также была отмечена преступным использованием инициативы сокрытия информации или подачи сообщений в рамках Европола.³⁷ Путем внесения изменений в первичный образ злоумышленники могли создавать свой собственный образ в соответствии со своими схемами. В то же время разработчики, которые не проявляют осторожности при работе с нейронными сетями, могут не знать, что их личная информация внедряется теми же нейронными сетями.

Тенденции, обнаруженные на подпольных форумах

Тенденции на подпольных форумах раскрывают возможный ландшафт злонамеренного злоупотребления ИИ в ближайшем будущем.

Чтобы проследить за актуальными тенденциями обсуждения, мы проанализировали наиболее популярные темы, связанные с ИИ, МО и

нейронными сетями на подпольных форумах. Хотя количество тем пока не является значительным, есть некоторые специфические приложения, которые заслуживают упоминания.

Олицетворение человека на платформах социальных сетей

Один тип приложений включает интеллектуальные системы, которые способны имитировать поведение человека, чтобы обмануть системы обнаружения ботов на платформах социальных сетей. Одной из основных целей этих интеллектуальных систем является монетизация с помощью мошенничества песен в таких сервисах, как Spotify. В системах есть армия пользователей-ботов, которые потребляют указанные песни, сохраняя при этом человеческую модель использования и, таким образом, генерируя трафик для конкретного исполнителя.

На аналогичной ноте обсуждение на форуме `blackhatworld[.]com` рассказывает о возможности разработки бота Instagram для создания учетных записей как поддельных подписчиков, так и детей, создания лайков, запуска последующих ссылок и т.д., как показано на следующем скриншоте. В этом случае предполагается, что ИИ можно использовать для преодоления капчи, имитации движений устройства, таких как выбор или перетаскивание, и управления браузерами в человеческом поведении.

На форуме `nulled[.]to`, поддерживаемый ИИ бот Spotify ([https://spotify\[.\]private-cdn\[.\]live](https://spotify[.]private-cdn[.]live)) рекламируется как имеющая возможность выдавать себя за нескольких пользователей Spotify, параллельно используя несколько прокси-серверов, чтобы избежать обнаружения и запрета системой. Он также может генерировать трафик и выбирать целевые песни, чтобы увеличить их количество воспроизведения и улучшить монетизацию. Чтобы монетизировать это, система ИИ используется для создания списков воспроизведения с различными песнями, чтобы лучше выдавать себя за пользователя. Выбор песни соответствует некоторым человеческим музыкальным вкусам, а не случайному выбору песен, которые могут

предупредить компанию о том, что объект, стоящий за выбором, является ботом.

The screenshot displays the Spotify Bot website, which promotes a service for automating Spotify account activities. The main heading is "SKY ROCKET YOUR SPOTIFY MULTI-THREADED BOT". Below this, the "FEATURES" section lists several capabilities:

- ADD FOLLOWERS TO ANY ARTIST:** "Add any amount of followers to your artist account. Specify an amount and an artist link to add your followers. It's required to use your own account link." An example shows adding 322,851,675 followers to Billie Eilish's account.
- ADD PLAYS TO ANY SONG:** "Add any amount of plays to your songs. Specify an amount and a song link to add your plays. It's required to use your own account link." An example shows adding 688,479,437 plays to a song.

The "PRICE" section offers three subscription tiers:

Spotify BOT Basic	Spotify BOT Pro	Spotify BOT Business
\$25 /mo	\$100 /mo	\$180 /mo
24/7 Bot runs 1000 plays 1000 followers 1000 likes 1000 reposts 1000 shares 1000 comments 1000 retweets 1000 mentions 1000 tags 1000 emojis 1000 stickers 1000 GIFs 1000 videos 1000 images 1000 audio 1000 text 1000 links 1000 mentions 1000 tags 1000 emojis 1000 stickers 1000 GIFs 1000 videos 1000 images 1000 audio 1000 text 1000 links	24/7 Bot runs 1000 plays 1000 followers 1000 likes 1000 reposts 1000 shares 1000 comments 1000 retweets 1000 mentions 1000 tags 1000 emojis 1000 stickers 1000 GIFs 1000 videos 1000 images 1000 audio 1000 text 1000 links 1000 mentions 1000 tags 1000 emojis 1000 stickers 1000 GIFs 1000 videos 1000 images 1000 audio 1000 text 1000 links	24/7 Bot runs 1000 plays 1000 followers 1000 likes 1000 reposts 1000 shares 1000 comments 1000 retweets 1000 mentions 1000 tags 1000 emojis 1000 stickers 1000 GIFs 1000 videos 1000 images 1000 audio 1000 text 1000 links 1000 mentions 1000 tags 1000 emojis 1000 stickers 1000 GIFs 1000 videos 1000 images 1000 audio 1000 text 1000 links

Additional features highlighted include:

- HUMAN BEHAVIOUR USING A.I.:** "Using A.I. we simulate human behaviour to all your accounts list so that it can appear as a real person was using the account." (Illustrated with a green brain icon).
- NOTIFICATIONS USING A VIRTUAL ASSISTANT:** "With this feature we improve your experience and usability throughout our software, making it much easier to use our BOT." (Illustrated with a blue microphone icon).
- CUSTOM PROXY LIST:** "You can use your custom proxy list to achieve more quality in your streams and follows." (Illustrated with a green world map icon).

Информация о ценах на Spotify

Чит-коды для онлайн-игр

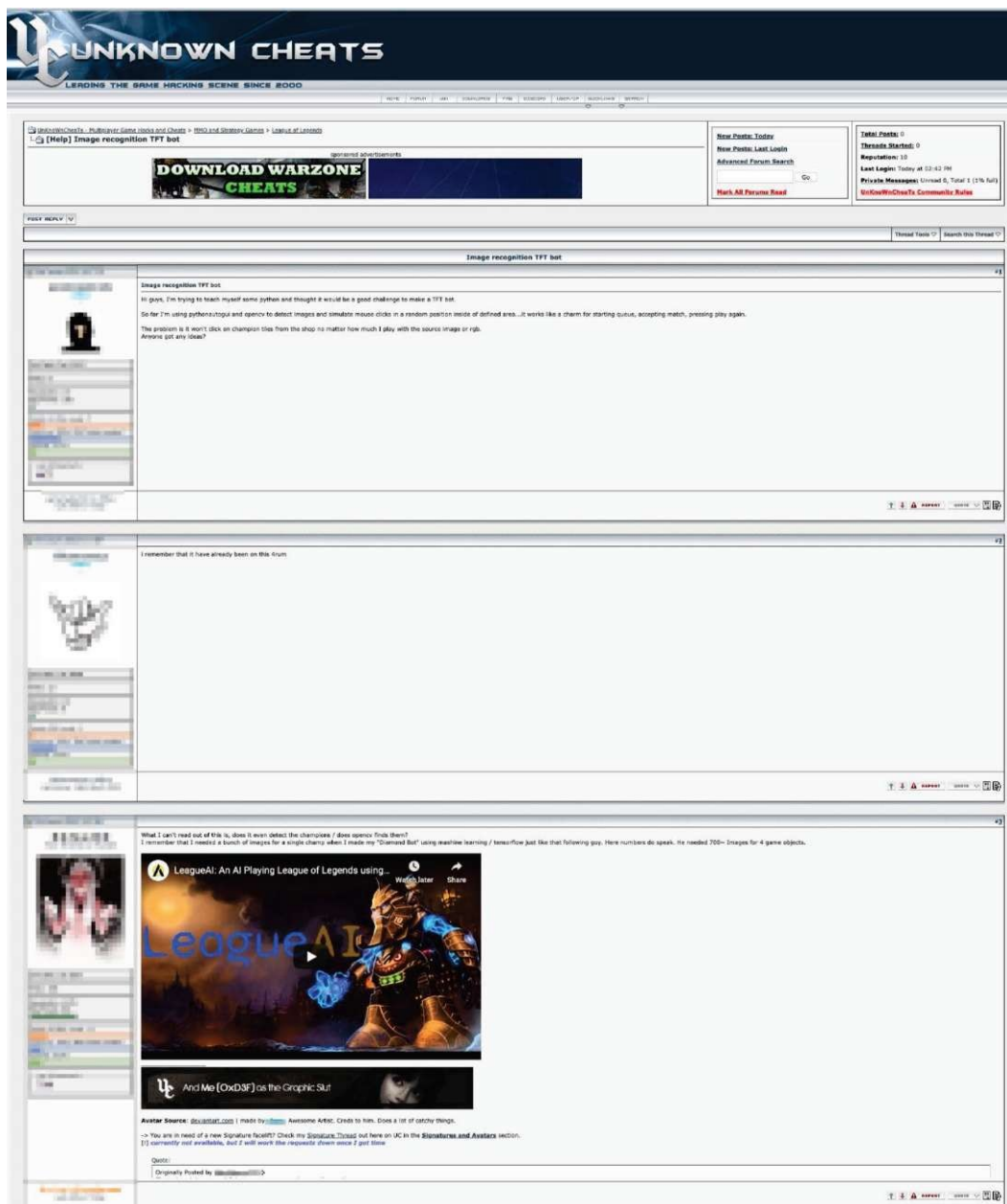
ИИ также используется в области игровых читов, особенно для игр, где монетизация существенна. Как подробно описано в статье Trend Micro под названием «Читы, взломы и кибератаки: угрозы киберспортивной индустрии в 2019 году и далее», рост индустрии киберспорта означает, что она также становится все более интересной целью для возможных атак.

Кроме того, эти атаки выходят за рамки простых игр для хобби. На самом деле они ориентированы на профессиональный и прибыльный киберспорт. Поскольку соревнования по онлайн-играм включают значительные суммы

призовых, они являются основной мишенью для мошеннических атак, в которых используются новейшие доступные технологии. На самом деле преступники уже используют онлайн-игры в целях отмывания денег.

На нескольких различных форумах часто встречаются сообщения о разработке «aimbots» (программного обеспечения, которое использует алгоритмы компьютерного зрения для автоматического наведения на цели). Эти боты используют алгоритмы глубокого обучения, чтобы быть способными не только играть в такие игры, как League of Legends, но и разрабатывать читы для таких игр.

Как свидетельствуют следующие скриншоты, исследование, проведенное на подпольных форумах, раскрывает потенциал использования алгоритмов ИИ для получения прибыли от киберспорта.



Сообщение на форуме о жульничестве ИИ в киберспорте

Взлом с поддержкой ИИ

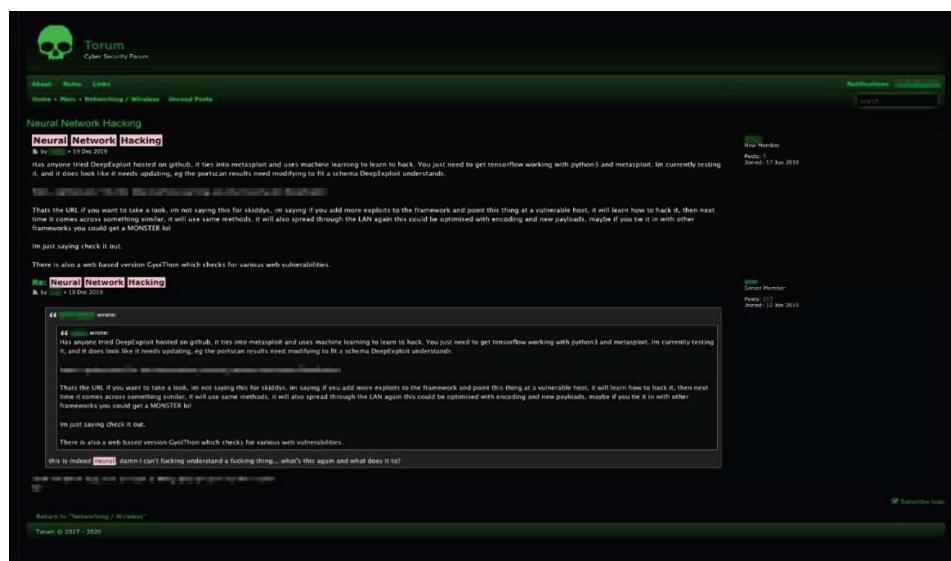
Несколько дискуссий на подземных форумах дают дополнительную информацию о практическом использовании таких фреймворков, как DeepHack и DeepExploit.

В Торуме, форуме darknet, один пользователь утверждает, что ищет информацию о том, как успешно использовать DeepExploit и как взаимодействовать с Metasploit.

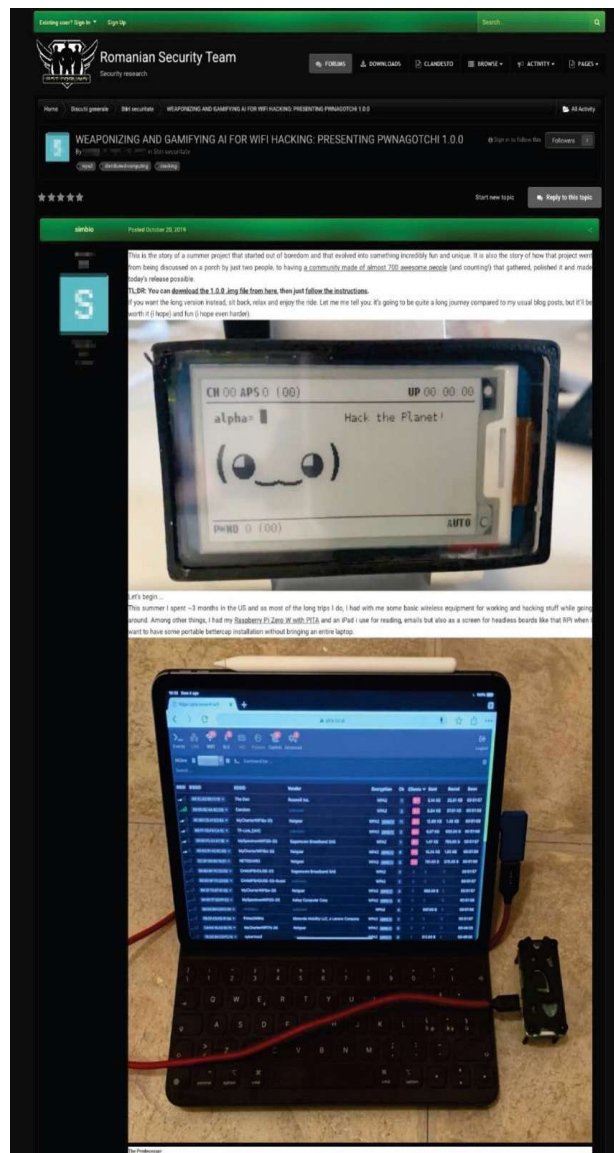
Как упоминалось ранее, инструменты для подбора паролей с поддержкой ИИ также уже используются хакерами и другими преступниками. Например, на

форуме cracked[.]to в посте, в котором перечислена коллекция хакерских инструментов с открытым исходным кодом, содержится программное обеспечение на основе ИИ, способное анализировать большой набор паролей, восстановленных в результате публичных утечек. Инструмент способен обучить GAN, который понимает, как люди склонны изменять и обновлять пароли (например, изменяя «hello123» на «h@llo123», а затем на «h@llo!23»), что значительно помогает в оптимизации инструментов для подбора паролей.

Тем временем в дискуссионной теме на rstforums[.]com обсуждается инструмент под названием «PWnagotchi 1.0.0.» Последняя версия этого инструмента, которая первоначально была разработана для выполнения взлома Wi-Fi с помощью атак деаутентификации, основана на модели нейронной сети для оптимизации ее производительности с использованием стратегии геймификации. Благодаря этой стратегии система вознаграждается за каждую успешную попытку деаутентификации. В результате система, основанная на стимулах, заставляет инструмент автономно улучшать свои показатели.



Пользователь на форуме, задающий вопрос об использовании DeepExploit



Пост описания PWNagotchi 1.0.0 Торговля криптовалютами с поддержкой ИИ

Инструменты на основе ИИ также доступны в то время, когда назрела необходимость разработки новых схем монетизации для злонамеренного использования. Чтобы проиллюстрировать, хотя инструменты на основе ИИ для финансовой торговли известны уже некоторое время, форум в [blackhatworld\[.\]com](http://blackhatworld[.]com) содержит ссылку на ботов с поддержкой ИИ, которые специально предназначены для торговли криптовалютами. Боты, упомянутые на этом форуме, могут изучать успешные торговые стратегии на основе исторических данных и применять их для лучшего прогнозирования более прибыльных сделок. Примечательно, что, хотя сообщение датируется апрелем 2019 года, и с тех пор соответствующий репозиторий GitHub не проявлял

большой активности, все еще показательно, что криптовалюта является прибыльной возможностью для получения финансовой выгоды преступниками.

SEO Форум Blackhat

«Обновление учетной записи участников Партнерств

Следуйте За Мо, Когда Я Создаю AI-Торгового Бота Для Bitcoin I Crypto
«Е

Использует в основном прямолинейную алгебру. Я создал несколько прогнозистов и продвинутую машину, которая позволила им решить проблему торговли. Криптография очень изменчива. Именно по этой причине бот должен быть в

У меня действительно есть некоторые собственные усилия, но я подумал про себя: «Кто-то должен поставить sc

- CCXT для данных OHLCV и торговых действий.
- Websockets? Разберусь в этом
- Большая часть объема, вероятно, находится на бирже Bitmex. Давайте

воспользуемся этим

- Это, конечно, должно быть сделано на Python.
- Давайте начнем с некоторых вариантов LSTM. Может быть, дифференцируемый нейронный компьютер? хмм.

• Как насчет агента оптимизатора? Q Ученик? Генетический? Байесовский? Давайте попробуем.

- Создайте репо на Github. Это будет в первую очередь.

• * Лаборатория Google Co может помочь всем следовать за вами. Не у всех есть графический процессор Titan, как

в этом путешествии, даже я не увлекаюсь трендами и криптографией.»

Социальная инженерия

По мере развития технологий ИИ развиваются и различные схемы социальной инженерии, которые могут извлечь выгоду из такой технологии.

Это важно отметить, поскольку Европол отметил в 2020 году, что социальная инженерия по-прежнему является главной угрозой, которая используется для содействия другим формам киберпреступности.⁴⁴ В подтверждение этих схем на различных подпольных форумах было обнаружено несколько интересных дискуссий, связанных с инструментами, использующими ИИ для улучшения задач социальной инженерии.

На форуме French Freedom Zone было заявлено, что инструмент разведки под названием «Eagle Eyes» способен находить все учетные записи в социальных сетях, связанные с определенным профилем. Соответственно, инструмент может даже сопоставлять профили с разными именами, сравнивая фотографии профиля пользователя с помощью алгоритмов распознавания лиц.

GitHub является домом для более чем 40 миллионов разработчиков, работающих вместе, чтобы размещать и просматривать код, управлять проектами и создавать программное обеспечение вместе.

Преследуй своих друзей. Найдите их профили в Instagram, FB и Twitter, используя распознавание изображений и обратный поиск изображений.

У вас есть по крайней мере одно изображение человека, которого вы ищете, и подсказка об его имени. Instagram, Facebook, Youtube и Twitter-профили этого человека. Вы вводите эти данные в Eagle Eye, и он пытается найти профили этого человека в Instagram, Youtube, Facebook и Twitter.

Кроме того, сообщение на rstforums[.]com рекламирует инструмент, способный выполнять клонирование голоса в реальном времени. С помощью всего лишь пятисекундной записи голоса цели злоумышленник уже может клонировать свой голос.

Клонирование Голоса В Реальном Времени

Этот репозиторий представляет собой реализацию обучения передаче от проверки динамиков к синтезу текста в речь с несколькими динамиками (SV2TTS) с вокодером, который работает в режиме реального времени. Не стесняйтесь проверять тезисы, если вам интересно или если вы ищете

информацию, которую я не задокументировал (не стесняйтесь создавать проблемы и для этого), В основном я бы рекомендовал быстро взглянуть на цифры за пределами введения.

SV2TTS-это трехэтапная система глубокого обучения, которая позволяет создать числовое представление голоса из нескольких секунд аудио и использовать его для создания модели преобразования текста в речь, обученной обобщать на демонстрацию новых голосов:



Реклама инструмента для клонирования голоса

Будущие сценарии злонамеренного использования и злоупотребления ИИ

Развитие новых технологий, включая системы ИИ, способствует повышению автоматизации, что может привести к делегированию контроля на некоторых уровнях общества.

В дополнение к постепенной интеграции методов ИИ для расширения масштабов и масштабов своих кибератак, киберпреступники могут использовать ИИ как вектор атаки, так и поверхность атаки. Благодаря ранее описанной модели криминального бизнеса, основанной на услугах, разработка и распространение наборов инструментов с поддержкой ИИ на подпольных рынках может предоставить менее технологически подкованным преступникам расширенные возможности.

Этот второй раздел предлагает всесторонний обзор, основанный на современном опыте, и исследует вероятные сценарии, в которых преступники могут злоупотреблять технологиями ИИ для облегчения своей деятельности. Наряду с этим злоупотреблением мы ожидаем, что киберпреступники будут постоянно использовать методы ИИ, чтобы скрыть свою незаконную деятельность в Интернете и избежать обнаружения.

Другие авторы предложили интересный ракурс для анализа этой темы. С этой точки зрения, если рассмотреть различные фазы в анатомии кибератаки, можно было бы теоретизировать как киберпреступники могут использовать методы ИИ для улучшения любого из них. Естественно, есть место для оптимизации или улучшения всех классических шагов атаки: разведка, проникновение, боковое перемещение, командование и управление (C&C) и эксфильтрация. (Полное исследование возможных атак с использованием ИИ можно найти в исследовании Шведского агентства оборонных исследований (FOI)).

Социальная инженерия в масштабе

Мошенничество с социальной инженерией не ново, но оно остается успешным, потому что оно питается врожденным доверием людей, страхом и уважением к авторитету. Мошенники, с другой стороны, пытаются убедить своих жертв в том, что их схема законна и что целевые жертвы могут безопасно переводить деньги под любым предлогом, к которому призывает мошенничество. Тем не менее, эти мошенничества трудоемки, потому что преступнику приходится тратить значительное количество времени, пытаясь убедить жертву в том, что история правдива.

Однако, инновационный мошенник может внедрить системы ИИ для автоматизации и увеличения скорости обнаружения, с которой жертвы попадают в мошенничество или выходят из него, что позволит им сосредоточиться только на тех потенциальных жертвах, которых легко обмануть. Какое бы ложное притворство ни выбрал мошенник, чтобы убедить цель принять участие, алгоритм мл сможет предвидеть наиболее распространенные ответы цели на выбранное притворство. Затем алгоритм может действовать в соответствии с этими ожидаемыми ответами, чтобы продолжить весь акт и придать достоверность истории. Что еще более тревожно, этот относительно простой процесс требует только достаточного

количества письменного материала, чтобы определить наиболее вероятные ответы на каждое входящее сообщение.

К сожалению, преступники уже имеют достаточный опыт и образцы текстов для построения своих операций. Даже если бы мошенники автоматически классифицировали тексты только в рамках первых трех обменов электронной почтой – независимо от того, заплатит потенциальная жертва или нет, – мошенники смогли бы сэкономить значительные усилия и время, поскольку этих трех писем было бы достаточно для того, чтобы алгоритм мл продолжил работу.

Фактически, такое использование систем ИИ было бы эквивалентно автоматизации технической поддержки первого уровня в корпорации. Простое сосредоточение внимания на этих трех обменах электронной почтой даст наибольшую отдачу от инвестиций мошенников, поскольку они смогут автоматизировать остальные. Таким образом, автоматизировав один реальный значительный расход в таком преступном начинании — в данном случае людей — мошенник мог бы снизить большую часть затрат и сохранить свой доход, тем самым увеличив вероятность успеха своей мошеннической операции.

Другой вероятный сценарий будет связан с системами онлайн-верификации, которые используются в банковском секторе для проверки создателя нового счета. Эти системы проверки основаны на принципах «Знай своего клиента» (также известных как KYC). Как ни странно, были атаки, когда видеопотоки манипулировали в режиме реального времени, чтобы вставить поддельные идентификаторы, чтобы обмануть систему проверки. В сочетании с возможностью синтеза голосов и лиц преступники могут злоупотреблять этими системами в масштабах, например, для создания новых банковских счетов или авторизации платежей. Это важно отметить, поскольку аутентификация на основе видео в настоящее время является одним из рекомендуемых механизмов идентификации для банковской отрасли в нескольких странах.

При попытке защитить системы верификации от подобных атак существует два подхода: Первый – попытаться определить, было ли сообщение сфабриковано (в случае КУС это будет включать проверку видео). Второй подход заключается в том, чтобы выяснить, заслуживает ли источник доверия, чтобы узнать, является ли сообщение подлинным или нет.

В зависимости от характера сообщения обнаружение атаки может быть затруднено. Например, в случае вышеупомянутого мошенничества с электронной почтой достаточно простого спам-фильтра. С другой стороны, манипулируемую фотографию или видео может быть невозможно обнаружить как подделку.

Тем не менее, объединение обоих подходов против такого рода атак, вероятно, будет наиболее успешной стратегией. Было бы важно не только определить, было ли сообщение сгенерировано автоматически, но и определить, является ли источник сообщения ненадежным. Представьте, например, как один и тот же IP-адрес, отправляющий файлы аутентификации КУС с разными изображениями, скорее всего, вызовет красный флаг в принимающей организации.

Генерация контента

Генерация контента относится к способности алгоритма генерировать произвольный контент, который выглядел бы созданным человеком. Этот алгоритм также позволит установить ограничения при создании контента и создать систему, имитирующую или клонирующую определенные аспекты существующего фрагмента контента. Есть утверждения, что генераторы контента на основе ИИ становятся настолько мощными, что их публикация для общественности считается риском.

Такие утверждения относятся к Генеративному предварительно обученному трансформатору 3 (GPT-3), технологии синтеза текста, выпущенной OpenAI в июне 2020 года, которая использует глубокое обучение для создания человекоподобного текста и способна адаптировать и точно

настроить его поведение с добавлением нескольких примеров, относящихся к конкретной области. Обладая емкостью более 175 миллиардов параметров машинного обучения (в 10 раз больше, чем у его ближайшего конкурента, Microsoft Turing NLG), эта технология способна синтезировать не только английский текст, но и код на нескольких языках программирования и даже гитарные табулатуры, что позволяет использовать такие приложения, как:

Создание полного, звучащего по-человечески текста простого заголовка

Превращение текстового описания приложения в рабочий код

Изменение стиля написания текста при сохранении содержания

Прохождение теста Тьюринга для чат-бота с человеческим голосом

Таким образом, преступники могут использовать мл для создания и распространения нового контента, такого как (полу) автоматически создаваемые и высококачественные (копейные) фишинговые и спам-письма на менее популярных языках. По сути, это еще больше автоматизирует и расширит масштабы распространения вредоносных программ по всему миру.

Кроме того, такие методы создания контента могут значительно улучшить кампании по дезинформации, автоматически комбинируя законную и ложную информацию, а также изучая, какие виды контента работают лучше всего и какие являются наиболее широко распространенными.

Способность генерировать рабочий код из простого текстового описания снижает барьер знаний, необходимый для того, чтобы стать программистом, и может способствовать появлению нового поколения «детей-скриптов», то есть людей с низкими техническими знаниями, но злонамеренными намерениями, которые используют готовые инструменты для выполнения вредоносных действий.

Синтез текстового контента также может быть использован для создания семантически обоснованного контента для поддельных веб-сайтов и, что более важно, для воспроизведения определенного стиля текста. В частности, синтез текста с сохранением стиля – это метод, который использует систему ИИ для создания текста, имитирующего стиль письма человека. Примечательно, что

система ИИ не нуждается в большом объеме, а скорее в нескольких образцах стиля письма человека для ее обучения.

Действительно, возможности таких технологий, как GPT-3, заставляют нас поверить, что это действительно может быть, с точки зрения воздействия, следующим *deepfake*.

Анализ содержимого

Анализ содержимого – это деятельность, обычно связанная с необходимостью идентификации и извлечения структурированной информации из неструктурированных документов. Более широкое определение может включать распознавание объектов и извлечение из статических изображений или видео, которые де-факто подпадают под распознавание лиц, описанное в следующих разделах этого документа.

Более конкретно, существует приложение, связанное с анализом текста, называемое Распознаванием именованных сущностей (NER). Вкратце, NER пытается идентифицировать семантически значимые части и фрагменты индивидуальной информации, такие как имена, адреса и номера в произвольном тексте. Кроме того, он имеет возможность различать кредитные карты, номера телефонов и валюты, среди прочего. Было проведено значительное количество исследований, направленных на совершенствование технологии более точного определения семантических фрагментов текста. Чтобы проиллюстрировать, этот вид идентификации включает в себя способность различать, используя только контекст, если слово «буш» относится к растительности или к «Джорджу Бушу», бывшему президенту США, и, в частности, к какому Джорджу Бушу относится текст.

Аналогичным образом злоумышленники работали над разработкой вредоносных программ для очистки документов, которые, будучи установлены на целевой машине, будут проходить через все документы на этой машине и либо эксфильтровать весь документ (что рискованно и очень неэффективно с ростом числа документов), либо искать определенные биты информации.

Например, установленная вредоносная программа для очистки документов может искать все личные данные сотрудников на сервере отдела кадров.

Однако с развитием более эффективных методов NER в будущем, вполне вероятно, что более сложные вредоносные программы для очистки смогут лучше идентифицировать релевантный контент и даже выполнять целевой поиск.

Например, злоумышленники могут эволюционировать от очистки сервера компании для «всего, что напоминает номер телефона», до очистки «всех номеров экстренных контактных телефонов для всех менеджеров высшего звена.»

Помимо синтеза текста, за последние несколько лет было достигнуто несколько достижений в области синтеза изображений, вплоть до того, что система ИИ теперь способна синтезировать фотореалистичное изображение из текстового описания.⁵³ Излишне говорить, что способность генерировать произвольные изображения, которые неразличимы невооруженным глазом, является мощным инструментом для социальных инженеров. Это еще одна область, в которой использование моделей GAN уже может дать высококачественные результаты. GAN для синтеза изображений более подробно обсуждаются в тематическом исследовании, которое следует за этой статьей.

Улучшено старение социальных профилей для форумов и ботнетов

Еще одна интригующая возможность вредоносного использования систем ИИ связана с эмуляцией поведения пользователей. Одним из непосредственных вариантов было бы использовать его для уклонения от системы безопасности, основанной на обнаружении биометрических данных и специфического поведения человека. Например, Typing DNA⁵⁵ предлагает коммерческий продукт с возможностью блокировки компьютера всякий раз, когда он обнаруживает шаблон ввода, отличный от обычных шаблонов пользователя. Отчеты также предполагают, что системы обнаружения ИИ уже используются

финансовыми учреждениями для обнаружения ненормальных трат или схем денежных переводов. В то же время есть сообщения о том, что компании уже начинают видеть, как преступники пытаются обойти эти системы, пытаясь придерживаться «нормальных» параметров использования, таких как рабочее время и поддельные IP-адреса в пределах доверенных диапазонов при работе со взломанными банковскими счетами или украденными кредитными картами.

Еще одно примечательное приложение включает в себя использование этих методов для создания ложного поведения в украденных учетных записях, чтобы избежать их закрытия. Эта концепция используется в учетных записях Spotify для автоматизации мошеннической монетизации, избегая при этом обнаружения системой, как описано ранее в подразделе под названием «Олицетворение человека на платформах социальных сетей.»

То же самое может быть достигнуто для поддержания активности на украденных аккаунтах Twitter или Facebook, которые преступники могут монетизировать дольше. Качество этих услуг, которые продают лайки или подписчиков, можно измерить по тому, насколько «старыми» являются эти учетные записи «ботов». Естественно, любая техника, которая помогает преступникам дольше сохранять поддельные учетные записи законными, обязательно станет для них ценным инструментом.

Для того, чтобы улучшить коммуникацию C&C ботнета, в одном предложении предлагает метод на основе роя, основанный на принципах ИИ. В этом методе при присоединении к ботнету зараженный узел отправляет многоадресный сигнал. Более старые участники узнают о присутствии этого нового бота и обновляют свои таблицы, чтобы отразить это. Команды отправляются случайным узлам, которые после расшифровки могут быть перенаправлены на соседние узлы для повторения процесса. Это новый способ управления ботнетом. Однако на данный момент это всего лишь теоретическая возможность.

Методы, основанные на ИИ, которые пытаются имитировать реальное поведение пользователей, нелегко идентифицировать. Возможно, используя

методы ИИ, специально направленные на их обнаружение, можно было бы увеличить вероятность разоблачения этих человекоподобных ботов. С другой стороны, описанные бессерверные методы C&C кажутся гораздо более сложными для понимания. Важно подчеркнуть, что с точки зрения стороннего наблюдателя, сетевых следов было бы недостаточно, чтобы собрать воедино то, что происходит на самом деле. К сожалению, сетевые следы часто являются единственным, что доступно защитнику. Как и во всех случаях, чем больше данных у защитника, тем выше вероятность того, что он узнает, что происходит на самом деле. Как в головоломке, чем больше кусочков у вас в руках, тем яснее становится картина.

Робоколлинг v2.0

Так называемая операция робокаллинга может быть оптимизирована во многом аналогичным образом. Робокаллинг стал способом совершения фишинг-мошенничества с помощью обычного телефона. В этом виде мошенничества автоматический вызывающий абонент доставляет голосовое фишинговое сообщение и пытается соблазнить жертву посетить вредоносный веб-сайт. Добавив интеллектуальную автоматизацию в такую систему мошенничества, преступники могут отслеживать, является ли мошенничество успешным или нет, и какие аргументы и логика наиболее убедительны для потенциальных жертв. Настраивая эти параметры, они могут усовершенствовать мошенничество, чтобы со временем получить лучшие результаты. Имея достаточное количество точек данных, злоумышленники могут использовать их в качестве функций для обучения постепенно улучшающихся моделей мл для усиления своих атак.

Еще одна интригующая возможность улучшения робокаллинга может включать использование звуковых глубоких подделок, чтобы обмануть пользователя, думая, что он имеет дело с человеком, которого он знает. Например, когда-то популярная афера заключалась в получении какой-либо формы общения от знакомого, который находится в бедственном положении.

Якобы они застряли где-то в мире без денег и нуждаются в помощи, чтобы добраться до своего отеля. Этот вид мошеннического сигнала бедствия может быть сделан гораздо более убедительным, однако, если он использовал голосовое сообщение от знакомого, чтобы лучше передать их беспокойство и панику.

Обнаружение сфабрикованных звуковых файлов так же сложно, как и попытка проверить любой другой вид *deepfake*. Дополнительным слоем здесь является то, что мошенничество с робокалливированием совершается через обычный телефон. Это означает, что до тех пор, пока они звучат убедительно для пользователя, они могут быть незаметны, потому что нет автоматизированного уровня, который мог бы пометить это как мошенничество. Напротив, любая другая афера, в которой звуковой файл имеет шанс пройти фильтр обнаружения, может стать шансом для защитника узнать, что файл является подделкой, а не реальным звуковым файлом.

Криминальная бизнес-разведка

Тот же подход к использованию ИИ для повышения и оптимизации эффективности преступных операций может быть применен и к любому другому мошенничеству, например, к обычному фишингу электронной почты. мл, в частности, уже применяется для повышения показателей успеха любого корпоративного предприятия – от продаж до маркетинга. Предположительно, поэтому мл может быть столь же эффективным для злонамеренного использования.

Представьте, например, что обычная фишинговая операция, направленная на банки, добавляет небольшой тег в электронные письма или встроенные фишинговые ссылки. Всякий раз, когда потенциальная жертва получает электронное письмо, мошенник будет знать, видел ли его получатель (в противном случае следует предположить, что оно попало в папку со спамом), или если ссылка была нажата. Кроме того, мошенник узнает, была ли введена

какая-либо личная информация на фишинг-странице, а также качество этой информации.

Сопоставляя все эти данные, мошенник может получить хорошее представление о том, какие электронные письма более успешны для каждого банка. Более того, они узнают, какие базы данных электронной почты с большей вероятностью будут иметь хорошие показатели успеха по сравнению с теми базами данных, которые неоднократно использовались повторно и, следовательно, больше не будут давать хороших результатов.

Автоматизируя все это, вся операция по мошенничеству может быть самодостаточной, и мошенник может удалить базы данных электронной почты, которые вряд ли будут доставлять фишинговые письма, создавать новые электронные письма, которые с большей вероятностью будут успешными, и отправлять их на те адреса, которые принадлежат более восприимчивым людям. Эти механизмы также могут создавать новые базы данных электронной почты, в которых перечислены потенциальные жертвы, которые с большей вероятностью попадут в большее число мошенников или идентифицируют тех людей, которые открывают мошеннические электронные письма и выбирают ссылки в них.

Примечательно также, что современные маркетинговые и рекламные компании уже используют подобные системы. Поэтому не было бы скачком воображения предполагать, что организации, основанные на вредоносных программах, начали использовать те же методы, основанные на ИИ, для повышения своей собственной эффективности.

На том же фронте расширения возможностей криминальной бизнес-аналитики ИИ также может быть использован для оптимизации потока данных в системах распределения трафика (TDS).

TDS служат посредниками, которые покупают и продают трафик между веб-сайтами. Основные функции TDS включают контроль и фильтрацию веб-трафика (которые включают выбор ссылки) и сбор соответствующей статистики. Они также могут фильтровать трафик на основе предпочтений

киберпреступника, таких как веб-браузер пользователя и местоположение (например, через IP-адрес).

Например, киберпреступник может установить параметры в TDS, чтобы перенаправить пользователей в США на банковские трояны, но доставить вымогателей в другие страны и избежать развертывания их вредоносных программ в Содружестве Независимых Государств (СНГ). Системы распределения трафика являются основным средством распространения вредоносных программ с помощью наборов эксплойтов и загрузок с диска и могут выступать в качестве службы для массового маркетинга вредоносных программ. Поставщики TDS продают трафик с момента, когда жертва нажимает на ссылку.

В то время как TDS в настоящее время полагаются на жестко закодированные правила, установленные злоумышленниками, легко понять, как система ИИ может быть использована для оптимизации самого распределения трафика, чтобы максимизировать доход злоумышленника.

В целом, вполне вероятно, что в ближайшем будущем эволюция этой концепции будет включать все больше и больше услуг с поддержкой ИИ. Это связано с тем, что деятельность, связанная с ИИ, как правило, требует значительных первоначальных инвестиций с точки зрения человеческих знаний и учебных наборов данных, которые могут быть легко восстановлены путем предоставления услуг ИИ нескольким заинтересованным сторонам.

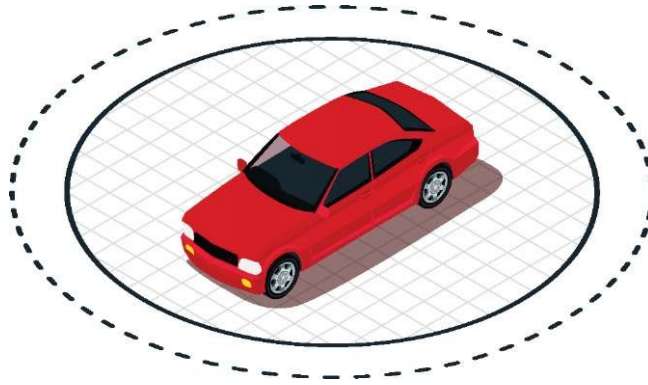
Злоупотребление системами распознавания изображений автономных автомобилей

Транспортные средства используют различные датчики. К ним относятся камеры для распознавания знаков и других элементов окружающей среды, таких как пешеходы и линии разделения полос движения, с помощью мл-ориентированного изображения. Эти модели мл также определяют соответствующее поведение для транспортного средства. В автономном автомобиле эпохи 4G это не было бы серьезной проблемой, поскольку из-за

низкого качества данных водитель должен постоянно контролировать движение на дорогах общего пользования, а полная автономия транспортного средства невозможна. Однако в полностью автономном транспортном средстве, подключенном к 5G, визуальные датчики могут иметь возможность как воспринимать линии движения, так и выполнять навигацию на основе того, что они видят.

Кроме того, автономные автомобили «обучаются» подчиняться закону (например, не пересекать сплошные линии движения), но это обучение также может быть использовано для атаки моделей ИИ автономного автомобиля. «Автономная ловушка 001», произведение перформанса Джеймса У., основано именно на такой концепции. Ловушка, проиллюстрированная на следующем рисунке, включает в себя рисование двух концентрических кругов в соли, причем внутренний круг рисуется сплошными линиями, а внешний круг – ломаными. Поскольку автономное транспортное средство обучено тому, что оно может пересекать разбитые линии движения, оно может легко войти в круг. Однако одно и то же транспортное средство не сможет покинуть круг, не пересекая сплошные линии полосы движения, что означало бы нарушение закона.

Размещение ловушек (таких как этот круг) в удобных для преступлений местах, которые покрыты глушителем (или в месте без покрытия сотовой связи), таким образом, соберет ослепленные глушителем автономные автомобили, такие как липучка – возможность для хаоса, которой преступники могли бы воспользоваться.



Автономные автомобили могут быть захвачены кольцом знаков «объезд слева» или другими подобными методами, как показано в работе Джеймса Бридла, в которой использовались руны движения, нарисованные солью.

Аналогичный состязательный метод, который нарушает правила моделей датчиков изображений мл, может быть достигнут путем создания других знаков или меток. В одном примере требуется только небольшой кусок черной ленты на знаке ограничения скорости. В полевых испытаниях лента использовалась для изменения цифры «3» в знаке ограничения скорости 35 миль в час. В результате модель распознавания изображений ошибочно распознала «35 миль в час» как «85 миль в час». Следовательно, он совершал эту ошибку в 58% случаев, в результате чего автомобиль разгонялся до запрещенной скорости.

Подобные атаки также могут быть использованы против систем ИИ без изменения самих дорожных знаков, а просто их размещения.

Чтобы проиллюстрировать это, злоумышленники могут сделать круг шириной в километр, построенный с помощью дорожных знаков, таких как «объезд налево», который в конечном итоге поймает все автономные транспортные средства, следующие за знаком объезда. Единственный способ спастись — это чтобы их водители — люди подсказали «незаконный» способ покинуть круг, в конечном итоге проигнорировав, то есть «не подчинившись» знакам «объезд слева». Здесь также важно отметить, что автоматизированные дорожные знаки позволят осуществить эту атаку на автономные транспортные средства с помощью хакера, вносящего изменения в дисплеи знаков.

Что еще более важно, эти возможные методы управления системами ИИ, которые управляют автономным движением транспортных средств, могут быть

использованы для управления даже неавтономным движением транспортных средств, используя совместимые автономные транспортные средства в качестве препятствий. Такие методы, применяемые к неавтономным транспортным средствам, могут привести к значительным задержкам во времени реагирования полиции и аварийных служб. В сочетании с массовым вредоносным обменом (использование многих украденных кредитных карт для затопления района транспортными средствами для обмена) это может привести к интенсивному движению в центре города. В результате трафик может быть увеличен, чтобы усилить эффект террористического события, дать больше времени для продолжения ограбления банка или продлить продолжительность пожара, задержав прибытие пожарных.

Изопренность и инвестиции в такого рода атаки минимальны, и все же способность предотвратить их сложна. Эти атаки, на самом деле, требуют лишь небольшого количества автономных транспортных средств, чтобы стать полностью эффективными: будет достаточно менее одного из 100 транспортных средств, и еще меньше в напряженное время дня.

Атаки с использованием ИИ, направленные на экосистемы подключенных транспортных средств, включая телематику, информационно-развлекательные системы (например, подкасты) и системы связи «Транспортное средство-все» (V2X), могут привести к обездвиживанию транспортных средств, дорожно-транспортным происшествиям, финансовым потерям и раскрытию конфиденциальных или персональных данных. Что еще хуже, это может поставить под угрозу безопасность участников дорожного движения. Примечательно, что по мере повышения уровня автоматизации транспортных средств и постепенной интеграции различных технологий в транспортные средства потенциальная поверхность атаки также становится шире.

Наконец, поскольку та же или аналогичная технология используется в автономных грузовиках или автономных грузовых судах, потенциальные цели для злонамеренного использования также увеличиваются.

Дроны, связанное небо и Интернет летающих вещей

Высококачественные данные управления воздушным пространством, включая телеметрию с устройств с поддержкой SIM-карт, используются для управления беспилотными летательными аппаратами Интернета летающих вещей (IoFT). Приложение поддержки полетов «управление полетом», в частности, управляется моделями мл, которые реагируют на телеметрию, полученную от бортовых датчиков IoFT, включая камеры распознавания изображений.

Эти камеры и другая телеметрия передаются обратно в их собственную версию управления трафиком: в данном случае это своего рода управление воздушным движением, которое требует высококачественных данных на уровне 5G. Для того чтобы беспилотные дроны могли реализовать свой потенциал, они должны иметь обратную связь в режиме реального времени, которая требует оптимизированного для мл 5G.

Для работы за пределами прямой видимости (BVLOS) этим беспилотным летательным аппаратам также требуется несколько механизмов, связанных с ИИ: от компьютерного зрения для выявления препятствий и ссылок на пути до тщательной оптимизации их безопасного полета (за счет предотвращения столкновений), безопасной доставки (включая доставку людей) и безопасной навигации (путем следования зонам «без полетов»). Тем не менее, эти модели поведения могут быть использованы путем применения пробелов в основных моделях поведения. Часто эти поведенческие пробелы могут быть вызваны передачей ложной информации по радио в сочетании с отображением ложной информации, такой как большая буква «Н» в круге на земле, которая может обмануть «потерянный» вертолет при посадке. Этот спекулятивный подход, скорее всего, захватит оптимизированные для мл беспилотные летательные аппараты таким же образом, как Иран захватил беспилотный летательный аппарат ЦРУ США Sentinel.

Захват автономных беспилотных летательных аппаратов представляет особый интерес, поскольку они, как правило, содержат множество интересных полезных нагрузок, таких как интеллектуальная собственность и живые ключи шифрования. Они могут быть использованы против ранее захваченного исходящего радиопотока, перехваченного с беспилотника, а затем могут быть использованы для передачи плохих разведданных и ложной телеметрии в вышестоящие системы. Эти вышестоящие системы, в свою очередь, могут быть отравлены ложной телеметрией, включая изображения, побуждающие к тактическим или исполнительным действиям по неправильным целям. Некоторые методы включают в себя взлом SIM-карт или обмен SIM-картами (использование социальной инженерии или других методов для получения контроля над SIM-картой и всеми ее данными), управление парком беспилотных летательных аппаратов, захват учетных записей (управление парками SIM-карт и их данными) и другие.

В случае дронов доставки Amazon они могут быть использованы для создания более крупных финансовых стимулов, поскольку цель (в данном случае посылка) будет легче монетизировать. В результате это создало бы эквивалент «пиратов крыльца» или, в данном контексте, «пиратов дронов».

Кроме того, беспилотник может быть вызван для доставки наркотиков или антибиотиков в местах, контролируемых преступниками. Опять же, с помощью взлома SIM-карт или обмена SIM-картами беспилотник может быть захвачен, а полезная нагрузка заменена на другую, чтобы облегчить незаметную контрабанду.

Беспилотный летательный аппарат с одноплатным компьютером UDOO (SBC) также может использоваться для сбора паролей Wi-Fi. Когда беспилотник летит, он сканирует сигналы Wi-Fi, и как только он находит маршрутизатор, он пытается взломать пароль маршрутизатора. Когда это оказывается невозможным, SBC дрона прибегает к сохранению пакета, содержащего данные, для последующего взлома с помощью Aircrack-ng.

Что касается смягчения последствий, то в настоящее время разрабатываются новые правила ЕС для беспилотных летательных аппаратов. Одно из них включает ограничение размера отдельных ударов беспилотных летательных аппаратов с высоким ущербом с использованием потребительских беспилотных летательных аппаратов. По сути, новый лимит также снизит пропускную способность беспилотника для взрывчатых веществ, хотя существует некоторая задержка в осуществлении этих правил из-за пандемии Covid-19.

Выход из системы распознавания изображений

Системы распознавания образов, такие как социальный скоринг и технологии социального кредитования, подпадают под категорию «репутационные системы». Они очень похожи на системы бизнес-аналитики и маркетинга, называемые анализом и склонностями социальных сетей (SNAP), которые анализируют поведение личных социальных сетей, чтобы определить вероятность совершения определенного действия, такого как покупка продукта. Эти системы SNAP также определяют, кто является «влиятельным лицом» или лидером группы. Аналогичные национальные разведывательные системы используются для выявления преступной деятельности, такой как привычки террористов-изготовителей бомб или членов организованных групп, таких как преступники. Без инструментов обнаружения социальных сетей, таких как SNAP, было бы трудно в полной мере использовать инструменты распознавания изображений для выявления того, кто является лидером преступной группы, или даже если такая «группа» существует.

Тем не менее, использование таких систем распознавания изображений правительством страны может представлять серьезную угрозу неприкосновенности частной жизни и правам человека. Например, в середине 2010-х годов российская компания NtechLab, специализирующаяся на распознавании лиц, выпустила приложение FindFace. С помощью этого приложения пользователь может сфотографировать любого человека, а затем

сопоставить его лицо со своими учетными записями в социальных сетях. В наши дни FindFace больше не является потребительским приложением, а инструментом для амбициозного многомиллионного проекта слежки в Москве. Используя распознавание в реальном времени и в режиме реального времени, эта технология может выделять отдельные лица среди толпы и мгновенно сопоставлять их с записями преступников в полицейской базе данных.

Что касается беспилотных летательных аппаратов, преступники уже успешно использовали их – без распознавания лиц – доставить контрабанду в тюрьмы и стрелять из перцового баллончика по протестующим лейбористам. Интересно, что использование этих «парамилитаризованных» функций в последнем примере можно рассматривать как эволюцию деятельности полиции по борьбе с протестами в черных масках, подобной той, которую осуществлял BlackLock во время событий G20. Поскольку организаторы протестов в черных масках прячутся в толпе и пытаются спровоцировать более агрессивную реакцию полиции на камеру, им нужен способ избежать попадания своих собственных дронов с перцовым баллончиком, когда они пытаются эскалировать конфликт. Необходимость в этой дополнительной мере защиты от собственных беспилотников обусловлена тем, что в соответствии с иерархией террористических ячеек многие члены этой группы не знают друг друга. В результате они не могут обмениваться фотографиями с лицами друг друга для загрузки в список безопасности дронов.

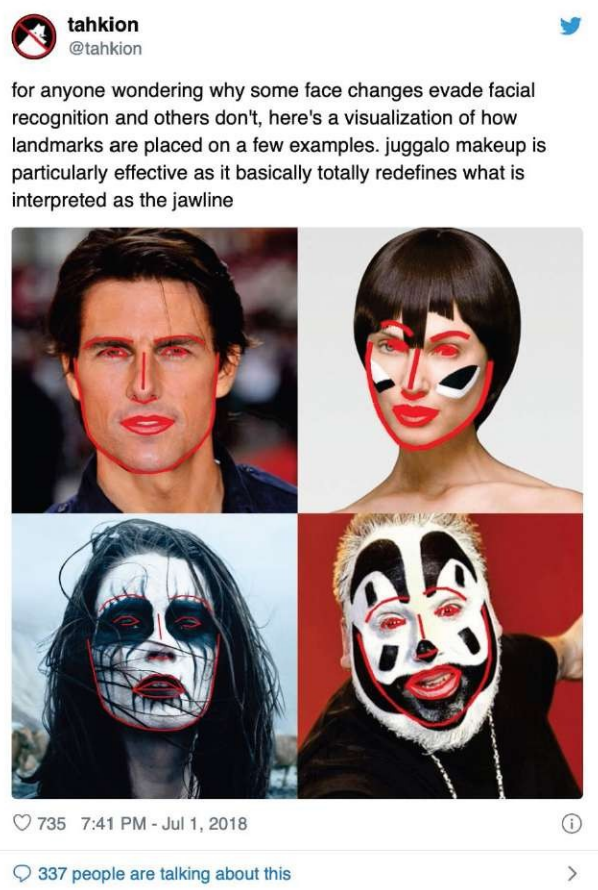
Тем не менее, конкретные рисунки на масках для лица (например, черно-красное счастливое лицо) могут быть быстрым и удобным способом попасть в список безопасности. Поскольку преступная группа может присвоить уникальный логотип маски каждому из своих членов (которые могут быть частью списка безопасности), было бы трудно обнаружить и идентифицировать отдельного члена, даже с использованием распознавания изображений для логотипов масок. Тем не менее, если бы такой список лиц-участников, принадлежащих к преступной группе, существовал, он мог бы служить источником доказательств, если бы он был обнаружен полицией. По этой

причине участники организованной преступности должны были бы внести в список конкретных лиц-агностиков, которые затем могут быть утилизированы или стерты после события.

Здесь также стоит отметить, что методы уклонения от распознавания лиц используются не только членами «черных масок», но и другими протестующими. Племенные знания об этих методах были распространены в Интернете еще в 2010 году.

Эти одноразовые косметические «виртуальные лица» могут защитить людей, заботящихся о конфиденциальности. Тем не менее, они также могут быть использованы вредоносными группами, чтобы избежать возможных арестов полицией или микронаправленных ударов беспилотников, когда любой из них основан на распознавании лиц. Аналогичным образом, эти виртуальные лица также используются отдельными лицами в странах, которые запретили ношение масок на публике. Обличий для уклонения от распознавания лиц, вероятно, будут еще некоторое время, поскольку материалы для их создания дешевы, легко скрываются, легко дезавуируются и легко доступны. В конце концов, маловероятно, что обычные потребительские товары для создания таких личин (например, губная помада или краска для лица) будут объявлены вне закона.

Другие маскировки, такие как макияж клоуна Джаггало (используемый поклонниками хип-хоп дуэта The Insane Clown Posse), продвигаются как способ обойти распознавание лиц. Благодаря макияжу Джаггало у человека могут быть неправильно классифицированы части лица, такие как глаза или линия подбородка. Фактически, система распознавания лиц мл-классификатора должна быть «настроена», в частности, для обнаружения этих лиц в дополнение к обычным человеческим лицам. Для этого федеративное ведение журнала в нескольких дисгармоничных системах может объединить анонимизированного человека с чем-то идентифицируемым, таким как метаданные сотовой связи или номерной знак транспортного средства.



Коллекция изображений, проанализированных с помощью распознавания лиц: верхнее левое лицо правильно классифицировано; верхнее правое лицо имеет неправильно классифицированные глаза и правую бровь; нижнее левое лицо имеет неправильно классифицированные глаза и брови; нижнее правое лицо имеет неправильно классифицированные линию подбородка, нос, рот и брови.⁸³ Фото: Ian O'Neill/Twitter

Хотя вышеупомянутые средства считаются эффективными для работы с распознаванием лиц, некоторые системы (такие как Apple Face ID) используют восприятие глубины для идентификации ямочек и других черт лица. Однако эта эволюция в системе менее эффективна при использовании на расстоянии и только с одной камерой.

Наконец, есть еще один неотложный вопрос, касающийся дронов для распознавания лиц и опасности, которую они могут представлять. Действительно, следует отметить, что сегодня разрабатываются крошечные беспилотные летательные аппараты для распознавания лиц, несущие грамм взрывчатого вещества, специально для микроцелевых взрывов с одним человеком. Эти дроны выглядят как насекомые или маленькие птицы. Учитывая природу этой технологии, подобные устройства управляются через

сотовый Интернет. Также можно с уверенностью предположить, что в ближайшее время они будут переданы в руки преступников – либо путем покупки, либо разработки.

Таким образом, решение проблемы уклонения от распознавания лиц может быть затруднено для организаций, использующих системы наблюдения на основе мл, – проблема, которая также рассматривается в рамках вышеупомянутой инициативы CUIING. В частности, для этих дискуссий крайне важно определить этический процесс идентификации людей, которые хотят избежать распознавания лиц.

Алгоритмы безопасности

В случае алгоритмов безопасности при обнаружении мошеннической деятельности система борьбы с мошенничеством обычно блокирует ее и выдает предупреждение. В модели мл по борьбе с мошенничеством действия, которые провоцируют ответные меры по борьбе с мошенничеством, могут быть очень четко и эффективно отображены. Фактически, пороговые значения предупреждения о мошенничестве могут быть встроены в модель угроз злоумышленника (основанную на журналах криминальных успехов) и представлены в их инфраструктуре атаки, архитектуре и расходах.

Теоретически эти алгоритмы безопасности потенциально могли бы использовать аналитический инструмент маркетинговых коммуникаций на основе бизнес-аналитики, такой как Cloudera, для отправки конкретных преступных «маркетинговых кампаний» выбранной целевой жертве или «клиенту.» Когда мл используется для точного отображения систем борьбы с мошенничеством конкретной группы жертв, «маркетинговая кампания», упомянутая ранее, может быть очень точно настроена, чтобы всегда оставаться чуть ниже пороговых значений по борьбе с мошенничеством. По сути, это максимизирует прибыльность инвестиций в криминальное предприятие.

Таким образом, используя мл для определения слепых зон в порогах обнаружения (на основе того, что было заблокировано, а что нет),

злоумышленник может оставаться в этой слепой зоне и оставаться незамеченным и постоянным до тех пор, пока правила обнаружения не будут обновлены. Обнаружить это может быть так же просто, как использовать закон Бенфорда, который гласит, что в естественных наборах чисел наиболее частая ведущая цифра, скорее всего, будет наименьшей. Поэтому, если статистика натуральных чисел, используемая в ИИ, ведет себя не так, как ожидалось статистически, могут быть запущены правила безопасности, такие как оповещение. Поиск «математически непротиворечивых аномалий», которые действуют чуть ниже общих порогов оповещения, на самом деле является статистическим правилом, используемым при обнаружении мошенничества. В то же время более продвинутые актуарные методы, заимствованные из финансовой отрасли, позволяют выявлять более сложные виды мошенничества, такие как отмывание денег.

Манипулирование фондовым рынком с помощью ИИ

Алгоритмические системы торговли акциями с поддержкой ИИ в последние годы заменили многих трейдеров-людей, особенно в области высокочастотной торговли (HFT). Однако возросшая зависимость от таких систем также открыла возможности для манипулирования фондовым рынком с помощью ИИ. Действительно, уже есть случаи, когда частным инвесторам удавалось взломать алгоритмы крупного электронного рынка и манипулировать ценами акций в свою пользу. Эксперименты с состязательными ботами также демонстрируют, что с помощью автоматизации можно манипулировать фондовым рынком.

Алгоритмы HFT преувеличивают и ускоряют рыночную активность. Аналогичным образом, «внезапные обвалы» фондового рынка и потребление стоимости активов за счет серии небольших доходов от более крупных торговых сборов могут привести к внезапным непредсказуемым изменениям на рынке с последующим изменением доверия инвесторов. Это может привести к

дальнейшему системному риску для инвесторов, если они будут эксплуатироваться преступниками.

Вариации этих подходов к «обнаружению преступных МО» безопасности могут быть использованы для обнаружения взлома SIM-карт (путем обнаружения статистически аномального поведения), массового прослушивания (путем оповещения о неожиданных телекоммуникационных метаданных, таких как неустойчивая задержка) и других оптимизированных телекоммуникационных злоупотреблений. Традиционные виды мошенничества, такие как мошенничество с кредитными картами, также можно найти в нетрадиционных методах, таких как автоматическая оплата SIM-карты и другие методы отмывания денег.

Компрометация бизнес-процессов и внедрение спутниковой телеметрии

Телефоны, ноутбуки, устройства Интернета вещей (IoT) и большинство других типов технологий, подключенных к Интернету, являются мощными устройствами наблюдения и могут рассматриваться как датчики. Информация об удаленных датчиках (называемая телеметрией датчиков) используется корпорациями, которые управляют устройствами. Например, устройством можно управлять, следуя такой инструкции, как «Когда телеметрия датчика X сообщает, что произошло событие Y, выполните событие Z.» Тем не менее, в то время как большинство организаций имеют сетевую и программную безопасность, очень немногие имеют безопасность данных. Этот пробел означает, что атаки, основанные на манипулировании данными (например, мошенничество), относительно незаметны, когда они также используют незащищенные конечные точки, такие как датчики.

Примечательно, что эта телеметрия очень важна. Очень большие и сложные сенсорные сети, такие как телефонные сети операторов связи или национальные корпоративные IT-сети, могут принимать выгодные решения на основе детальной информации, полученной с помощью анализа телеметрии на основе ИИ. Когда несколько датчиков скомпрометированы (с помощью

ботнета, взлома SIM-карт или других средств), модель мл получит ложную телеметрию датчиков, и поскольку она основывает свои решения на этой телеметрии, она невольно открывает дверь для крупномасштабного постоянного мошенничества. К сожалению, модель мл должна быть обманута лишь небольшое количество раз с помощью ложной телеметрии, прежде чем она будет переобучена и обновлена, таким образом, включив эту ложную телеметрию в свои модели. В этот момент злоумышленник может отступить и позволить алгоритму ИИ, так сказать, сделать свою работу за них.

ИИ обычно выполняется в автономном режиме. Когда ИИ выполняется в производственной сети реального времени, например в 5G, активы, ранее контролируемые моделями ИИ, затем будут контролироваться злоумышленником. Кроме того, поскольку эти данные не фильтруются и не защищаются на уровне данных, они могут быть изменены, чтобы спровоцировать катастрофу, вызвав неожиданные ответы разработчиков систем, в том числе те, которые могут быть нарушены.

Пример компромисса бизнес-процесса: Пробел в дизайне навигации транспортного средства

В феврале 2020 года художник Саймон Векерт провел инсталляцию и перформанс, заполнив небольшую красную тележку законными, активными, полностью включенными телефонами.⁹⁰ Затем он протащил свою тележку, полную 99 активных телефонов, по нескольким немецким улицам. Google Maps рассматривал это как 99 автомобилей, движущихся медленно, так как в Google Maps медленное движение стольких телефонов интерпретировалось как состояние с высоким трафиком. Хотя описание работы Векерта по установке и производительности ссылается на статью Морица Алерта «Сила виртуальных карт», в которой анализируются отношения власти в картографии и виртуальные изменения в том, как люди взаимодействуют с картами с тех пор, как появились карты Google, результат его работы имеет некоторое отношение к пробелу в дизайне навигации транспортных средств. Действительно, в результате законной деятельности Векерта (воспринимаемой Google как

трафик) Google Maps затем перенаправила другие реальные автомобили на другие, гораздо менее эффективные и более отдаленные пути.

Хотя деятельность г-на Векерта была законным и очень оригинальным примером искусства исполнения и инсталляции, поэтому легко понять, как этим можно злоупотреблять. Использование 99 мобильных телефонов для отправки 99 автомобилей (или даже больше) в определенное место в определенное время достигло бы того же результата, с дополнительным воздействием физических автомобилей, фактически присутствующих и, таким образом, образующих реальное препятствие. Можно предположить, например, что достаточная плотность движения создаст проблему для реагирования пожарных, скорой помощи и полиции.

Инсайдерские атаки: банковские и торговые площадки

Некоторые задачи банка в значительной степени автоматизированы с целью улучшения и ускорения их выполнения. Ценностное предложение одного банка по сравнению с другим может определяться бесконечно малым улучшением скорости торговли (которое включает в себя использование множества крошечных возможностей), или самым незначительным улучшением точности торговли акциями (которое включает в себя множество крошечных сокращений убытков). Все крупнейшие банки имеют лучшие развертывания торговых залов, что приводит к почти идентичным системам между банками и почти идентичным развертываниям финансового ИИ. Эти развертывания ИИ в финансовых технологиях (FinTech), в свою очередь, имеют почти идентичные математические количественные или «квантовые» модели, что приводит к жестокой конкуренции за самый талантливый персонал и самое крошечное конкурентное преимущество, которое они могут представлять.



Художник Симон Векерт проводит свой перформанс и инсталляцию с тележкой, полной 99 активных телефонов, на улицах Германии.

В этом контексте ИИ важен для банков, поскольку он может повысить их прибыльность. Между тем, человеческие кванты, которые используют результаты моделей ИИ, также важны, поскольку они являются движущими силами глобального фондового рынка.

Однако существует более глубокий уровень, на котором специалисты по обработке данных ИИ отвечают за создание, настройку и реализацию самих моделей. По соображениям конкуренции личности этих лиц держатся в секрете в соответствии с контрактом, который также содержит карательные положения, запрещающие им работать в конкурирующем банке в течение длительного периода после того, как они покинут свой текущий банк.

Тем не менее, следует отметить, что эти специалисты по данным торговых алгоритмов находятся в сильном положении, поскольку они контролируют способность банка совершать как сделки, так и прибыль. Часто их модели мл и код настолько неясны, что выполнение этих моделей трудно перепроектировать даже для их создателя. Эти ученые обязаны (как их основная функция работы) манипулировать присутствием своего банка на фондовых рынках. Они являются бесспорными тайными хозяевами единственной машины, которая в наибольшей степени отвечает за способность банка зарабатывать деньги. По сути, никто не знает, как они делают то, что они делают, кроме них самих.

Предполагается, что эта возможность была бы слишком заманчивой, чтобы ее игнорировать. Для некоторых способность разрабатывать модели ИИ, которые являются «самореализующимися финансовыми пророчествами», позволяет почти незаметно осуществлять инсайдерскую торговлю, поскольку это поведение не вызывает тревоги. Банковские ученые такого рода могут диктовать огромную покупательную способность банка, покупать то, что, как они знают, банк купит, а затем продавать после того, как банк его купит. Поскольку они продают после того, как банк купил, акции ученого растут в цене, и они могут обналичить разницу. Эта инсайдерская торговля может быть незаметной, если она осуществляется в рамках так называемых пробелов аудиторского контроля в рамках Борьбы с отмыванием денег/финансированием терроризма (ПОД/ФТ) или контроля за инсайдерской торговлей (которые ищут только конкретные показатели). Действительно, ни один из этих средств контроля не может обнаружить инсайдерскую торговлю, основанную на ML, если будут предприняты определенные шаги. Если правила ИИ рассматривают поведение, подлежащее аудиту, как запрещенное, это будет означать, что будут выполняться только инсайдерские сделки, для которых нет других функций аудита. Кроме того, следует отметить, что этой практике преступники уже следуют сегодня, хотя и вручную, путем обратного аудита глобальных стандартов.

Эта практика также может быть реверсирована для короткой торговли акциями. Спекулятивно, алгоритмический тип отмывания денег ИИ отслеживает эти сделки и разделяет их результаты, чтобы они оставались ниже пороговых значений финансового обнаружения.

Заражение данных резидентным вредоносным ПО

Современные вредоносные программы могут изменять общее поведение компьютерной системы, вводя внешний код в библиотеки резидентного программного обеспечения, например, для сохранения перезагрузки или запуска поведения при определенных условиях. Такой детерминированный

механизм в настоящее время относительно легко обнаружить, поскольку он включает в себя либо инъекцию кода с определенной и обнаруживаемой сигнатурой, либо даже в случаях сильно зашифрованного кода поведенческую сигнатуру, которая может быть обнаружена внешним агентом.

Однако ML и, в частности, системы глубокого обучения не следуют философии классической компьютерной программы. Как мы упоминали в предыдущем подразделе под названием «Вредоносное ПО ИИ», решения не кодируются в детерминированных и явных строках кода, а скорее кодируются в виде обученных моделей, которые представлены матрицами чисел.

Этот подход поднимает известную проблему объяснимости моделей машинного обучения: всякий раз, когда система мл принимает решение, для многих других алгоритмов мл это решение принимается черным ящиком предварительно обученных матриц без четкой, объяснимой причины для принятия конкретного решения. Например, спам-фильтр, основанный на глубоком обучении для идентификации спам-контента, будет очень трудно отладить в случае ложного срабатывания, поскольку в нем не будет какой-либо конкретной информации, объясняющей, почему он классифицировал данное сообщение как спам.

Преступники, со своей стороны, могут рассматривать эту конкретную проблему как источник будущих возможностей. В то время как изменение библиотеки может быть обнаружено либо по коду, либо по поведенческим сигнатурам – например, путем поиска аномалий в коде или его выполнении, — аномалии в модели мл будет труднее идентифицировать из-за отсутствия объяснимости модели.

Таким образом, будущие вредоносные программы могут воздействовать на хост-машины, ориентируясь, эксфильтрируя или изменяя модель мл напрямую, а не библиотеки или программные API. В конце концов, разница в модели мл, вызванная вредоносным ПО, вряд ли будет отличима от разницы, вызванной обновлением модели.

Таким образом, первостепенное значение будет иметь обеспечение создания механизмов проверки целостности моделей мл, используемых в производстве, в виде алгоритмов контрольной суммы или последовательностей обучения целостности с известными результатами для модели.

Вымогатели с поддержкой ИИ

Кибератаки WannaCry и NotPetya 2017 года распространились по всему миру с беспрецедентной скоростью из-за самовоспроизводящихся функций.^{96, 97} В свете распространения атак вымогателей и постоянных усилий киберпреступников по оптимизации их эффективности атаки вымогателей, поддерживаемые ИИ, могут появиться с функциями самораспространения. В будущем глубокими нейронными сетями можно будет злоупотреблять для улучшения выбора цели на основе predetermined атрибутов или использовать для отключения мер безопасности целевой системы, которые облегчат боковое перемещение. Кроме того, ИИ может усугубить атаки вымогателей с помощью интеллектуального таргетинга и интеллектуального уклонения. Интеллектуальный таргетинг позволит находить новые уязвимости, используя комбинацию методов атаки и используя наиболее эффективные с точки зрения проникновения.

Последствия такой атаки вымогателей с поддержкой ИИ на международные частные компании могут быть разрушительными, поскольку их глобальные филиалы могут быть быстро заражены. Недавний показывает, что киберпреступники постепенно используют программы-вымогатели для нацеливания на целые муниципалитеты или общегородские службы, что может сделать потенциальные атаки программ-вымогателей, поддерживаемые ИИ, еще более разрушительными. Такая же возможность применима к критической инфраструктуре и основным услугам, которые также очень уязвимы для таких атак и могут привести к серьезным последствиям в реальной жизни с большим географическим охватом.

ИИ предоставляет злоумышленникам широкие возможности для повышения их операционной безопасности и избежания обнаружения. Семейства вымогателей, такие как Cerber, уже использовали многоуровневый подход к обнаружению вредоносных программ, включая отдельный загрузчик, предназначенный для уклонения от решений мл. Злоумышленники также могут использовать ИИ для защиты своей преступной инфраструктуры путем обнаружения злоумышленников или подозрительных сред, или они могут создать механизм саморазрушения в вредоносном ПО, который запускается при обнаружении подозрительного поведения. Такие формы стирания доказательств значительно затрудняют процессы сбора доказательств, расследования и технической атрибуции.

Банки используют голосовую аутентификацию для различных целей, особенно для обеспечения своего «самого мягкого периметра безопасности», человеческого элемента центров обработки вызовов. В колл-центрах по своей природе работают сотрудники, которые подвергаются самым изощренным практикам мошенничества в мире, которые, в свою очередь, часто обладают очень продвинутыми навыками озвучивания.

Целью этих злоумышленников является захват учетной записи. Они стремятся получить доступ к учетным данным учетной записи или запросить сброс пароля. В противном случае они стремятся получить доступ к учетной записи целевого клиента. Эта учетная запись используется для истощения ресурсов учетной записи жертвы или даже для получения доверия к еще более крупным атакам. Эта вторичная цель может также включать санкционирование крупных платежей от компании жертвы.

Внедрение голосовой биометрии снизило частоту мошенничества в отношении клиентов на целых 90% в случаях захвата учетных записей. Точность этих моделей обусловлена совпадением двух типов систем ИИ: голосовой аналитики и аналитики мошенничества. Когда обе системы функционируют должным образом, эти системы ИИ объединяются, чтобы уменьшить атаки колл-центров на банки, а также онлайн-атаки из различных

источников. Они включают в себя выполнение атак грубой силы на учетные записи клиентов (поскольку теперь им требуется сопоставление голоса), компрометацию базы данных PIN-кодов (в которой PIN-код является голосом), совместное использование учетных данных (люди не могут делиться своим голосом), фишинг или вишинг (человек не может легко отказаться от своего голоса) и функцию сброса PIN-кода (сброшенный PIN-код по-прежнему является голосом человека и его все равно будет трудно скопировать).

Голосовая аутентификация на базе мл по иронии судьбы обнаруживает в качестве своей слабости присущее технологам доверие к новым технологиям. Когда безопасность, представленная голосовой биометрией, подвергается риску с помощью инструментов клонирования голоса, таких как LyreBird или других, менее законных, активы, защищенные голосовой биометрией, подвергаются риску. В настоящее время эта ситуация с голосовой аутентификацией на основе мл представляет собой гонку вооружений, в которой точность моделей обнаружения, используемых в платформах голосовой безопасности (таких как Nuance), конкурирует с точностью моделей, генерируемых не защищенными, клонирующими голос ML.

В этом разделе доклада предлагается неисчерпывающий перечень рекомендаций о том, как повысить готовность к борьбе с нынешней и потенциальной будущей эволюцией злонамеренного использования и злоупотребления ИИ.

ИИ во благо:

Используйте потенциал технологии ИИ в качестве инструмента борьбы с преступностью, чтобы в будущем защитить индустрию кибербезопасности и полицию, опираясь на постоянные усилия по обеспечению того, чтобы ИИ был надежным: это означает, что он законен, этически последователен и технически надежен.

Содействовать ответственным инновациям в области ИИ и обмену передовым опытом на таких общественных форумах, как Глобальный саммит

«ИИ во имя добра,» Конференция по IoT и безопасности ИИ Европола-Европейского агентства по сетям и информационной безопасности (ENISA), Конференция по киберпреступности Европола-Интерпола и Глобальное совещание ЮНИКРИ-Интерпола по ИИ для правоохранительных органов.

Повышение киберустойчивости к нынешнему и будущему вредоносному использованию или злоупотреблениям ИИ путем разработки конкретных, углубленных, перспективных оценок угроз и постоянного картирования ландшафта угроз ИИ.

- ° Принять подходы к управлению рисками для классификации угроз, связанных с текущим и будущим использованием и злоупотреблениями ИИ, и соответствующим образом определить приоритеты мер реагирования.

Безопасные рамки проектирования ИИ:

- ° Содействовать разработке надежных систем ИИ в соответствии с принципами безопасности по дизайну и конфиденциальности по дизайну.

- ° Разработайте специальные рамки защиты данных, чтобы обеспечить непрерывную разработку, экспериментирование и обучение систем ИИ в защищенной и контролируемой среде для обеспечения максимальной точности систем ИИ.

- ° Поощрять принятие ориентированного на человека подхода к ИИ, такого как рамочная программа Европейского союза (ЕС) по надежному ИИ.

- ° Установить технические стандарты для содействия кибербезопасности ИИ, такие как усилия Группы отраслевых спецификаций Европейского института телекоммуникационных стандартов (ETSI) по обеспечению безопасности ИИ (SAI).

Политика:

- ° Деэскалация политически нагруженной риторики об использовании ИИ в целях кибербезопасности, которая может препятствовать способности

индустрии кибербезопасности и правоохранительных органов реагировать на события киберпреступности и опережать их.

- ° Содействовать разработке применимых на международном уровне и не зависящих от технологий мер реагирования в области политики для предотвращения потенциального злонамеренного использования ИИ, не препятствуя инновационному и позитивному применению ИИ.

- ° Признаем, что внедрение ИИ в целях кибербезопасности и борьбы с преступностью может оказать влияние на права личности; решаем эти проблемы путем систематического содействия информированным общественным дискуссиям в целях получения общественного согласия и разработки соответствующих мер.

- ° Обеспечить, чтобы использование ИИ в целях кибербезопасности, в том числе правоохранительными органами, соответствовало этическим принципам и подлежало эффективному надзору для обеспечения долгосрочной устойчивости.

Пропаганда:

- ° Использовать партнерские отношения между государственным и частным секторами и содействовать созданию междисциплинарных групп экспертов по кибербезопасности ИИ, таких как Специальная рабочая группа ENISA по кибербезопасности ИИ.¹¹⁸

- ° Повышение грамотности в области ИИ и кибергигиены путем разработки инновационных кампаний по профилактике и повышению осведомленности о киберугрозах ИИ, таких как веб-серия проекта Trend Micro 2020.¹¹⁹

- ° Содействовать наращиванию потенциала, межсекторальному сотрудничеству и обмену знаниями об инструментах, тактиках и методах сдерживания, выявления, смягчения последствий и расследования злоупотреблений ИИ.

Пример из практики: глубокое погружение в Deepfakes

Одним из наиболее заметных вредоносных применений ИИ (и, возможно, наиболее ощутимым, наносящим ущерб приложению) является феномен так

называемых глубоких подделок. Портфолио «глубокого обучения» и «поддельных медиа», deepfakes включает в себя использование методов ИИ для манипулирования или создания визуального и аудиоконтента, который людям или даже технологическим решениям трудно сразу отличить от подлинных.

Эта технология превозносится как мощное оружие в современных войнах за дезинформацию, в результате чего человек больше не может полагаться на то, что он видит или слышит. Кроме того, в сочетании с охватом и скоростью Интернета, социальных сетей и приложений для обмена сообщениями, deepfakes может быстро охватить миллионы людей за чрезвычайно короткий период. Что касается такого короткого периода времени, deepfakes представляют значительный потенциал для целого ряда злонамеренных и преступных целей, которые включают в себя:

- Разрушение имиджа и доверия к личности,
- Преследование или унижение людей в Интернете,
- Совершение вымогательства и мошенничества,
- Содействие мошенничеству с документами,
- Фальсификация онлайн-идентификаций и обман механизмов KYC,
- Фальсификация или манипулирование электронными доказательствами для расследований в области уголовного правосудия,
- Разрушение финансовых рынков,
- Распространение дезинформации и манипулирование общественным мнением,
- Подстрекательство к актам насилия в отношении групп меньшинств,
- Поддержка нарративов экстремистских или даже террористических групп, а также
- Разжигание социальных волнений и политической поляризации.

Другие, более надуманные примеры включают возможность спровоцировать международный дипломатический инцидент, например, путем публикации фейка с участием лидера страны с передовыми ядерными технологиями, объявляющего или призывающего к ядерному удару по другой

стране или территории. Хотя это вряд ли приведет к кинетической ответной реакции, это может на мгновение вызвать массовую панику и замешательство.

Одним из побочных эффектов использования deepfakes для дезинформации является снижение доверия граждан к власти и информационным средствам массовой информации. Наводненные все более генерируемым ИИ спамом и фальшивыми новостями, основанными на фанатичных текстах, поддельных видео и множестве теорий заговора, люди могут чувствовать, что значительному количеству информации, включая видео, просто нельзя доверять, что приводит к явлению, называемому «информационным апокалипсисом» или «апатией реальности.» На самом деле одним из наиболее разрушительных аспектов deepfakes может быть не дезинформация как таковая, но скорее принцип, согласно которому любая информация может быть поддельной. Это оставляет открытой дверь для еще одного вредного аспекта распространения глубоких фейков и фальшивых новостей в целом. В конечном счете, трудности с аутентификацией видео позволяют отрицать любую компрометирующую информацию, поскольку любой аудиовизуальный контент может быть сфабрикован. Другими словами, любой может заявить, что видео является поддельным, тем самым полностью уклоняясь от какой-либо ответственности за свои действия. Вот как deepfakes могут быть использованы не только для того, чтобы угрожать нашей и без того уязвимой информационной экосистеме, но и для того, чтобы подорвать возможность надежной общей «реальности.»

Deepfakes начали набирать популярность примерно в конце 2017 года, когда анонимный пользователь Reddit опубликовал видео с такими популярными фигурами, как Тейлор Свифт, Скарлетт Йоханссон, Обри Плаза, Галь Гадот и Мэйзи Уильямс, чьи лица были наложены на тела женщин в порнографических фильмах. Хотя эти видео были быстро сняты хостинговой платформой, потенциал этой неожиданной техники замены лица быстро привлек внимание средств массовой информации и общественности, что привело к ее распространению в Интернете.

Хотя в этом тематическом исследовании основное внимание уделяется злонамеренному использованию deepfakes, важно отметить, что существует также много положительных применений этой технологии. Например, синтез видео может позволить автомобильной промышленности имитировать автономные автомобильные аварии, чтобы учиться и избегать ошибок в будущем, в то время как киноиндустрия может использовать эту технологию для цифрового воскрешения умерших актеров. Синтез голоса в равной степени может быть использован для того, чтобы «позволить» людям говорить на нескольких языках своим собственным голосом, в то время как медицинские исследователи могут использовать его для разработки искусственных замен голоса для людей, потерявших способность говорить.

Deepfakes создаются путем манипулирования контентом, чтобы создать синтетический аудиовизуальный файл и сделать его подлинным. Технология, используемая для создания deepfake аудиовизуального контента, зависит от различных типов цифровых подделок. В общем, deepfake изображения и видео можно классифицировать следующим образом:

Замена лица: замена лица одного человека, сшивая его с лицом другого человека

Воссоздание лица: манипулирование чертами лица цели, чтобы сделать вид, что они говорят что-то, чего на самом деле нет.

Генерация лиц: создание убедительных, но полностью вымышленных синтетических образов людей

Синтез речи: генерация аудиоконтента с помощью алгоритмов и обучения для создания deepfake голоса или синтетического аудиофайла.

Мелкие подделки: создание аудиовизуальных подделок менее сложного характера с использованием элементарных методов редактирования

Независимо от формата (звук, изображение или видео), синтетические носители могут быть созданы с помощью GANs, инновационного приложения для машинного обучения без контроля, изобретенного в 2014 году, которое произвело революцию в области deepfake. Он состоит из двух искусственных

нейронных сетей: генеративная сеть генерирует новые данные с теми же характеристиками, что и обучающий набор, в то время как дискриминационная сеть отделяет сгенерированные данные от обучающих данных. Например, GAN, обученный на фотографиях, может генерировать новые данные, которые выглядят аналогично набору данных сохраненных фотографий. Дискриминатор получает случайно полученные генератором фотографии из реального набора данных и пытается их идентифицировать. Цель генератора состоит в том, чтобы обмануть сеть дискриминаторов, производя все более лучшие новые кандидаты, в то время как цель дискриминатора состоит в том, чтобы стать еще лучше в обнаружении синтезированных изображений. Этот циклический процесс приводит в действие как дискриминатор, так и генератор, чтобы улучшить создание реальных носителей. GAN также может быть обучен на тысячах фотографий лиц, чтобы создать новый портрет, который приближается к этим фотографиям, не будучи точной копией ни одной из них. Хотя deepfakes обычно требует большого количества изображений для создания реалистичной подделки, в ближайшем будущем GAN будут обучены на меньшем количестве информации и смогут обмениваться головами, целыми телами и голосами.¹²⁷

Другой метод использует более примитивный тип искусственной нейронной сети, называемый автоэнкодером. Эта нейронная сеть состоит из кодера, который сжимает изображение, уменьшая его до латентного пространства более низкого измерения, и декодера, который восстанавливает изображение из латентного представления. Deepfakes использует эту архитектуру, имея две пары кодер-декодер, где каждая пара используется для обучения набору изображений. В процессе обучения один кодировщик находит и изучает сходство между двумя лицами и сводит их к общим чертам, таким как положение глаз, носа и рта. Две сети используют один и тот же кодер, но разные декодеры для каждого изображения. Чтобы проиллюстрировать, один декодер восстанавливает лицо А, а другой декодер восстанавливает лицо В. Чтобы поменять местами лица, изображение лица А кодируется общим кодером и декодируется декодером лица В. Затем декодер восстанавливает

лицо Человека В с выражением и ориентацией Лица А. Чтобы создать убедительное видео, это должно быть сделано на каждом кадре.

Наряду с deepfakes с помощью ИИ на первый план вышли менее сложные формы манипуляций с аудио- и видеозаписями, которые создают возможность для преступного или злонамеренного использования. Как уже упоминалось выше, эти поверхностные подделки, также известные как дешевые подделки, представляют собой видео или аудиозаписи, которые либо представлены вне контекста, либо подделаны с помощью простых инструментов редактирования, таких как намеренное замедление или ускорение темпа, чтобы представить объект в недостоверном виде. Другие возможные манипуляции заключаются в выборочном соединении некоторых неизмененных фрагментов речи, удалении или смягчении тона текста или добавлении контекста путем изменения времени начала и окончания видео. Такое редактирование не изменяет оригинальное содержание, а лишь меняет представление зрителя о нем, тем самым приписывая новый смысл ранее безобидному видео. Хотя они отличаются от глубоких подделок тем, что при их создании не используется ИИ, мелкие подделки, тем не менее, заслуживают признания как часть более широкого явления глубоких подделок.

Как указывалось ранее, аудиозаписи также могут быть глубоко подделаны для создания аудиоконтента, например, «голосовых скинов» или «голосовых клонов» общественных деятелей. Для deepfake аудио движки машинного обучения обучаются на конференц-звонках, YouTube, обновлениях социальных сетей и даже выступлениях TED, чтобы скопировать голосовые модели некоторых целевых лиц, а затем генерировать новые аудиоклипы с теми же голосовыми характеристиками. Поскольку в этом методе можно использовать реальный голос целевых лиц и можно склеивать их слова, сгенерированные алгоритмом голоса могут звучать невероятно реально.

Создание deepfakes: Приложения и инструменты

Модели глубокого обучения могут использоваться злоумышленниками, которые, в свою очередь, могут легко использовать инструменты с открытым исходным кодом, доступные на GitHub, распространяемые на форумах или продаваемые на подпольных рынках. Кроме того, они могут использовать функции, доступные для коммерческого или личного использования крупными компаниями-разработчиками программного обеспечения, такими как NVIDIA и Google. Это означает, что, хотя для разработки такой методики требуются технические знания и понимание вычислительных параметров, большая часть программного обеспечения уже доступна общественности для общего использования.

Однако проблема с созданием глубоких подделок заключается в том, что глубокие затухания жаждут данных, то есть им требуется большое количество аудиовизуальных данных для обучения своих моделей. Однако уже появились порталы услуг, для которых требуется обучающий набор данных размером всего 250 фотографий, что меняет динамику и облегчает создание глубоких фейков. Например, Центр искусственного интеллекта Samsung создал модель, которой требуется всего несколько изображений, чтобы превратить ее в реалистичное видео «говорящей головы.»

Приложения для мобильных телефонов также изменили процесс создания глубоких подделок, позволяя пользователям создавать deepfakes с помощью простых приложений. Примерами этого являются FakeApp, первое приложение, которое внесло значительный вклад в шумиху deepfake в январе 2018 года, FaceApp, российское приложение для старения лица, которое стало вирусным в 2019 году, но быстро столкнулось с проблемами политики конфиденциальности, которые препятствовали его успеху, и, наконец, Zao, китайское приложение, которое позволяет пользователям менять свои лица с персонажами фильмов или телепередач, на которых была обучена система. Есть также несколько веб-сайтов, которые предлагают глубокое создание поддельных по запросу в качестве услуги.

Растущая коммерциализация инструментов, сервисов и приложений помогает обойти техническую сложность этой технологии. Что еще более важно, это снизит барьер для отдельных лиц или групп, которые не являются экспертами, для создания deepfakes, увеличивая возможности для их вредоносного использования в ближайшем будущем.

Как уже отмечалось, ранние примеры глубоких подделок включали актеров, чьи лица были вплетены в порнографические видео. Однако, с тех пор использование deepfakes расширилось, включив политических лидеров, комиков, артистов и так далее.

Помимо их использования в целях порнографии без согласия, было на удивление мало сообщений о злонамеренном использовании deepfakes. На самом деле, некоторые из самых плодотворных видео deepfake, которые стали вирусными, были созданы не со злым умыслом, а скорее с целью повышения осведомленности с целью освещения потенциальной угрозы этой технологии. Так было в случае с видео deepfake бывшего президента США Барака Обамы, который использовал голос актера и режиссера Джордана Пила, который сотрудничал с BuzzFeed в апреле 2018 года для создания видео. Еще один пример – Марк Цукерберг, генеральный директор Facebook, в котором он, по-видимому, говорит о фальшивой политике Facebook в отношении видео.

Хотя существует не так много сообщений об использовании deepfakes в злонамеренных или специально преступных целях, были выявлены некоторые примеры использования deepfake (или предполагаемого использования deepfake). Эти примеры, в частности, наблюдались на политической арене. В 2019 году в Малайзии появилось видео с глубоким фейком, в котором был задействован политический помощник, который, как оказалось, признался в гомосексуалистах с министром кабинета министров. Кроме того, видео включало призыв к тому, чтобы министр был расследован по обвинению в коррупции. Хотя мотив, стоящий за этим видео (помимо диффамации персонажей), остается неясным, ему удалось нанести политический ущерб и дестабилизировать коалиционное правительство.

В конце 2018 года было опубликовано видео с Али Бонго Ондибой, президентом Габона, которого не видели на публике в течение нескольких месяцев и который, как считалось, находился в плохом состоянии здоровья или даже, возможно, умер. По сути, это вызвало национальный кризис.¹⁴¹ Убеждение в том, что видео было *deepfake*, казалось, подтверждало теории о том, что правительство держало истинное состояние президента в тайне. В ответ на это габонские военные устроили неудачный государственный переворот. Хотя кажется, что видео, скорее всего, не было сфабриковано, потенциальные политические последствия в этом случае явно очевидны.

Помимо политической сферы, *deepfakes* продемонстрировали потенциал для нанесения ущерба в других областях. Недавно известный малайзийский актер Зул Ариффин заявил, что вирусное секс-видео в социальных сетях с мужчиной, похожим на него, было *deepfake*.¹⁴² Помимо клеветы, публикация этого видео была тревожной, поскольку распространение или демонстрация порнографии в Малайзии является уголовным преступлением. То, что видео, о котором идет речь, является *deepfake*, однако, не было проверено, и, таким образом, это остается только предполагаемым случаем. Это иллюстрирует один из самых разрушительных аспектов распространения *deepfakes* – и фальшивых новостей в целом, если на то пошло, — который заключается в том, что любой может заявить, что компрометирующий аудиовизуальный контент является поддельным.

Было также замечено, что мелкие подделки представляют политическую угрозу. Например, видеозапись спикера Палаты представителей Соединенных Штатов Нэнси Пелоси была замедлена, создавая ложное впечатление, что она была пьяна или ослаблена. В этом случае мелкий фейк вызвал столько же неистовства в средствах массовой информации и политических споров, сколько и его *deepfake*.

Как уже отмечалось, технология *deepfake* может создавать убедительные, но полностью вымышленные фотографии с нуля. В 2019 году в Интернете появился фантомный профиль LinkedIn под именем «Кэти Джонс». Кэти,

молодой специалист лет тридцати, работала в Центре стратегических и международных исследований и была связана с рядом правительственных чиновников США. Эксперты, изучившие профиль Кэти, отметили его как вымышленный и установили, что ее фотография в профиле была сгенерирована ИИ, созданным, скорее всего, в попытке заманить людей, представляющих интерес, и собрать информацию, возможно, в рамках операции по сбору разведывательных данных.

Одним из наиболее заметных способов использования deepfakes для получения крупной преступной финансовой выгоды является использование deepfake аудио, а не видео или изображения. В марте 2019 года глава британской энергетической компании перевел почти 200 000 британских фунтов стерлингов (примерно 260 000 долларов США на момент написания статьи) на счет в венгерском банке после того, как ему позвонил человек, который использовал технологию deepfake audio, чтобы выдать себя за голос генерального директора фирмы. Тем временем Symantec сообщила, что она видела три успешных поддельных аудио-атаки (включая случай в Великобритании) на частные компании, которые использовали звонок «генерального директора» старшему финансовому сотруднику с просьбой о срочном переводе денег. В другом примере с июля 2020 года NISOS обнаружила аналогичную атаку, направленную на сотрудника технологической компании, с целью убедить их отправить злоумышленникам деньги.¹⁴⁸ Преступники также создали фальшивую аудиозапись бывшего исполнительного директора Европола для получения крупной финансовой выгоды, используя аудиозапись предыдущего публичного интервью, и попытались вымогать 10 000 евро (примерно 11 790 долларов США на момент написания статьи) у жертвы. Однако схема была раскрыта, когда жертва позвонила в Европол, чтобы узнать о возвращении своих денег. Следовательно, им также было рекомендовано больше не отправлять мошеннику деньги.

Аудиозапись deepfake также недавно была передана в суд в Великобритании в битве за опеку над детьми. Фальшивый аудиоклип, в

котором якобы демонстрировался отец, угрожающий другой стороне, был использован в попытке подорвать дело отца, чтобы обеспечить заключение под стражу. Фабрикация доказательств, конечно, является преступным деянием. Уникальные события в этом деле вызывают особенно серьезную озабоченность в отношении достоверности и допустимости аудиовизуальных материалов в качестве электронных доказательств в судах и служат важным предупреждением для следователей по уголовным делам, судов и юристов в отношении аудиовизуальных доказательств в эпоху глубоких подделок.

Еще одним злонамеренным злоупотреблением *deepfakes*, получившим ограниченное внимание средств массовой информации, но имеющим серьезные последствия для безопасности, являются так называемые измененные фотографии паспортов, также известные как «морфинговая атака» или тенденция к изменению лица. Морфинг лица позволяет использовать один паспорт двум или более лицам с помощью процесса, который объединяет их фотографии в одну, которая сохраняет сходство с обоими или всеми лицами. Процесс морфинга становится возможным благодаря использованию GAN, которые могут обмануть как экспертов по идентификации человека, так и ИИ, стоящий за биометрическими системами распознавания лиц. В настоящее время измененные фотографии лиц используются двумя способами, а именно на обманном путем полученных подлинных (поддельных) документах или фальсифицированных удостоверениях личности.

В первом векторе атаки преступники нацеливаются на процесс выдачи удостоверения личности в печатном виде, который требует, чтобы заявитель представил распечатанную фотографию вместе со своей формой заявки (в отличие от процесса регистрации в реальном времени, в котором делается фотография заявителя во время самого процесса подачи заявки). Тем не менее, можно подать мошенническое заявление на продление удостоверения личности, используя измененную фотографию лиц, которые выглядят похожими. Поскольку размер требуемой фотографии невелик, трудно определить, является ли она измененной фотографией. В результате эта

фальсификация способствует выдаче фальшивых документов. Кроме того, поскольку процесс создания поддельных документов, в свою очередь, облегчается фальшивыми документами заводчика (то есть документами, которые служат основой для получения других удостоверений личности, таких как свидетельства о рождении)¹⁵², идентификация поддельного документа на более поздних этапах процесса подачи заявки становится все более сложной задачей.

Второй вектор атаки включает использование измененной фотографии во время процесса фальсификации удостоверения личности. В этом случае фальсификатор может напечатать измененную фотографию поверх подлинной фотографии оригинального удостоверения личности. Этот тип фальсификации также может быть осуществлен с помощью украденных удостоверений личности.

К сожалению, текущие исследования показывают, что как эксперты-люди без специальной подготовки, так и традиционные решения для выявления атак морфинга имеют низкие показатели обнаружения морфинга лица. Следует подчеркнуть, что эта тенденция использования идентификационных данных на основе морфов для мошенничества значительно подрывает работу органов, выдающих паспорта, пограничный контроль и процедуры проверки личности.

Возможные причины низкой скорости принятия Deepfakes

Хотя вышеупомянутые примеры злонамеренного злоупотребления deepfakes были засвидетельствованы, технология создания deepfake еще не используется преступниками в больших масштабах, по сравнению с тем, что можно было бы ожидать с момента ее появления в 2017 году. Основное использование глубоких подделок по-прежнему в подавляющем большинстве случаев используется в порнографических целях без согласия. Примечательно, что в сентябре 2019 года ИИ-фирма Deeptrace обнаружила в Интернете 15 000 видео deepfake, из которых 96% были порнографическими и 99% из которых использовали нанесенные на карту лица женщин-знаменитостей на

порнографических актерах. То, что нет больше доказательств злоупотребления deepfakes преступниками или в преступных целях, поднимает важный вопрос о том, почему, несмотря на потенциальную возможность злонамеренного использования deepfakes, не было больше таких случаев? Вот некоторые потенциальные причины:

Короткий срок службы.

При высокой степени глобализации и взаимосвязанности поддельный контент, как правило, помечается в течение короткого периода времени. В этом отношении поддельный контент, скорее всего, только вызовет немедленный шок, который вскоре исчезнет.

Технические навыки. Хотя технология доступна в инструментах и пакетах, которые облегчают ее реализацию, тем не менее она требует высокой степени мастерства для создания очень убедительных видеороликов. Также вполне вероятно, что преступники будут из всех сил пытаться найти необходимые данные для создания правдоподобного deepfake. С другой стороны, дешевые подделки, которые не требуют таких навыков, как правило, более легко идентифицируются и, следовательно, будут менее эффективным инструментом для преступников.

Более эффективные «традиционные» атаки. Современные криминальные инструменты, технологии и подходы, вероятно, будут более эффективными и простыми в использовании. Они также будут менее дорогостоящими и менее трудоемкими. Следовательно, если цель состоит в том, чтобы получить прибыль от использования deepfakes, есть более простые способы заработать деньги легально, а не незаконно.

Хотя понимание этих факторов может пролить свет на то, почему было относительно мало случаев злонамеренного злоупотребления deepfakes, было бы разумно отметить, что эти факторы не всегда могут быть применимы. По мере того, как технологии становятся более доступными для широкой публики

и становится возможным создавать более убедительные синтетические средства массовой информации, в конечном итоге преступникам больше не потребуется такой высокий уровень технического мастерства, и deepfakes может стать более эффективным средством атаки.

Некоторые возможные будущие угрозы Deepfakes

Рост манипулирования изображениями и голосами в преступных целях был определен многими как потенциальная угроза, включая Европол^{161, 162, 163} и ЮНИКРИ. Хотя крупномасштабное злоупотребление deepfakes злоумышленниками и преступниками, в частности, до настоящего времени не наблюдалось, важно отметить потенциал этой технологии в сочетании с растущей доступностью и повышением качества инструментов для создания deepfake. В результате этих двух условий есть смысл выдвинуть гипотезу о возможных сценариях, которые предполагают, как преступники могут вскоре использовать эту технологию.

В качестве возможных возможностей в ближайшем будущем были определены следующие:

Сценарий 1: Кампании по дезинформации. Deepfakes могут использоваться в рамках массовых аудиовизуальных кампаний дезинформации для создания правдоподобного поддельного контента, который выдает себя за высокопоставленных государственных чиновников, чтобы вмешаться в выборы или спровоцировать насилие или общественные беспорядки. Например, видеоролики deepfake, на которых политик берет взятку или признается в соучастии в преступлении, или солдаты, совершающие военные преступления, или террористическая группа, жестоко нападающая на представителей меньшинства, могут, вероятно, вызвать внутренние беспорядки, протесты и срывы выборов. Они могут даже привести к международным конфликтам, если правительства мобилизуют военных на основе таких сфабрикованных видеозаписей. Хотя deepfakes в целях дезинформации имеют кратковременное эффективное окно, злоумышленники все еще могут использовать их, если само

окно учитывается и строится как часть *modi operandi*. Например, короткое окно *deepfakes* может сделать их решающими инструментами в ситуациях, когда может быть вызвана немедленная эмоциональная реакция, или в обстоятельствах, когда время имеет важное значение, например, в последний день перед выборами.

Сценарий 2: Мошенничество с ценными бумагами. *deepfakes* могут быть использованы в рамках манипуляций на фондовом рынке или мошеннических схем. Например, *deepfake*, на котором главный исполнительный директор говорит расистские или женоненавистнические оскорбления, объявляет о фальшивом слиянии или делает ложные заявления о финансовых потерях или банкротстве, может повлиять на стоимость фондового рынка, а также на *deepfakes*, которые изображают главного исполнительного директора, по-видимому, совершающего преступление. Хотя стоимость акций может значительно восстановиться, как только видео будет заявлено как поддельное, оно, тем не менее, может торговаться по гораздо более низкой цене, чем раньше, что приведет к потере компании, которая растет по мере того, как опасения по поводу подлинности видео сохраняются во всем мире. *Deepfake* высокопоставленного должностного лица компании также может быть использовано для содействия краже коммерческой тайны или конфиденциальных данных компании путем обмана сотрудников компании для отправки конфиденциальной информации преступникам. В свою очередь, такого рода украденная информация может серьезно повлиять на финансы и стоимость акций компании.

Сценарий 3: Вымогательство. Угрожая выпустить поддельное видео, которое нанесет ущерб репутации или авторитету физического или юридического лица, преступники также могут использовать *deepfakes* для вымогательства выкупа или конфиденциальной информации. Также возможно, что использование *deepfakes* в целях вымогательства может сочетаться с атакой вымогателей.¹⁶⁸ Кибератаки, связанные с вымогательством, часто требуют выплаты выкупа в криптовалютах,¹⁶⁹ что еще больше ограничивает

возможности расследования компетентными органами. Поэтому вполне вероятно, что преступники будут сочетать злоупотребление сложными аудиовизуальными глубокими замираниями с вымогательскими атаками, требующими альтернативных средств оплаты, чтобы избежать обнаружения.

Сценарий 4: Преступления в Интернете против детей. Технология Deepfake может быть использована для создания поддельной личности в целях ухода за детьми или для преобразования лица взрослого в лицо ребенка или несовершеннолетнего, что вызывает озабоченность по поводу совершенно новых проблем, связанных с предотвращением сексуальной эксплуатации детей в Интернете. Кроме того, deepfake могут использоваться преступниками для облегчения сексуального принуждения и вымогательства детей в Интернетепутем создания сексуально откровенных материалов, изображающих жертву или других детей. Преступники также могут использовать deepfakes для изображения сексуального контакта в автономном режиме, а затем требовать оплаты, угрожая публично опубликовать обнаженную фотографию или видео deepfake. Эта возможность вызывает особую озабоченность в связи с растущей доступностью самогенерируемых откровенных материалов (SGEM) в Интернете, что обусловлено растущим доступом несовершеннолетних к высококачественным смартфонам и недостаточной осведомленностью о рисках, связанных с распространением откровенного сексуального контента на онлайн-платформах.¹⁷³

Сценарий 5: Воспрепятствование правосудию. Аудиовизуальный контент Deepfake может быть злонамеренно представлен в качестве «законных» доказательств в попытке сорвать уголовные расследования и судебные разбирательства, тем самым ставя под сомнение аудиовизуальные доказательства как целую категорию доказательств. Это создало бы новые юридические препятствия для следователей и адвокатов, не говоря уже о том, чтобы подорвать доверие к судебным разбирательствам, включая учреждения и отдельных лиц, участвующих в одном и том же разбирательстве. Даже административная или судебная система также может быть поставлена под

сомнение. Кроме того, существующие уязвимости в широко используемых технологиях, таких как камеры видеонаблюдения или камеры полицейских органов, могут быть использованы для замены реальных кадров deepfakes. В конечном счете, возможность взлома систем камер наблюдения открывает двери для сокрытия любых видов преступной и вредоносной практики.

Сценарий 6: Кripto-взлом. Это может быть связано с распространением вредоносных программ и криптоджекингом. Недавно исследователи выявили веб-сайт, посвященный deepfakes, который использовался в рамках схемы крипто-взлома. Любители Deepfake, которые относятся к категории людей, которые, как правило, с большей вероятностью имеют мощные компьютеры, чем средний человек, таким образом, являются идеальными целями для криптоджекинга, вредоносной атаки, которая, по-видимому, использует вычислительные ресурсы для добычи ценных криптовалют, таких как биткойн.¹⁷⁴

Сценарий 7: Незаконные рынки. Распространение наборов инструментов и услуг для разработки deepdale в вредоносных целях может стать новым прибыльным подпольным рынком. Услуги по созданию Deepfake уже предлагаются на различных подпольных рынках и онлайн-форумах по скромным ценам. В сочетании с бизнес-моделью SaaS¹⁷⁶ в ближайшем будущем, вероятно, будет расти распространение инструментов и услуг «по найму» в Интернете.

Противодействие Deepfakes

В то время как deepfakes предоставляют возможности для тех, кто имеет злой умысел, и могут стать серьезной угрозой для общества, политических систем, бизнеса и кибербезопасности, все еще существуют меры, которые могут быть приняты для контроля и ограничения воздействия этой технологии. К ним относятся разработка технологий обнаружения глубоких подделок, аутентификация контента и предотвращение глубоких подделок, а также принятие политики, законодательства и нормативных актов, кампании по

повышению осведомленности, наращивание потенциала и создание каналов для сообщения о deepfakes (включая процедуры уведомления и снятия).

Кроме того, ручное обнаружение видео и изображений deepfake, особенно менее сложных, все еще возможно, особенно если модераторы контента или компетентные органы должным образом обучены обнаруживать несоответствия сцены (такие как странные тени и освещение, фоновый шум на открытом воздухе в видео в помещении, внезапные изменения фона и другие примечательные детали), а также несоответствия динамиков (например, отсутствие или нерегулярное мигание, микро-выражения лица, несоответствие между тонами кожи лица и тела, движение губ и несоответствие звука и другие подобные несоответствия). Однако ручная идентификация и фильтрация аудиовизуальных несоответствий человеком возможна только для небольшого объема подозрительных файлов. Он не является ни масштабируемым, ни устойчивым в свете всплеска глубоких подделок и экспоненциально растущего объема аудиовизуальных электронных доказательств в современных расследованиях. Более того, ручной анализ может привести к ошибкам, поскольку люди имеют психологическую предрасположенность верить в аудиовизуальный контент и склонность к истине по умолчанию.

В свете различных инструментов и методов создания глубоких подделок в настоящее время не существует технологического решения, которое было бы адаптировано для эффективного обнаружения всех типов цифровых подделок. По иронии судьбы, ИИ сам по себе может быть ответом на обнаружение deepfake. Наиболее популярным методом является использование алгоритмов, аналогичных тем, которые использовались для создания рассматриваемых глубоких фейков. Цель этой техники состоит в том, чтобы распознать шаблоны, показывающие, как были созданы эти поддельные изображения и аудиоклипы, а также выявить тонкие несоответствия. Инструменты с поддержкой ИИ, такие как FaceForensics++, уже показали точность обнаружения, которая превышает возможности ручного анализа.

Однако обнаружение не лишено своих проблем. В 2018 году американские исследователи обнаружили, что лица на видео *deepfake* обычно не мигают. Поскольку большинство изображений в Интернете показывают людей с открытыми глазами, алгоритмы не были обучены моргать.¹⁷⁹ К сожалению, как только это исследование было опубликовано, появились новые *deepfakes* с улучшенной способностью моргать. Действительно, обнаружение *deepfakes* становится сложнее по мере выявления слабых мест и обнаружения технологий их устранения. Поэтому автоматизированные системы, которые были разработаны для проверки видео на наличие ошибок (таких как нерегулярное мигание, несогласованные схемы освещения, искажения¹⁸⁰ или характерные выражения лица), вскоре устарели.¹⁸¹ Еще одна слабость систем обнаружения ИИ заключается в том, что они лучше всего работают для знаменитостей, поскольку большая часть обучения происходит в базах данных с открытым исходным кодом, и в Интернете значительно больше изображений знаменитостей, чем обычных людей.

Альтернативный метод состоит в применении блокчейна для проверки источника СМИ. С помощью этой технологии будут одобрены только видео из надежных источников, что уменьшит распространение потенциально вредных средств массовой информации *deepfake*. Инструмент *fvb*, использующие блокчейн *Ethereum* с открытым исходным кодом была установлена аутентификация в виде хэшей или мобильных приложений для облегчения надежной цепочки хранения захваченных изображений или видео. Другое решение предлагает использовать блокчейн – подход *Ethereum* для доказательства подлинности цифрового контента, а также для идентификации исходного источника цифрового контента. Он также может быть использован для оценки того, был ли цифровой контент получен из авторитетного источника. Также появились другие решения для проверки происхождения, такие как *Truepic*.

Что касается звуковых вызовов *deepfake*, самый архаичный, но эффективный способ проверить звонящего-повесить трубку и перезвонить.

Если за этим не стоит очень изощренная хакерская группа, которая может перенаправлять телефонные звонки, обнаружить поддельные телефонные звонки может быть относительно легко. Как и при использовании ИИ для обнаружения deepfakes, то же самое можно применить для отличия реальных аудиоклипов от поддельных. Детектор глубоких нейронных сетей использует визуальные представления аудиоклипов, называемые спектрограммами, которые также используются для обучения моделей синтеза речи. В то время как для ничего не подозревающего уха эти клипы звучат в основном одинаково, спектрограммы реального звука по сравнению с поддельным звуком имеют разные представления.

Примечательно, что несколько международных организаций, правительств, университетов и технологических фирм финансируют исследования для разработки инструментов, которые могут обнаруживать deepfakes. Например, ЮНИКРИ организовал вызов на Хакатоне за мир, справедливость и безопасность в 2019 году, чтобы бросить вызов исследовательским группам по всему миру в поисках решения для обнаружения глубоких фейков. Microsoft, Facebook и Amazon также запустили задачу обнаружения Deepfake в 2019 году в надежде ускорить технологию идентификации манипулируемого контента. В сентябре 2020 года Microsoft выпустила Microsoft Video Authenticator, новый инструмент обнаружения deepfake, который был протестирован с использованием наборов данных задачи обнаружения Deepfake. Несмотря на то, что аутентификатор не публикуется публично, он анализирует контент и предоставляет пользователю оценку уверенности в том, что видео могло быть обработано или нет.

Кроме того, индустрия социальных сетей также может сыграть ключевую роль в улучшении обнаружения deepfakes, предоставляя общедоступные данные для надлежащего обучения инструментам обнаружения с поддержкой ИИ. Например, Facebook выпустила базу данных с открытым исходным кодом из 100 000 deepfakes для обучения моделей ИИ распознавать манипулируемые видео.

По мере того, как степень сложности для создания технологии deepfake продолжает улучшаться, в конечном итоге потребуется комбинация методов, позволяющих обнаруживать синтетические или поддельные видео и изображения.

Вредоносный потенциал высокореалистичных видео deepfake вызвал тревогу в цифровом мире. В результате несколько крупных социальных платформ пересмотрели свои пользовательские политики, чтобы запретить пользователям загружать или делиться deepfakes.

После скандала с участием нескольких известных женских личностей, породивших феномен deepfake, Reddit запретил «поддельное порно» – фотографии и видео deepfake, которые накладывают лицо субъекта на откровенную фотографию или видео без разрешения субъекта – в начале 2018 года. Совсем недавно, в январе 2020 года, платформа обновила свою политику, чтобы запретить выдавать себя за физическое или юридическое лицо вводящим в заблуждение или вводящим в заблуждение способом, охватывающим deepfakes. Новые правила были приняты всего через несколько дней после того, как Facebook обновил свою собственную политику платформы, чтобы запретить видео deepfake. Новая политика Facebook, однако, не распространяется на видео, манипулируемые ради пародии или сатиры, и не включает в себя поверхностные фейки, такие как спорное видео Пелоси, факт, который вызвал критику со стороны политических лидеров и цифровых экспертов. Вскоре после этого Twitter также решил отметить манипулируемые видео, включая deepfakes и shallowfakes, а также запретить аккаунтам «обманчиво» делиться отредактированными видео с намерением причинить вред. Тем не менее, одна из проблем, связанных с принципами разработки политики для deepfakes, заключается в том, что со временем станет труднее провести грань между тем, что причиняет вред, и тем, что является законной пародией.

Тем временем на национальном уровне велись дискуссии о регулировании глубоких подделок, главным образом на основе опасений, что

технология может быть использована для вмешательства в политические процессы.

Например, Канадское учреждение безопасности связи опубликовало отчет, в котором говорится, что deepfakes могут быть использованы для вмешательства в канадскую политику, в частности для дискредитации политиков и влияния на избирателей. Несмотря на отсутствие конкретных канадских законов, запрещающих deepfakes, было высказано предположение, что существующие причины действий могут быть применимы для устранения нарушений, совершенных в результате злоупотребления технологией deepfake, а именно: «нарушение авторских прав, диффамация, нарушение неприкосновенности частной жизни, присвоение личности, уголовный кодекс, жалоба на права человека, умышленное причинение психических страданий и преследование.»

В ноябре 2019 года китайские регуляторы объявили о новых правилах, регулирующих видео- и аудиоконтент в Интернете, включая запрет на публикацию и распространение ложной информации или фальшивок в Интернете без надлежащего раскрытия того, что данный пост был создан с помощью технологии ИИ или виртуальной реальности (VR). Китайское правительство заявило, что начиная с 2020 года нераскрытие информации об использовании такой технологии для создания поста будет считаться уголовным преступлением.

В США несколько законопроектов, касающихся регулирования deepfakes, были представлены федеральным законодателям, хотя ни один из этих законопроектов не был принят.

По мере развития решений для обнаружения глубоких подделок лица и группы, стоящие за злоупотреблением deepfakes, в равной степени должны адаптировать свои методы работы с целью уклонения от обнаружения и обучения своих моделей мерам противодействия обнаружению. В этой связи Deepfakes может стать серьезной проблемой для современных методов судебного аудиовизуального анализа и аутентификации, используемых

промышленными предприятиями, компетентными органами, специалистами средств массовой информации и гражданским обществом.

Необходимо не только сделать значительные инвестиции в разработку новой технологии скрининга для выявления фальсификаций и нарушений в визуальном и аудиоконтенте, но и обеспечить, чтобы такие инструменты поддерживались в актуальном состоянии с учетом развития технологии создания глубоких подделок, включая текущее и возможное в будущем злонамеренное злоупотребление такой технологией. Эффективные алгоритмы обнаружения также должны опираться на существующие решения и цифровые методы судебной экспертизы в области анализа фотографий, аудио и видео. Для оптимизации усилий и устранения существующих пробелов разработка систем борьбы с глубокими подделками должна осуществляться на основе сотрудничества между промышленностью и конечными пользователями из компетентных органов (таких как полицейские следователи, криминалисты, судебные эксперты и эксперты и другие аналогичные агенты) и в сочетании с обучением и наращиванием потенциала этих компетентных органов, модераторов контента и других соответствующих специалистов.

Кроме того, следует применять новые меры политики и законодательства. В частности, эта политика должна быть технологически не зависящей от того, чтобы быть эффективной в долгосрочной перспективе и избегать необходимости регулярно пересматривать и заменять их по мере развития технологии, лежащей в основе создания и злоупотребления deepfakes. Тем не менее, такие меры также должны избегать препятствий для позитивного применения GANS.

Следует также рассмотреть вопрос о включении злоупотребления deepfakes в текущие и будущие инициативы на международном уровне по борьбе с незаконным контентом в Интернете, уделяя особое внимание партнерству между государственным и частным секторами, процедурам уведомления и изъятия, активному использованию технологий обнаружения и более тесному сотрудничеству с компетентными органами.

Понимание технологии создания deepfakes на ранней стадии значительно улучшит возможности как для обнаружения, так и для разработки эффективных контрмер. Таким образом, выполнение функции технологического наблюдения имеет решающее значение. Роль академического и исследовательского сообщества также жизненно важна, поскольку существует растущая потребность в исследованиях по deepfakes. Специалисты отрасли и правоохранительных органов также должны постоянно разрабатывать прогнозные оценки угроз в отношении текущего и будущего злоупотребления deepfakes злоумышленниками. Лучшее понимание угрозы может также послужить основой для потенциальных политических и законодательных мер по решению общих проблем, связанных с расследованием и противодействием злоупотреблениям deepfake.

Несмотря на возросшую популярность глубоких подделок, исследования показывают, что многие люди по-прежнему в значительной степени не знают о существовании таких цифровых подделок. Поэтому активизация усилий по повышению осведомленности о подделках и связанных с ними угрозах имеет решающее значение для расширения возможностей общественности, различных отраслей промышленности и компетентных органов по выявлению поддельных изображений, видео и аудиофайлов. В долгосрочной перспективе повышенная осведомленность будет способствовать дальнейшему ограничению потенциального злоупотребления deepfakes.

Этот совместный доклад Trend Micro, ЮНИКРИ и Европола был инициирован простым вопросом: «Был ли кто-нибудь свидетелем каких-либо примеров злоупотребления преступниками искусственным интеллектом?» В стремлении ответить на этот вопрос в докладе удалось обеспечить коллективное понимание текущего и будущего злонамеренного использования и злоупотреблений ИИ.

ИИ обещает большую эффективность и более высокий уровень автоматизации и автономности. Будучи подотраслью информатики, имеющей множество взаимосвязей с другими дисциплинами, ИИ по своей сути является

технологией двойного назначения, лежащей в основе так называемой четвертой промышленной революции. В результате этой двойственности, в то время как она может принести огромную пользу обществу и помочь решить некоторые из самых больших проблем, с которыми мы сталкиваемся в настоящее время, ИИ также может обеспечить целый ряд цифровых, физических и политических угроз. Поэтому необходимо хорошо понимать риски и потенциальные преступные злоупотребления системами ИИ, чтобы защитить не только общество, но и важнейшие отрасли промышленности и инфраструктуры от злоумышленников.

Основываясь на имеющихся данных, исследованиях и структурированном анализе с открытым исходным кодом, этот отчет охватывал нынешнее состояние злонамеренного использования и злоупотреблений ИИ, включая вредоносное ПО ИИ, угадывание паролей с поддержкой ИИ, шифрование с помощью ИИ и атаки социальной инженерии. В нем также описывались конкретные сценарии будущего, начиная от автоматизированного создания и анализа контента, разведки с использованием ИИ, интеллектуальных и подключенных технологий, таких как беспилотные летательные аппараты и автономные автомобили, до манипуляций на фондовом рынке с использованием ИИ, а также методов обнаружения и защиты систем на основе ИИ.

Используя одно из наиболее заметных вредоносных применений ИИ – феномен так называемых глубоких подделок – в докладе далее подробно описывается тематическое исследование использования методов ИИ для манипулирования или создания визуального и аудиоконтента, который людям или даже технологическим решениям было бы трудно сразу отличить от подлинных.

Как предполагается в этой статье, преступники, скорее всего, будут использовать ИИ для облегчения и улучшения своих атак, максимизируя возможности для получения прибыли в течение более короткого периода, эксплуатируя больше жертв и создавая новые, инновационные модели

преступного бизнеса — все это в то же время снижает их шансы быть пойманными. Следовательно, по мере того, как «ИИ как услуга» становится все более распространенным, он также снизит барьер для входа, сократив навыки и технические знания, необходимые для облегчения атак. Короче говоря, это еще больше усугубляет возможность злоупотребления ИИ преступниками и превращения его в движущую силу будущих преступлений.

Хотя атаки, описанные здесь, в основном теоретические, созданные в качестве доказательств концепции на данном этапе, и использование ИИ для повышения эффективности вредоносных программ все еще находится в зачаточном состоянии, вполне вероятно, что разработчики вредоносных программ уже используют ИИ более запутанными способами, не будучи обнаруженными исследователями и аналитиками. Например, разработчики вредоносных программ уже могут полагаться на методы, основанные на ИИ, чтобы обойти фильтры спама, избежать функций обнаружения антивирусного программного обеспечения и сорвать анализ вредоносных программ. На самом деле DeepLocker, инструмент, недавно представленный IBM и обсуждаемый в этой статье, уже демонстрирует эти возможности атаки, которые защитнику было бы трудно остановить.

Кроме того, ИИ может также улучшить традиционные методы взлома, введя новые способы выполнения атак, которые людям будет трудно предсказать. Они могут включать полностью автоматизированное тестирование на проникновение, улучшенные методы подбора паролей, инструменты для взлома систем безопасности капчи или улучшенные атаки социальной инженерии. Что касается инструментов с открытым исходным кодом, предоставляющих такие функциональные возможности, в документе обсуждались некоторые из них, которые уже были представлены, такие как DeepHack, DeepExploit и XEvil.

Между тем широкое использование помощников ИИ также создает возможности для преступников, которые могут использовать присутствие этих помощников в домашних хозяйствах. Например, преступники могут

проникнуть в умный дом, похитив систему автоматизации с помощью открытых аудиоустройств.

Учитывая обсуждения в этой статье, можно с уверенностью предположить, что киберпреступники будут постепенно интегрировать методы ИИ для расширения масштабов и масштабов своих атак, тем самым используя ИИ как вектор атаки, так и поверхность атаки, дополнительно опираясь на ранее упомянутую модель преступного бизнеса на основе услуг. Следовательно, во второй части доклада рассматривались вероятные будущие сценарии, в которых преступники могут злоупотреблять технологиями ИИ для облегчения своей деятельности и повышения успеха своих атак.

Например, ожидается, что ИИ поможет преступникам проводить социальную инженерию в масштабе, автоматизируя первые шаги атаки с помощью создания контента с помощью ИИ, улучшая сбор бизнес-аналитики и ускоряя скорость обнаружения как потенциальных жертв, так и атак на компрометацию бизнес-процессов.

Более того, поскольку оборонные системы все чаще используют подходы, основанные на ИИ, преступники, как ожидается, либо непосредственно атакуют эти системы, либо также используют ИИ для эмуляции поведения пользователей.

Растущая автоматизация многочисленных аспектов повседневной жизни с использованием систем, основанных на ИИ, также неизбежно влечет за собой возможную потерю контроля над теми же аспектами. Это не только расширило бы поверхность атаки, но и создало бы новые типы атак. Одним из новых примеров является манипулирование фондовым рынком с использованием ИИ, которое использует преимущества использования алгоритмов на основе ИИ в области высокочастотной торговли.

Между тем, текущие и будущие технологии, такие как 5G, в сочетании с ИИ будут формировать промышленность и в дальнейшем стимулировать автоматизацию и интеллектуальные технологии в масштабе.

Предположительно, преступники, также нацелены на эти технологии или манипулируют ими.

В ответ на различные сценарии, описанные здесь, в докладе были определены возможные контрмеры, способы снижения рисков и ряд других рекомендаций.

Следует также подчеркнуть, что тесное сотрудничество с промышленностью и научными кругами является необходимым условием для развития знаний и повышения осведомленности о потенциальном использовании и злоупотреблении ИИ преступниками. Такое сотрудничество не только позволит предвидеть злонамеренные и преступные действия, которым способствует ИИ, но и предотвратит, отреагирует или смягчит последствия этих атак упреждающим образом.

Наконец, несмотря на то, что в этом докладе основное внимание уделяется неправильному использованию ИИ, нельзя отрицать, что ИИ обладает огромным потенциалом для позитивного применения, включая критическую поддержку в расследовании преступлений всех видов и выполнении важных проектов организациями с глобальным воздействием. Однако потенциал этих приложений лучше всего подробно изучить в другой статье, и мы с нетерпением ждем возможности прочитать и извлечь уроки из будущей работы, которая будет представлена как промышленностью, так и научными кругами, о бесконечных возможностях, которыми обладает ИИ на благо всех.

Действительно, понимание возможностей, сценариев и векторов атак является ключом к повышению готовности, повышению устойчивости и обеспечению позитивного использования ИИ для раскрытия потенциала, которым он обладает.

Ссылки

<https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe>.

High-Level Expert Group on Artificial Intelligence (AI HLEG). (Apr. 8, 2019). *European Commission*. «Ethics Guidelines for Trustworthy AI.» Accessed on Jul. 16, 2020, at <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top>.

Joint Research Centre (JRC). (2020). *European Commission*. «AI Watch Defining Artificial Intelligence.» Accessed on Jul. 16, 2020, at <https://ec.europa.eu/jrc/en/publication/ai-watch-defining-artificial-intelligence>.

Europol. (Jul. 18, 2019). *Europol*. «Do Criminals Dream of Electric Sheep?» Accessed on Oct. 14, 2020, at <https://www.europol.europa.eu/newsroom/news/do-criminals-dream-of-electric-sheep-how-technology-shapes-future-of-crime-and-law-enforcement>.

Europol. (Oct. 5, 2020). *Europol*. «Internet Organised Crime Threat Assessment (IOCTA) 2020.» Accessed on Oct. 14, 2020, at <https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020>.

High-Level Expert Group on Artificial Intelligence (AI HLEG). (Apr. 8, 2019). *European Commission*. «Ethics Guidelines for Trustworthy AI.» Accessed on Jul. 16, 2020, at <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top>.

Europol. (Feb. 28, 2017). *Europol*. «European Union Serious and Organised Crime Threat Assessment (SOCTA).» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/activities-services/main-reports/european-union-serious-and-organised-crime-threat-assessment-2017>.

Europol. (Sep. 29, 2014). *Europol*. «The Internet Organised Crime Threat Assessment (IOCTA) 2014.» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2014>.

Europol. (Feb. 28, 2017). *Europol*. «European Union Serious and Organised Crime Threat Assessment (SOCTA).» Accessed on Jul. 20, 2020, at

<https://www.europol.europa.eu/activities-services/main-reports/european-union-serious-and-organised-crime-threat-assessment-2017>.

Europol and Eurojust. (July 5, 2019). *Europol and Eurojust*. «Common Challenges in Combating Cybercrime.» Accessed on Oct. 14, 2020, at <https://www.europol.europa.eu/publications-documents/common-challenges-in-combating-cybercrime>.

Europol. (Jul. 18, 2019). *Europol*. «Do Criminals Dream of Electric Sheep?» Accessed on Oct. 14, 2020, at <https://www.europol.europa.eu/newsroom/news/do-criminals-dream-of-electric-sheep-how-technology-shapes-future-of-crime-and-law-enforcement>.

Europol. (n.d.). *Europol*. «Joint Cybercrime Action Taskforce (J-CAT).» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/activities-services/services-support/joint-cybercrime-action-taskforce>.

International Crime Court. (n.d.). *International Crime Court*. «About.» Accessed on Jul. 20, 2020, at <https://www.icc-cpi.int/about>.

Europol. (n.d.). *Europol*. «EC3 Partners.» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/about-europol/european-cybercrime-centre-ec3/ec3-partners>.

European Commission. (Feb. 19, 2020). *European Commission*. «White Paper on Artificial Intelligence – a European approach to excellence and trust.» Accessed on Oct. 1, 2020, at <https://ec.europa.eu/digital-single-market/en/news/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence>.

Sean Palka and Damian McCoy. (n.d.). *Usenix*. «Fuzzin E-mail Filters with Generative Grammars and N-Gram Analysis.» Accessed on Jul. 20, 2020, at <https://www.usenix.org/system/files/conference/woot15/woot15-paper-palka.pdf>.

Ankit Singh and Vijay Thaware. (Jul. 26, 2017). *Black Hat USA*. «Wire Me Through Machine Learning.» Accessed on Jul. 20, 2020, at <https://www.blackhat.com/us-17/briefings/schedule/#wire-me-through-machine-learning-6749>.

Chanil Jeon, et al. (Jul. 27, 2019). *Black Hat USA*. «AVPASS: Leaking and Bypassing Antivirus Detection Model Automatically.» Accessed on Jul. 20, 2020, at

<https://www.blackhat.com/us-17/briefings/schedule/#avpass-leaking-and-bypassing-antivirus-detection-model-automatically-7354>.

Hyrum Anderson. (Jul. 27, 2017). *Black Hat USA*. «Bot vs. Bot for Evading Machine Learning Malware Detection.» Accessed on Jul. 20, 2020, at <https://www.blackhat.com/us-17/briefings/schedule/#bot-vs-bot-for-evading-machine-learning-malware-detection-6461>.

Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. (May 1, 1996). *Journal of Artificial Intelligence Research*. «Reinforcement Learning: A Survey.» Accessed on Oct. 13, 2020, at <https://www.jair.org/index.php/jair/article/view/10166>.

Dan Petro and Ben Morris. (2017). *DefCon*. «Weaponizing Machine Learning: Humanity was Overrated Anyway.» Accessed on Jul. 20, 2020, at <https://www.defcon.org/html/defcon-25/dc-25-speakers.html#Petro>.

Do Son. (Apr. 7, 2018). *Penetration Testing*. «Deep Exploit: Fully automatic penetration test tool using Machine Learning.» Accessed on Jul. 20, 2020, at <https://securityonline.info/deep-exploit/>.

Arthur Juliani. (Dec. 17, 2016). *Medium*. «Simple Reinforcement Learning with Tensorflow Part 8: Asynchronous Actor-Critic Agents (A3C).» Accessed on Jul. 20, 2020, at <https://medium.com/emergent-future/simple-reinforcement-learning-with-tensorflow-part-8-asynchronous-actor-critic-agents-a3c-c88f72a5e9f2>.

Dhilung Kirat, Jiyong Jang, and Marc Ph. Stoecklin. (Aug. 9, 2018). *Black Hat USA*. «DeepLocker — Concealing Targeted Attacks with AI Locksmithing.» Accessed on Jul. 20, 2020, at <https://www.blackhat.com/us-18/briefings/schedule/index.html#deeplocker---concealing-targeted-attacks-with-ai-locksmithing-11549>.

Virus Total. (n.d.). *VirusTotal*. «YARA — The pattern matching swiss knife for malware researchers (and everyone else).» Accessed on Jul. 20, 2020, at <https://virustotal.github.io/yara/>.

Stephen Hilt. (Aug. 27, 2017). *Trend Micro*. «The Sound of a Targeted Attack.» Accessed on Jul. 20, 2020, at [https:// documents.trendmicro.com/assets/pdf/The-Sound-of-a-Targeted-Attack.pdf](https://documents.trendmicro.com/assets/pdf/The-Sound-of-a-Targeted-Attack.pdf).

Craig S. Smith. (May 10, 2018). *The New York Times*. «Alexa and Siri Can Hear This Hidden Command. You Can't.» Accessed on Jul. 20, 2020, at <https://www.nytimes.com/2018/05/10/technology/alexa-siri-hidden-command-audio-attacks.html>.

Stephen Hilt, et al. (Mar. 5, 2019). *Trend Micro*. «Cybersecurity Risks in Complex IoT Environments: Threats to Smart Homes, Buildings and Other Structures.» Accessed on Jul. 20, 2020, at https://documents.trendmicro.com/assets/white_papers/wp-cybersecurity-risks-in-complex-iot-environments.pdf?_ga=2.20408120.1686390182.1575332885-267397288.1557129106.

Hashcat (n.d.). *Hashcat*. «Hashcat Advanced Password Recovery.» Accessed on Aug. 26, 2020, at <https://hashcat.net/hashcat/>.

GitHub. (n.d.) *GitHub, Inc.* «openwall/john.» Accessed on Aug. 26, 2020, at <https://github.com/openwall/john>.

Jason Brownlee. (Jul. 12, 2019). *Machine Learning Mastery*. «Impressive Applications of Generative Adversarial Networks (GANs).» Accessed on Jul. 20, 2020, at <https://machinelearningmastery.com/impressive-applications-of-generative-adversarial-networks/>.

Briland Hitaj, et al. (Feb. 14, 2019). *arXiv*. «PassGAN: A Deep Learning Approach for Password Guessing.» Accessed on Jul. 20, 2020, at <https://arxiv.org/pdf/1709.00440.pdf>.

Botmaster Labs. (n.d.). *Botmaster Labs*. «XEvil.» Accessed on Aug. 26, 2020, at <https://xevil.net/en/>.

Europol (Jan. 18, 2019). *Europol*. «First report of the observatory function on encryption.» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/publications-documents/first-report-of-observatory-function-encryption>.

Martfn Abadi and David G. Andersen. (Oct. 24, 2016). *arXiv*. «Learning to Protect Communications with Adversarial Neural Cryptography.» Accessed on Jul. 20, 2020, at <https://arxiv.org/pdf/1610.06918.pdf>.

Europol and Eurojust. (Feb. 18, 2020). *Europol and Eurojust*. «Second report of the observatory function on encryption.» Accessed on Oct. 15, 2020, at <https://www.europol.europa.eu/publications-documents/second-report-of-observatory-function-encryption>.

Criminal Use of Information Hiding (CUING). (n.d.). *Europol European Cybercrime Centre (EC3)*. «Steganography to cybercriminals exploitation.» Accessed on Jul. 20, 2020, at <https://CUING.org/>.

Europol and Eurojust. (Feb. 18, 2020). *Europol and Eurojust*. «Second report of the observatory function on encryption.» Accessed on Oct. 15, 2020, at <https://www.europol.europa.eu/publications-documents/second-report-of-observatory-function-encryption>.

Mayra Rosario Fuentes and Fernando Mercés. (Oct. 29, 2019). *Trend Micro*. «Cheats, Hacks, and Cyberattacks: Threats to the Esports Industry in 2019 and Beyond.» Accessed on Jul. 20, 2020, at https://documents.trendmicro.com/assets/white_papers/wp-threats-to-the-esports-industry-in-2019-and-beyond.pdf.

Jennifer Carole. (Dec. 21, 2018). *Bromium*. «Laundering Via In-Game Currency and Goods is on the Rise | Part 2.» Accessed on Jul. 20, 2020, at <https://threatresearch.ext.hp.com/laundrying-via-gaming-currency-and-goods-part-2/>.

Zachary Baker. (Feb. 28, 2020). *Journal of Law and International Affairs*. «Gaming the System: Money Laundering through Microtransactions and In-Game Currencies.» Accessed on Jul. 20, 2020, at <https://sites.psu.edu/jlia/gaming-the-system-money-laundering-through-microtransactions-and-in-game-currencies/>.

Anton Moiseienko and Kayla Izenman. (Oct. 11, 2019). *RUSI Registered Charity*. «Gaming the System: Money Laundering Through Online Games.» Accessed on Jul. 20, 2020, at <https://rusi.org/publication/rusi-newsbrief/gaming-system-money-laundering-through-online-games>.

OliverAi. (Nov. 26, 2019). *Youtube*. «LeagueAI: An AI Playing League of Legends using Deep Learning.» Accessed on Jul. 20, 2020, at https://www.youtube.com/watch?time_continue=92&v=iB4PoNJUXzc.

Europol. (Oct. 5, 2020). *Europol*. «Internet Organised Crime Threat Assessment (IOCTA) 2020.» Accessed on Oct. 14, 2020, at <https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020>.

GitHub. (n.d.) *GitHub, Inc.* «ThoughtfulDev/EagleEye.» Accessed on Jul. 20, 2020, at <https://github.com/ThoughtfulDev/EagleEye>.

GitHub. (n.d.). *GitHub, Inc.* «CorentinJ/Real-Time-Voice-Cloning.» Accessed on Jul. 20, 2020, at <https://github.com/CorentinJ/Real-Time-Voice-Cloning>.

Europol's European Cybercrime Centre (EC3). (2019). *2019 Predictions — The Year of Artificial Intelligence*. Europol.

Erik Zouave, et al. (Mar. 4, 2020). *Swedish Defence Research Agency (FOI)*. «Artificially Intelligent Cyberattacks.» Accessed on Jul. 20, 2020, at https://www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf.

Tom Simonite. (Jun. 11, 2020). *Wired*. «OpenAI's Text Generator Is Going Commercial.» Accessed on Jul. 20, 2020, at <https://www.wired.com/story/openai-text-generator-going-commercial/>.

OpenAI. (n.d.). *OpenAI*. Accessed on Jul 10, 2020 at <https://openai.com/>.

ESET. (n.d.). *ESET*. «Can Artificial Intelligence Power Future Malware?» Accessed on Jul. 20, 2020, at <https://www.eset.com/me/whitepapers/can-artificial-intelligence-power-future-malware/>.

ESET. (n.d.). *ESET*. «Can Artificial Intelligence Power Future Malware?» Accessed on Jul. 20, 2020, at <https://www.eset.com/me/whitepapers/can-artificial-intelligence-power-future-malware/>.

Han Zhang, et al. (Dec. 10, 2016). *arXiv*. «StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks.» Accessed on Jul. 20, 2020, at <https://arxiv.org/abs/1612.03242>.

Jason Brownlee. (Jul. 12, 2019). *Machine Learning Mastery*. «Impressive Applications of Generative Adversarial Networks (GANs).» Accessed on Jul. 20, 2020, at <https://machinelearningmastery.com/impressive-applications-of-generative-adversarial-networks/>.

TypingDNA. (n.d.). *TypingDNA*. Accessed on Aug. 26, 2020, at <https://www.typingdna.com/>.

Aniello Castiglione, et al. (2014). *ScienceDirect*. «A botnet-based command and control approach relying on swarm intelligence.» Accessed on Aug. 26, 2020, at <https://doi.org/10.1016/j.jnca.2013.05.002>.

Aniello Castiglione et al. (2014). *ScienceDirect*. «A botnet-based command and control approach relying on swarm intelligence.» Accessed on Aug. 26, 2020, at <https://doi.org/10.1016/j.jnca.2013.05.002>.

European Cybercrime Centre (EC3). (May 7, 2015). *Europol*. «Cybercrime Dependencies Map.» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/publications-documents/cybercrime-dependencies-map>.

James Bridle. (Mar. 14, 2017). *James Bridle*. «Autonomous Trap 001.» Accessed on Sep. 16, 2020, at <https://jamesbridle.com/works/autonomous-trap-001>.

James Bridle. (Mar. 14, 2017). *James Bridle*. «Autonomous Trap 001.» Accessed on Sep. 16, 2020, at <https://jamesbridle.com/works/autonomous-trap-001>.

Katyanna Quach. (Feb. 20, 2020). *The Register*. «Researchers trick Tesla into massively breaking the speed limit by sticking a 2-inch piece of electrical tape on a sign.» Accessed on Jul. 20, 2020, at https://www.theregister.com/2020/02/20/tesla_ai_tricked_85_mph/.

European Network and Information Security Agency (ENISA). (Nov. 25, 2019). *European Network and Information Security Agency (ENISA)*. «ENISA good practices for security of smart cars.» Accessed on Jul. 20, 2020, at <https://www.ENISA.europa.eu/publications/smart-cars>.

SAE International (Jun. 15, 2018) *SAE International*. «Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.» Accessed on Jul. 20, 2020, at https://www.sae.org/standards/content/j3016_201806/.

Evelyn Cheng. (Nov. 22, 2019). *CNBC*. «Self-driving trucks likely to hit the roads before passenger cars.» Accessed on Jul. 20, 2020, at

<https://www.cnn.com/2019/11/22/self-driving-trucks-likely-to-hit-the-roads-before-passenger-cars.html>.

Bernard Marr. (Jun. 11, 2019). *Forbes*. «The Incredible Autonomous Ships of the Future: Run By Artificial Intelligence Rather than a Crew.» Accessed on Jul. 20, 2020, at <https://www.forbes.com/sites/bernardmarr/2019/06/05/the-incredible-autonomous-ships-of-the-future-run-by-artificial-intelligence-rather-than-a-crew/#2481d4fe6fbf>.

Tiffany Blake. (Ed.). (Feb. 11, 2020). *National Aeronautics and Space Administration (NASA)*. «What is Unmanned Aircraft Systems Traffic Management?» Accessed on Aug. 27, 2020, at <https://www.nasa.gov/ames/utm/>.

GSMA. (Nov. 23, 2018). *GSMA*. «Using Mobile Networks to Coordinate Unmanned Aircraft Traffic.» Accessed on Jul. 20, 2020, at <https://www.gsma.com/iot/wp-content/uploads/2018/11/Mobile-Networks-enabling-UTM-v5NG.pdf>.

Lily Hay Newman. (Jun. 20, 2019). *Wired*. «The Global Hawk Drone Iran Shot Down Was a \$220M Surveillance Monster.» Accessed on Jul. 20, 2020, at <https://www.wired.com/story/iran-global-hawk-drone-surveillance/>.

Europol. (n.d.). *Europol*. «Sim Swapping — A Mobile Phone Scam.» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/activities-services/public-awareness-and-prevention-guides/sim-swapping-%E2%80%93-mobile-phone-scam>.

Frederic Lardinois. (Jun. 5, 2019). *TechCrunch*. «A first look at Amazon's new delivery drone.» Accessed on Jul. 20, 2020, at <https://techcrunch.com/2019/06/05/a-first-look-at-amazons-new-delivery-drone/>.

Marco Mancini. (Jan. 8, 2015). *Hackster.io*. «Wifi Hacking Drone.» Accessed on Oct. 26, 2020, at <https://www.hackster.io/MakerMark/wifi-hacking-drone-045041>.

European Union. (May 24, 2019). *EUR-Lex*. «Consolidated text: Commission Implementing Regulation (EU) 2019/947 of 24 May 2019 on the rules and procedures for the operation of unmanned aircraft (Text with EEA relevance).» Accessed on Aug. 27, 2020, at https://eur-lex.europa.eu/eli/reg_impl/2019/947/.

European Union. (Jun. 4, 2020). *EUR-Lex*. «Commission Implementing Regulation (EU) 2020/746 of 4 June 2020 amending Implementing Regulation (EU) 2019/947 as

regards postponing dates of application of certain measures in the context of the COVID-19 pandemic (Text with EEA relevance).» Accessed on Aug. 27, 2020, at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32020R0746>.

Dominik Gerstner and Dietrich Oberwittler. (Jan. 14, 2018). *SAGE Journals*. «Who's hanging out and what's happening? A look at the interplay between unstructured socializing, crime propensity and delinquent peers using social network data.» Accessed on Jul. 20, 2020, at <https://doi.org/10.1177/1477370817732194>.

Thomas Brewster. (Jan. 29, 2020). *Forbes*. «Remember FindFace? The Russian Facial Recognition Company Just Turned On A Massive, Multimillion-Dollar Moscow Surveillance System.» Accessed on Oct. 26, 2020, at <https://www.forbes.com/sites/thomasbrewster/2020/01/29/findface-rolls-out-huge-facial-recognition-surveillance-in-moscow-russia/#4309c5a3463b>.

Lauren M. Johnson. (Aug. 27, 2019). *CNN*. «A drone was caught on camera delivering contraband to an Ohio prison yard.» Accessed on Jul. 20, 2020, at <https://edition.cnn.com/2019/09/26/us/contraband-delivered-by-drone-trnd/index.html>.

Nicole Hensley. (Jun. 18, 2020). *Daily News*. «Drones being deployed to South Africa mines equipped with lasers, paintball guns.» Accessed on Jul. 20, 2020, at <https://www.nydailynews.com/news/world/drones-deployed-south-africa-mines-article-1.1834649>.

Sandrine Amiel. (May 3, 2019). *Euronews*. «Yellow vests, Black Blocs, casseurs: what's the difference? Euronews explains.» Accessed on Aug. 27, 2020, at <https://www.euronews.com/2019/05/02/yellow-vests-black-blocs-casseurs-what-s-the-difference-euronews-explains>.

James Pasley. (Oct. 8, 2019). *Business Insider*. «12 US states and 7 countries that have barred protesters from wearing masks.» Accessed on Aug. 27, 2020, at <https://www.businessinsider.com/countries-states-where-protesters-cant-wear-masks-2019-10>.

Ming Lee Newcomb. (Jul. 9, 2019). *Consequence of Sound*. «It turns out that Juggalo makeup blocks facial recognition technology.» Accessed on Jul. 20, 2020, at <https://consequenceofsound.net/2019/07/juggalo-makeup-facial-recognition/>.

Geordie Gray. (Jul. 5, 2018). *Tone Deaf*. «Facial recognition technology doesn't recognise Juggalo Makeup.» Accessed on Jul. 20, 2020, at <https://tonedeaf.thebrag.com/facial-recognition-technology-juggalo/>.

Ian O'Neill. (Jul. 2, 2018). *Twitter*. «for anyone wondering why some face changes evade facial recognition and others don't, here's a visualization of how landmarks are placed on a few examples.» Accessed on Sep. 22, 2020, at <https://twitter.com/tahkion/status/1013568373622362113>.

Ian O'Neill. (Jul. 2, 2018). *Twitter*. «for anyone wondering why some face changes evade facial recognition and others don't, here's a visualization of how landmarks are placed on a few examples.» Accessed Sep. 22, 2020, at <https://twitter.com/tahkion/status/1013568373622362113>.

David Cardinal. (Sep. 14, 2017). *ExtremeTech*. «How Apple's iPhone X TrueDepth Camera Works.» Accessed on Sep. 4, 2020, at <https://www.extremetech.com/mobile/255771-apple-iphone-x-truedepth-camera-works>.

theworacle. (Jul. 16, 2009). *Youtube*. «US Air Force Flapping Wing Micro Air Vehicle.» Accessed on Jul. 20, 2020, at https://www.youtube.com/watch?list=RDCMUCZDh1CUNrvzefaLU0n6qZoA&v=_5YkQ9w3PJ4&feature=emb_rel_end.

Cloudera. (n.d.). *Cloudera*. Accessed on Jul. 10, 2020, at <https://www.cloudera.com>.

Tirthajyoti Sarkar. (Oct. 26, 2018). *Towards Data Science*. «What Is Benford's Law and why is it important for data science?» Accessed on Jul. 20, 2020, at <https://towardsdatascience.com/what-is-benfords-law-and-why-is-it-important-for-data-science-312cb8b61048>.

Algorithm&Blues. (Oct. 15, 2010). *Algorithm&Blues*. «Update: Day Traders Crack The Timber Hill Trading System.» Accessed on Jul. 20, 2020, at

<https://algosandblues.wordpress.com/2010/08/17/norwegian-day-traders-crack-the-timber-hill-trading-system/>.

Manceps. (Jul. 14, 2020). *Manceps*. «Could an Adversarial Bot Manipulate the Stock Market?» Accessed on Jul. 20, 2020, at <https://www.manceps.com/articles/experiment/beat-the-bots>.

Simon Weckert. (n.d.). *Simon Weckert*. «Google Maps Hacks.» Accessed on Jul. 20, 2020, at <http://www.simonweckert.com/googlemapshacks.html>.

Moritz Ahlert. (2019). *Hamburger Journal für Kulturanthropologie*. «The Power of Virtual Maps.» Accessed on Sep. 15, 2020, at <https://journals.sub.uni-hamburg.de/hjk/article/view/1395>.

Simon Weckert. (n.d.). *Simon Weckert*. «Google Maps Hacks.» Accessed on Jul. 20, 2020, at <http://www.simonweckert.com/googlemapshacks.html>.

Strategy-Stocks. (n.d.). *Strategy-Stocks*. «What is micro-trading?» Accessed on Aug. 27, 2020, at <https://www.strategystocks.co.uk/micro-trading.html>.

CFM. (Jul. 5, 2018). *CFM*. «Artificial Intelligence: perspectives from the quant coalface.» Accessed on Sep. 2, 2020, at <https://www.cfm.fr/insights/artificial-intelligence-perspectives-from-the-quant-coalface/>.

International Monetary Fund (IMF). (Sep. 2016). *International Monetary Fund (IMF)*. «Detailed Assessment Report on Anti- Money Laundering and Combating the Financing of Terrorism.» Accessed on Jul. 20, 2020, at <https://www.imf.org/external/pubs/ft/scr/2016/cr16294.pdf>.

Europol. (n.d.) *Europol*. «WannaCry Ransomware.» Accessed on Aug. 26, 2020, at <https://www.europol.europa.eu/wannacry-ransomware>.

Europol. (Jun. 28, 2017). *Europol*. «New Wave of Ransomware Affecting Businesses: What to Do?» Accessed on Aug. 26, 2020, at <https://www.europol.europa.eu/newsroom/news/new-wave-of-ransomware-affecting-businesses-what-to-do>.

Erik Zouave, et al. (Mar. 4, 2020). *Swedish Defence Research Agency (FOI)*. «Artificially intelligent cyberattacks.» Accessed on Jul. 20, 2020, at https://www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf.

FileCloud. (Aug. 27, 2018). *FileCloud*. «Machine Vs Machine: A Look at AI-Powered Ransomware.» Accessed on Jul. 20, 2020, at <https://www.getfilecloud.com/blog/2018/08/machine-vs-machine-a-look-at-ai-powered-ransomware/#.XuIHkkVLiUk>.

Harshajit Sarmah. (Aug. 22, 2019). *Analytics India Magazine*. «What If Artificial Intelligence Becomes Ransomware's Sidekick?» Accessed on Jul. 20, 2020, at <https://analyticsindiamag.com/what-if-artificial-intelligence-becomes-ransomwares-sidekick/>.

ESET. (n.d.). *ESET*. «Can Artificial Intelligence Power Future Malware?» Accessed on Jul. 20, 2020, at <https://www.eset.com/me/whitepapers/can-artificial-intelligence-power-future-malware/>.

Scott Ferguson. (Apr. 28, 2020). *Dark Reading*. «Ransomware Attacks Target Public & Government Orgs with More Frequency, Ferocity.» Accessed on Jul. 20, 2020, at <https://www.darkreading.com/application-security/ransomware/ransomware-attacks-target-public-and-government-orgs-with-more-frequency-ferocity/a/d-id/746811>.

Theo Douglas. (Apr. 11, 2018). *Government Technology*. «Phishing, Malware, Ransomware Among Top Public-Sector Threats, Reports Find.» Accessed on Jul. 20, 2020, at <https://www.govtech.com/security/Phishing-Malware-Ransomware-Among-Top-Public-Sector-Threats-Reports-Find.html>.

David Canellos. (Apr. 23, 2020). *Threatpost*. «Public Sector Ransomware Attacks Rage On: Can Your Organization Repel Them?» Accessed on Jul. 20, 2020, at <https://threatpost.com/public-sector-ransomware-attacks-rage/155086/>.

Bill Siegel. (Nov. 1, 2019). *Security Boulevard*. «Ransomware Payments Rise as Public Sector is Targeted, New Variants Enter the Market.» Accessed on Jul. 20, 2020, at <https://securityboulevard.com/2019/11/ransomware-payments-rise-as-public-sector-is-targeted-new-variants-enter-the-market/>.

Trend Micro. (Apr. 26, 2017). *Trend Micro Security Intelligence Blog*. «Machine Learning and the Fight Against Ransomware.» Accessed on Jul. 20, 2020, at <https://blog.trendmicro.com/machine-learning-and-the-fight-against-ransomware/>.

ESET. (n.d.). *ESET*. «Can Artificial Intelligence Power Future Malware?» Accessed on Jul. 20, 2020, at <https://www.eset.com/me/whitepapers/can-artificial-intelligence-power-future-malware/>.

Pindrop. (n.d.). *Pindrop Security, Inc.* «Call Center Fraud Vectors & Fraudsters Analyzed: Summer Break Edition.» Accessed on Aug. 27, 2020, at <https://www.pindrop.com/resources/video/webinar/call-center-fraud-vectors-fraudsters-analyzed-summer-break-edition/>.

Rotem Shemesh. (Nov. 21, 2016). *NICE*. «4 Things You Must Have to Ensure Effective Voice Biometrics in a Contact Center.» Accessed on Aug. 27, 2020, at <https://www.nice.com/engage/blog/4-things-you-must-have-to-ensure-effective-voice-biometrics-in-a-contact-center-2224/>.

Descript. (n.d.). *Descript*. «Lyrebird AI.» Accessed on Aug. 27, 2020, at <https://www.descript.com/lyrebird>.

Nuance. (n.d.). *Nuance Communications, Inc.* «Index.» Accessed on Aug. 27, 2020, at <https://www.nuance.com/index.html>.

AI for Good Summit. (n.d.) *International Telecommunication Union*. «AI for Good Summit.» Accessed on Sep. 24, 2020, at <https://aiforgood.itu.int/>.

Europol. (Oct. 24 – 25, 2019). *Europol*. «3rd ENISA — Europol IoT Security Conference.» Accessed on Sep. 24, 2020, at <https://www.europol.europa.eu/events/3rd-ENISA-europol-iot-security-conference>.

Europol. (Oct. 11, 2019). *Europol*. «Fighting Cybercrime in a Connected Future.» Accessed on Sep. 24, 2020, at <https://www.europol.europa.eu/newsroom/news/fighting-cybercrime-in-connected-future>.

United Nations Interregional Crime and Justice Research Institute (UNICRI). (Jul. 3, 2019). *United Nations Interregional Crime and Justice Research Institute (UNICRI)*. «2nd INTERPOL – UNICRI Global Meeting on Artificial Intelligence for Law Enforcement.» Accessed on Sep. 24, 2020, at http://www.unicri.it/news/article/ai_UNICRI_INTERPOL_law_enforcement.

European Commission. (Jun. 2018). *European Commission*. «Ethics Guidelines for Trustworthy AI.» Accessed on Sep. 24, 2020, at <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top>.

European Telecommunications Standards Institute (ETSI). (n.d.). *European Telecommunications Standards Institute (ETSI)*. «Industry Specification Group (ISG) Securing Artificial Intelligence (SAI).» Accessed on Sep. 24, 2020, at <https://www.etsi.org/committee/1640-sai>.

European Network and Information Security Agency (ENISA). (n.d.). *European Network and Information Security Agency (ENISA)*. «Ad-Hoc Working Group on Artificial Intelligence Cybersecurity.» Accessed on Sep. 24, 2020, at https://www.EMSA.europa.eu/topics/iot-and-smart-infrastructures/artificial_intelligence/adhoc_wg_calls/.

Europol. (Sep. 25, 2013). *Europol*. «Facing Future Cyber Threats.» Accessed on Sep. 24, 2020, at <https://www.europol.europa.eu/newsroom/news/facing-future-cyber-threats>.

Tonya Riley. (Jun. 13, 2019). *The Washington Post*. «The Technology 202: It's time for Congress to address deepfakes, experts say.» Accessed on Jun. 19, 2020, at <https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2019/06/13/the-technology-202-it-s-time-for-congress-to-address-deepfakes-experts-say/5d01687aa7a0a4586bb2da85/>.

Donie O'Sullivan. (Aug. 10, 2019). *CNN Business*. «The Democratic Party deepfaked its own chairman to highlight 2020 concerns.» Accessed on Jun. 19, 2020, at <https://edition.cnn.com/2019/08/09/tech/deepfake-tom-perez-dnc-defcon/index.html>.

Oscar Schwartz. (Nov. 12, 2018). *The Guardian*. «You thought fake news was bad? Deep fakes are where truth goes to die.» Accessed on Jun. 19, 2020, at <https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth>.

Haya R. Hasan and Khaled Salah. (Mar. 18, 2019). *IEEEExplore*. «Combating Deepfake Videos Using Blockchain and Smart Contracts.» Accessed on Jun. 19, 2020, at <https://doi.org/10.1109/ACCESS.2019.2905689>.

Centre for Data Ethics and Innovation. (Sep. 2019). *Centre for Data Ethics and Innovation*. «Deepfakes and Audio-visual Disinformation.» Accessed on Jul. 20, 2020, at [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-_Deepfakes_and_Audiovisual_Disinformation.pdf)

[_Deepfakes_and_Audiovisual_Disinformation.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-_Deepfakes_and_Audiovisual_Disinformation.pdf).

Ian J. Goodfellow, et al. (Jun. 10, 2014). *arXiv*. «Generative Adversarial Networks.» Accessed on Jul. 1, 2020, at <https://arxiv.org/abs/1406.2661>.

Oscar Schwartz. (Nov. 12, 2018). *The Guardian*. «You thought fake news was bad? Deep fakes are where truth goes to die.» Accessed on Jun. 19, 2020, at <https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth>.

Joe Andrews. (Jul. 12, 2019). *CNBC*. «Fake news is real — AI is going to make it much worse.» Accessed on Jun 26, 2020, at <https://www.cnbc.com/2019/07/12/fake-news-is-real-ai-is-going-to-make-it-much-worse.html>.

Than Thi Nguyen, et al. (Jul. 28, 2020). *arXiv*. «Deep Learning for Deepfakes Creation and Detection: A Survey.» Accessed on Jul 3, 2020, at <https://arxiv.org/abs/1909.11573>.

Than Thi Nguyen, et al. (Jul. 28, 2020). *arXiv*. «Deep Learning for Deepfakes Creation and Detection: A Survey.» Accessed on Jul 3, 2020, at <https://arxiv.org/abs/1909.11573>.

Samantha Cole. (Jan. 24, 2020). *Vice*. «New Deepfake Method Can Put Words in Anyone's Mouth.» Accessed on Jun. 27, 2020, at https://www.vice.com/en_us/article/g5xvk7/researchers-created-a-way-to-make-realistic-deepfakes-from-audio-clips.

Henry Ajder. (Sep. 2019). *Deeptrace*. «The State of Deepfakes: Landscape, Threats, and Impact.» Accessed on Jun. 20, 2020, at <https://deeptracelabs.com/archive/>.

Campbell Kwan. (May 23, 2019). *ZDNet*. «Samsung uses AI to transform photos into talking head videos.» Accessed on Jun. 23, 2020, at <https://www.zdnet.com/article/samsung-uses-ai-to-transform-photos-into-talking-head-videos/>.

Adi Robertson. (2018, Feb 11). *The Verge*. «I'm using AI to face-Swap Elon Musk and Jeff Bezos, and I'm really bad at it.» Accessed on Jun. 29, 2020, at <https://www.theverge.com/2018/2/11/16992986/fakeapp-deepfakes-ai-face-swapping>.

Ashley Carman. (Jul. 17, 2019). *The Verge*. «FaceApp is back and so are privacy concerns.» Accessed on Jun 20, 2020, at <https://www.theverge.com/2019/7/17/20697771/faceapp-privacy-concerns-ios-android-old-age-filter-russia>.

Agence France-Presse in Shanghai. (Sep. 2, 2019). *The Guardian*. «Chinese deepfake app Zao sparks privacy row after going viral.» Accessed on Jun. 24, 2020, at <https://www.theguardian.com/technology/2019/sep/02/chinese-face-swap-app-zao-triggers-privacy-fears-viral>.

Descript. (n.d.). *Descript*. «Overdub.» Accessed on Jul. 20, 2020, at <https://www.descript.com/overdub?lyrebird=true>.

Haya R. Hasan and Khaled Salah. (Mar. 18, 2019). *IEEEExplore*. «Combating Deepfake Videos Using Blockchain and Smart Contracts.» Accessed on Jun. 19, 2020, at <https://doi.org/10.1109/ACCESS.2019.2905689>.

Aja Romano (Apr. 18, 2018). *Vox*. «Jordan Peele's simulated Obama PSA is a double-edged warning against fake news.» Accessed on Jun. 23, 2020, at <https://www.vox.com/2018/4/18/17252410/jordan-peelee-obama-deepfake-buzzfeed>.

Samantha Cole. (Jun. 12, 2019). *Vice*. «This Deepfake of Mark Zuckerberg Tests Facebook's Fake Video Policies.» Accessed on Jul. 3, 2020, at https://www.vice.com/en_us/article/ywyxex/deepfake-of-mark-zuckerberg-facebook-fake-video-policy.

Nic Ker. (Jun. 12, 2019). *Malay Mail*. «Is the political vide Viral sex video confession real or a Deepfake?» Accessed on Jul. 1, 2020, at <https://www.malaymail.com/news/malaysia/2019/06/12/is-the-political-aide-viral-sex-video-confession-real-or-a-deepfake/1761422>.

Sarah Cahlan. (Feb. 13, 2020). *The Washington Post*. «How misinformation helped spark an attempted coup in Gabon.» Accessed on Jul. 1, 2020, at

<https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/>.

Angelin Yeoh. (Dec. 12, 2019). *AsiaOne*. «Malaysian actor Zul Ariffin makes 'deepfake' claims after sex video goes viral on social media.» Accessed on Jul. 1, 2020, at <https://www.asiaone.com/digital/malaysian-actor-zul-ariffin-makes-deepfake-claims-after-sex-video-goes-viral-social-media>.

Oscar Schwartz. (Nov. 12, 2018). *The Guardian*. «You thought fake news was bad? Deep fakes are where truth goes to die.» Accessed on Jun. 20, 2020, at <https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth>.

Beatrice Dupuy (May 24, 2019). *AP News*. «NOT REAL NEWS: Altered video makes Pelosi seem to slur words.» Accessed on Jun. 22, 2020, at <https://apnews.com/4841d0ebcc704524a38b1c8e213764d0>.

Raphael Satter. (Jun. 13, 2019). *ABC News Network*. «Experts: Spy used AI-generated face to connect with targets.» Accessed on Jun. 26, 2020, at <https://abcnews.go.com/Technology/wireStory/experts-spy-ai-generated-face-connect-targets-63674174>.

Jesse Damiani (Nov. 9, 2019). *Forbes*. «A Voice Deepfake Was Used To Scam A CEO Out Of \$243,000.» Accessed on Jun. 25, 2020, at <https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/#5b10b1802241>.

Kaveh Waddell and Jennifer A. Kingson. (Jul. 19, 2019). *Axios*. «The coming of deepfakes threat to businesses.» Accessed on Jul. 1, 2020, at [https://www.axios.com/the-coming-deepfakes-threat-to-businesses-308432e8-f1d8-465e-b628-](https://www.axios.com/the-coming-deepfakes-threat-to-businesses-308432e8-f1d8-465e-b628-07498a7c1e2a.html?utm_source=twitter&utm_medium=social&utm_campaign=organic)

[07498a7c1e2a.html?utm_source=twitter&utm_medium=social&utm_campaign=organic](https://www.axios.com/the-coming-deepfakes-threat-to-businesses-308432e8-f1d8-465e-b628-07498a7c1e2a.html?utm_source=twitter&utm_medium=social&utm_campaign=organic).

Lorenzo Franceschi-Bicchierai. (Jul. 24, 2020). *Vice*. «Listen to This Deepfake Audio Impersonating a CEO in Brazen Fraud Attempt.» Accessed on Jul. 3, 2020, at https://www.vice.com/en_us/article/pkyqvb/deepfake-audio-impersonating-ceo-fraud-attempt.

Maeve Sheehan. (Feb. 16, 2020). *Independent.ie*. «Cyber scam stole voice of Europol's ex-boss to get cash.» Accessed on Aug. 26, 2020, at <https://www.independent.ie/irish-news/cyber-scam-stole-voice-of-europols-ex-boss-to-get-cash-38960406.html>.

Gabriela Swerling. (Jan. 31, 2020). *The Telegraph*. «Doctored audio evidence used to damn father in custody battle.» Accessed on Jul. 3, 2020, at <https://www.telegraph.co.uk/news/2020/01/31/deepfake-audio-used-custody-battle-lawyer-reveals-doctored-evidence/>.

Naser Damer, PhD, et al. (Oct. 2018). *IEEEExplore*. «MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network.» Accessed on Jul. 20, 2020, at <https://doi.org/10.1109/BTAS.2018.8698563>.

Arjo Solutions. «Digital technologies: the solution for breeder documents protection?» (Sep. 12, 2017). Arjo Solutions. Accessed on Aug. 26, 2020, at <https://www.arjo-solutions.com/en/2017/09/12/breeder-documents-protection/>.

University of Lincoln. (Aug. 1, 2019). *ScienceDaily*. «Two fraudsters, one passport: Computers more accurate than humans at detecting fraudulent identity photos.» Accessed on Jul. 20, 2020, at www.sciencedaily.com/releases/2019/08/190801104038.htm.

Naser Damer, PhD. (n.d.). *FraunhoferIGD*. «Face morphing: a new threat?» Accessed on Jul. 20, 2020, at <https://www.igd.fraunhofer.de/en/press/annual-reports/2018/face-morphing-a-new-threat>.

David J. Robertson, et al. (Jun. 27, 2018). *Springer Nature*. «Detecting morphed passport photos: a training and individual differences approach.» Accessed on Jul. 20, 2020, at <https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-018-0113-8>.

Robin S.S. Kramer, et al. (Jul. 29, 2019). *Springer Nature*. «Face morphing attacks: Investigating detection with humans and computers.» Accessed on Jul. 20, 2020, at https://link.springer.com/article/10.1186/s41235-019-0181-4?error=cookies_not_supported&code=bffbdaf-92aa-47c2-b4c1-37a7a51dd341.

David J. Robertson, et al. (Jun. 27, 2018). *Springer Nature*. «Detecting morphed passport photos: a training and individual differences approach.» Accessed on Jul. 20, 2020, at <https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-018-0113-8>.

Matteo Ferrara, Annalisa Franco, and Davide Maltoni. (Dec. 2014). *ResearchGate*. «The magic passport.» Accessed on Jul. 20, 2020, at https://www.researchgate.net/publication/283473746_The_magic_passport.

Giorgio Patrini. (Oct. 7, 2019) *DeepTrace*. «Mapping the Deepfake Landscape.» Accessed on Aug. 26, 2020, at [https:// deeptracelabs.com/mapping-the-deepfake-landscape/](https://deeptracelabs.com/mapping-the-deepfake-landscape/).

Europol, Trend Micro, and UNICRI. (Mar. 3, 2020). Criminal Misuse of AI and Deepfakes, The Hague, The Netherlands.

Europol European Cybercrime Centre (EC3). (2019). *2019 Predictions — The Year of Artificial Intelligence*. Europol.

Europol. (Oct. 9, 2019). *Europol*. «Internet Organised Crime Threat Assessment (IOCTA).» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2019>.

Europol. (Jul. 18, 2019). *Europol*. «Do Criminals Dream of Electric Sheep?» Accessed on Oct. 14, 2020, at [https://www. europol.europa.eu/newsroom/news/do-criminals-dream-of-electric-sheep-how-technology-shapes-future-of-crime-and-law-enforcement](https://www.europol.europa.eu/newsroom/news/do-criminals-dream-of-electric-sheep-how-technology-shapes-future-of-crime-and-law-enforcement).

United Nations Interregional Crime and Justice Research Institute (UNICRI). (Apr. 2, 2019). *United Nations Interregional Crime and Justice Research Institute (UNICRI)*. «High-Level Event: Artificial Intelligence and Robotics — Reshaping the Future of Crime, Terrorism and Security.» Accessed on Aug. 26, 2020, at http://www.UNICRI.it/news/article/AI_Robotics_Crime_Terrorism_Security.

Hany Farid. (Jun. 18, 2019). *Fox News*. «Hany Farid: Deepfakes give new meaning to the concept of 'fake news,' and they're here to stay.» Accessed on Jun. 20, 2020, at <https://www.foxnews.com/opinion/hany-farid-deep-fakes>.

Tom Taulli. (Jun. 15, 2019). *Forbes*. «Deepfake: What You Need To Know.» Accessed on Jun. 20, 2020, at <https://www.forbes.com/sites/tomtaulli/2019/06/15/deepfake-what-you-need-to-know/#52981563704d>.

Drew Harwell. (Jun. 12, 2019). *The Washington Post*. «Top AI researchers race to detect 'deepfake' videos: 'We are outgunned.'» Accessed on Jun. 20, 2020, at <https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/>.

Mayra Rosario Fuentes. (May 26, 2020). *Trend Micro*. «Shifts in Underground Markets: Past, Present, and Future.» Accessed on Jul. 20, 2020, at <https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/trading-in-the-dark>.

Europol. (Jun. 14, 2018). *Europol*. «French Coder Who Helped Extort British Company Arrested in Thailand.» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/newsroom/news/french-coder-who-helped-extort-british-company-arrested-in-thailand>.

Hany Farid. (Jun. 18, 2019). *Fox News*. «Hany Farid: Deepfakes give new meaning to the concept of 'fake news,' and they're here to stay.» Accessed on Jun. 20, 2020, at <https://www.foxnews.com/opinion/hany-farid-deep-fakes>.

Europol. (Jun. 19, 2017). *Europol*. «Online Sexual Coercion and Extortion as a Form of Crime Affecting Children: Law Enforcement Perspective.» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/publications-documents/online-sexual-coercion-and-extortion-form-of-crime-affecting-children-law-enforcement-perspective>.

Europol. (Oct. 9, 2019). *Europol*. «Internet Organised Crime Threat Assessment (IOCTA).» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2019>.

Europol. (Oct. 9, 2019). *Europol*. «Internet Organised Crime Threat Assessment (IOCTA).» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2019>.

Laura Hautala. (Feb. 11, 2018). *InformationSecurity.Report*. «If you like deepfakes, you might be mining cryptocurrency.» Accessed on Jun. 23, 2020, at <https://informationsecurity.report/news/if-you-like-deepfakes-you-might-be-mining-cryptocurrency/4825>.

Mayra Rosario Fuentes. (May 26, 2020). *Trend Micro*. «Shifts in Underground Markets: Past, Present, and Future.» Accessed on Jul. 20, 2020, at <https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/trading-in-the-dark>.

Europol. (Oct. 9, 2019). *Europol*. «Internet Organised Crime Threat Assessment (IOCTA).» Accessed on Jul. 20, 2020, at <https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2019>.

Timothy Levine. (Aug. 11, 2014). *SAGE Publications*. «Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection.» Accessed on Aug. 26, 2020, at https://www.researchgate.net/publication/273593306_Truth-Default_Theory_TDT_A_Theory_of_Human_Deception_and_Deception_Detection.

Andreas Rossler, et al. (Aug. 26, 2019). *arXiv*. «FaceForensics++, Learning to Detect Manipulated Face Images.» Accessed on Aug. 26, 2020, at <https://arxiv.org/abs/1901.08971v3>.

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. (Jun. 11, 2018). *arXiv*. «In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking.» Accessed on Aug. 26, 2020, at <https://arxiv.org/abs/1806.02877>.

Kara Manke. (Jun. 18, 2019). *Berkeley News*. «Researchers use facial quirks to unmask 'deepfakes'.» Accessed on Aug. 26, 2020, at <https://news.berkeley.edu/2019/06/18/researchers-use-facial-quirks-to-unmask-deepfakes/>.

Jim Brenner. (Nov. 25, 2019). *iProov*. «The Deepfake Age: Who Gatekeeps Digital Trust?» Accessed on Aug. 26, 2020, [https:// www.iproov.com/newsroom/blog/the-deepfake-age-who-gatekeeps-digital-trust](https://www.iproov.com/newsroom/blog/the-deepfake-age-who-gatekeeps-digital-trust).

Antonio Garda Martfnez. (Mar. 26, 2018). *Wired*. «The Blockchain Solution to Our Deepfake Problems.» Accessed on Jul. 20, 2020, at <https://www.wired.com/story/the-blockchain-solution-to-our-deepfake-problems/>.

Lily Hay Newman. (Feb. 11, 2019). *Wired*. «A New Tool Protects Videos From Deepfakes and Tampering.» Accessed on Jul. 20, 2020, at <https://www.wired.com/story/amber-authenticate-video-validation-blockchain-tampering-deepfakes/>.

eyeWitness. (n.d.). *eyeWitness*. «What we do.» Accessed on Jul. 20, 2020, at <https://www.eyewitness.global/our-work.html>.

Haya R. Hasan and Khaled Salah. (Mar. 18, 2019). *IEEEExplore*. «Combating Deepfake Videos Using Blockchain and Smart Contracts.» Accessed on Jun. 19, 2020, at <https://doi.org/10.1109/ACCESS.2019.2905689>.

Centre for Data Ethics and Innovation. (Sep. 2019). *Centre for Data Ethics and Innovation*. «Deepfakes and Audio-visual Disinformation.» Accessed on Jul. 20, 2020, at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-_Deepfakes_and_Audiovisual_Disinformation.pdf.

Kim Lyons. (Jan. 29, 2020). *The Verge*. «FTC says the tech behind audio deepfakes is getting better.» Accessed on Aug. 26, 2020, at <https://www.theverge.com/2020/1/29/21080553/ftc-deepfakes-audio-cloning-joe-rogan-phone-scams>.

Dessa News. (Sep. 28, 2019). *Medium*. «Detecting Audio Deepfakes With AI.» Accessed on Aug. 26, 2020, at <https://medium.com/dessa-news/detecting-audio-deepfakes-f2edfd8e2b35>.

Facebook AI. (Jun. 25, 2020). *Facebook AI*. «Deepfake Detection Challenge Dataset.» Accessed on Aug. 26, 2020, at <https://deepfakedetectionchallenge.ai>.

Ivan Mehta. (Sep. 2, 2020). *The Next Web*. «Microsoft's new deepfake detection tool rates bogus videos with a confidence score.» Accessed on Sep. 8, 2020, at <https://thenextweb.com/neural/2020/09/02/microsofts-new-deepfake-detection-tool-rates-bogus-videos-with-a-confidence-score/>.

Will Douglas Heaven. (Jun. 12, 2020). *MIT Technology Review*. «Facebook just released a database of 100,000 deepfakes to teach AI how to spot them.» Accessed on Aug. 26, 2020, at <https://www.technologyreview.com/2020/06/12/1003475/facebooks-deepfake-detection-challenge-neural-network-ai/>.

Leo Kelion. (Feb. 7, 2018). *BBC News*. «Reddit bans deepfake porn videos.» Accessed on Jun. 27, 2020, at <https://www.bbc.com/news/technology-42984127>.

Jay Peters. (Jan. 9, 2020). *The Verge*. «Reddit bans impersonation on its platform.» Accessed on Jun. 27, 2020, at <https://www.theverge.com/2020/1/9/21058803/reddit-account-ban-impersonation-policy-deepfakes-satire-rules>.

Tony Romm, Drew Harwell, and Isaac Stanley-Becker. (Jan. 8, 2020). *The Washington Post*. «Facebook bans deepfakes, but new policy may not cover controversial Pelosi video.» Accessed on Jun. 27, 2020, at <https://www.washingtonpost.com/technology/2020/01/06/facebook-ban-deepfakes-sources-say-new-policy-may-not-cover-controversial-pelosi-video/>.

Margi Murphy. (Feb. 4, 2020). *The Telegraph*. «Twitter bans 'deepfakes' and 'cheap fakes.'» Accessed on Jun. 27, 2020, at <https://www.telegraph.co.uk/technology/2020/02/04/twitter-bans-deepfakes-cheap-fakes/>.

Pablo Tseng. (Mar. 2018). *McMillan*. «What Can The Law Do About 'Deepfake'?» Accessed on Jun. 27, 2020, at <https://mcmillan.ca/What-Can-The-Law-Do-About-Deepfake>.

Nick Statt. (Nov. 29, 2019). *The Verge*. «China makes it a criminal offense to publish deepfakes or fake news without disclosure.» Accessed on Jun. 27, 2020, at <https://www.theverge.com/2019/11/29/20988363/china-deepfakes-ban-internet-rules-fake-news-disclosure-virtual-reality>.

Ben Sasse. (Dec. 21, 2018). *Library of Congress*. «S.3805 – Malicious Deep Fake Prohibition Act of 2018.» Accessed on Jun. 27, 2020, at [https://www.congress.gov/bill/115th-congress/senate-bill/3805#:~:text=%2F21%2F2018\)-,Malicious%20](https://www.congress.gov/bill/115th-congress/senate-bill/3805#:~:text=%2F21%2F2018)-,Malicious%20)

Deep%20Fake%20Prohibition%20Act%20of%202018,media%20records%20that%20appear%20realistic.

Yvette D. Clarke. (Jun. 12, 2019). *Library of Congress*. «H.R.3230 – Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019.» Accessed on Jun. 27, 2020, at <https://www.congress.gov/bill/116th-congress/house-bill/3230>.

Centre for Data Ethics and Innovation. (Sep. 2019). *Centre for Data Ethics and Innovation*. «Deepfakes and Audio-visual Disinformation.» Accessed on Jul. 20, 2020, at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-_Deepfakes_and_Audiovisual_Disinformation.pdf.

Centre for Data Ethics and Innovation. (Sep. 2019). *Centre for Data Ethics and Innovation*. «Deepfakes and Audio-visual Disinformation.» Accessed on Jul. 20, 2020, at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-_Deepfakes_and_Audiovisual_Disinformation.pdf.

European Commission. (Feb. 4, 2020). *European Commission*. «Illegal content on online platforms.» Accessed on Jul. 20, 2020, at <https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms>.

Europol European Cybercrime Centre (EC3). (2019). *DeepFake*. Europol Platform for Experts – EPE.

Europol. (Jul. 18, 2019). *Europol*. «Do Criminals Dream of Electric Sheep?» Accessed on Oct. 14, 2020, at <https://www.europol.europa.eu/newsroom/news/do-criminals-dream-of-electric-sheep-how-technology-shapes-future-of-crime-and-law-enforcement>.

Europol European Cybercrime Centre (EC3). (2019). *DeepFake*. Europol Platform for Experts – EPE.

Europol. (Oct. 5, 2020). *Europol*. «Internet Organised Crime Threat Assessment (IOCTA) 2020.» Accessed on Oct. 14, 2020, at

<https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2020>.

Trend Micro, мировой лидер в области кибербезопасности, помогает сделать мир безопасным для обмена цифровой информацией.

В Trend Micro Research работают эксперты, которые стремятся обнаружить новые угрозы, поделиться ключевыми знаниями и поддержать усилия по борьбе с киберпреступниками. Наша международная команда ежедневно помогает выявлять миллионы угроз, лидирует в отрасли по раскрытию уязвимостей и публикует инновационные исследования новых методов борьбы с угрозами. Мы постоянно работаем над тем, чтобы предвидеть новые угрозы и предоставлять аналитические материалы.



Искусственный интеллект-2021

Письмо Председателя и заместителя Председателя

США еще не осознали, насколько глубоко революция искусственного интеллекта (ИИ) повлияет на экономику, национальную безопасность и благосостояние, вместе с тем предстоит узнать о мощи и ограничениях технологий искусственного интеллекта.

Тем не менее, сейчас необходимо принять важные решения для ускорения инноваций в области ИИ на благо США, а также защиты от пагубного использования ИИ.

Дилеммой государственного управления, обозначенной ранее Генри Киссинджером: «Когда ваш простор для действий наибольший, знания, на которых вы можете основывать это действие, всегда минимальны. Когда ваши знания огромны, простор для действий часто исчезает.»

Простор для действий остается, но пространство для маневра Америки сокращается.

Так, двухпартийная комиссия состоящая из 15 технологов, специалистов по национальной безопасности, руководителей бизнеса и академического истеблишмента, представителей Комиссии национальной безопасности по искусственному интеллекту (NSCAI) сообщает о том, что Америка ещё не готова защищаться или конкурировать в эпоху искусственного интеллекта.

Это суровая реальность, с которой она должна столкнуться, и именно эта реальность требует всеобъемлющих, общегосударственных действий.

В заключительном докладе представлена стратегия защиты от угроз ИИ, а также использования ИИ в интересах национальной безопасности и возможной победы в более широком технологическом соревновании во имя процветания США.

Правительство США не может реализовать эту задачу в одиночку. Ему нужны преданные партнеры в промышленности, научных кругах и гражданском обществе. И Америка также должна заручиться поддержкой своих старейших союзников и новых партнеров, чтобы построить более безопасный и свободный мир в эру искусственного интеллекта.

ИИ – это вдохновляющая технология, которая станет самым мощным инструментом в будущем, направленная для блага человечества.

Учёные уже достигли поразительного прогресса в ходе использования искусственного интеллекта в самых разных областях – от биологии и медицины до астрофизики.

Эти достижения не являются частными научными экспериментами, они улучшают жизнь и открывают тайны природного мира и это такие открытия, для которых ярлык «изменение игры» не является клише.

Считается, что системы искусственного интеллекта также будут использоваться как в погоне за властью, так и как инструменты оружия первой инстанции в будущих конфликтах.

Но сам ИИ не останется только в сфере сверхдержав, или в сфере научной фантастики, он имеет двойное назначение, часто с открытым исходным кодом и быстро распространяется.

Противники государства уже используют дезинформационные атаки с использованием ИИ, чтобы посеять раскол в демократических институтах и поколебать у народа чувства реальности.

По нашему мнению, государства-преступники и террористы будут проводить кибератаки на основе искусственного интеллекта и соединять программное обеспечение искусственного интеллекта с коммерчески доступными беспилотными летательными аппаратами для создания «умного оружия».

Ни для кого не секрет, что и американские военные благодаря десятилетнему технологическому преимуществу Соединённых Штатов

интегрируют концепции ИИ в военных технологиях и военно-технологических платформах, чтобы противостоять современным вызовам.

Несмотря на перспективные эксперименты в разработке программ искусственного интеллекта, правительство США ещё далеко не готово к широкому применению «искусственного интеллекта» во всех государственных сферах. Бизнес-лидеры Комиссии больше всего разочарованы медлительностью с которым подходит правительство к прогрессу в применении ИИ в крупных учреждениях.

Интеграция ИИ сложна в любом секторе и в сфере национальной безопасности возникают некоторые проблемы. Тем не менее, лидеры государства, преданные своему народу и политическому курсу США, могут способствовать переменам.

Такие лидеры нужны в Пентагоне и во всем федеральном правительстве, чтобы выстроить техническую инфраструктуру и связать идеи и эксперименты с новыми политическими и военными концепциями.

К 2025 году Министерство обороны США и Разведывательное сообщество должны быть готовы к использованию ИИ и должны конкурировать в ИИ.

Тот факт, что ИИ проходит через множество смежных технологий и используется во многих областях, объясняет его мощь и неумолимо приводит к ещё одной критической точке: ИИ является частью более широкой глобальной технологической конкуренции.

Конкуренция должна ускорить инновации. Мы должны соревноваться вместе с партнёрами, когда конкуренция ИИ направлена на научные открытия, которые приносят пользу человечеству, например, открытие вакцин.

Но мы должны выиграть конкуренцию ИИ, которая усиливает стратегическую конкуренцию с Китаем. Планы, ресурсы и прогресс Китая должны волновать всех американцев. Он является сверстником ИИ во многих областях и лидером ИИ в некоторых приложениях. Мы серьёзно относимся к

стремлению Китая превзойти Соединённые Штаты в качестве мирового лидера ИИ в течение десятилетия.

Соревнование ИИ – это также соревнование ценностей. Внутреннее использование ИИ Китаем является пугающим прецедентом для любого человека во всём мире, который дорожит личной свободой.

Использование ИИ в качестве инструмента репрессий и слежки – как внутри страны, так и все чаще за рубежом – является мощным контрапунктом тому, как, по нашему мнению, следует использовать ИИ.

Будущее ИИ может быть демократическим, но мы достаточно узнали о силе технологий, чтобы укрепить авторитаризм за рубежом и подпитывать экстремизм внутри страны, чтобы знать, что мы не должны считать само собой разумеющимся, что будущие технологические тенденции будут укреплять, а не разрушать демократию.

Мы должны работать с другими демократиями и частным сектором над созданием стандартов защиты частной жизни для технологий ИИ и продвигать демократические нормы использования ИИ, чтобы демократии могли ответственно использовать инструменты ИИ в целях национальной безопасности.

Мы хотели бы выделить несколько областей, в которых необходимы срочные действия, поскольку ставки в этом соревновании очень высоки:

В конечном счёте, мы обязаны убедить лидеров в правительстве США принять трудное решение, чтобы выиграть эру искусственного интеллекта.

В Америке доллар останавливается на президенте, а стратегия ИИ начинается в Белом доме. Мы создали Совет национальной безопасности, чтобы противостоять вызовам послевоенной эпохи. Теперь нам необходимо создать Совет по конкурентоспособности технологий для разработки стратегии, учитывающей сложные проблемы безопасности, экономики и науки, связанные с ИИ и связанными с ним технологиями.

Этот императив лидерства должен распространяться на все важнейшие департаменты и агентства национальной безопасности.

Дефицит человеческих талантов является самым заметным дефицитом ИИ в правительстве и самым большим тормозом для покупки, создания и внедрения технологий с поддержкой ИИ в целях национальной безопасности.

Сейчас не время добавлять несколько новых должностей в департаментах и агентствах национальной безопасности для технологов Кремниевой долины. Нам нужно с нуля построить совершенно новые трубопроводы талантов. Мы должны создать новую Академию цифровой службы и гражданский национальный резерв, чтобы выращивать технические таланты с той же серьёзностью цели, с какой мы выращиваем военных офицеров.

Цифровой век требует цифрового корпуса. Не менее важно нам осознавать, что Соединённые Штаты должны выиграть международный конкурс талантов, улучшив как STEM-образование, так и нашу систему приёма и удержания высококвалифицированных иммигрантов.

Микроэлектроника питает весь ИИ, и Соединённые Штаты больше не производят самые сложные в мире чипы. Мы не хотим преувеличивать шаткость нашего положения, но, учитывая, что подавляющее большинство современных чипов производится на одном заводе, отделённом всего лишь 110 милями воды от нашего главного стратегического конкурента, мы должны пересмотреть значение устойчивости и безопасности цепочки поставок. Недавняя нехватка чипов для производства автомобилей дорого обошлась американской автомобильной компании, эта ситуация вошла в Заключительный отчёт и представляет собой важный шаг, но он не является заключительным актом NSCAI.

В течение оставшегося срока работы Комиссии наша работа будет сосредоточена на том, чтобы помочь Президенту и Конгрессу сделать необходимые инвестиции и принять меры, рекомендованные для победы в эру искусственного интеллекта. (5 млрд долларов).

Стратегическая блокировка обойдётся гораздо дороже и поставит под угрозу нашу безопасность. Федеральные инвестиции и стимулы, необходимые для оживления отечественного производства микрочипов (возможно 35 млрд

долларов) должны быть простым решением, когда альтернативой является использование другой страны для производства двигателей, которые будут питать машины, которые будут формировать будущее.

Мы обеспокоены тем, что только несколько крупных компаний и могущественных государств будут иметь ресурсы, чтобы сделать самые большие прорывы в области искусственного интеллекта.

Несмотря на распространение инструментов с открытым исходным кодом, потребности в вычислительных мощностях и массивах данных для улучшения алгоритмов стремительно растут на переднем крае инноваций. Федеральное правительство должно сотрудничать с американскими компаниями, чтобы сохранить американское лидерство и поддержать разработку разнообразных приложений ИИ, которые продвигают национальные интересы в самом широком смысле. Во всяком случае, этот отчёт недооценивает инвестиции, которые Америка должна будет сделать. (40 млрд долларов), которые мы рекомендуем потратить на расширение и демократизацию федеральных исследований и разработок в области ИИ, – это скромный аванс за будущие прорывы.

Нам также нужно будет создать безопасную цифровую инфраструктуру по всей стране, общий доступ к облачным вычислениям и умные города, чтобы действительно использовать ИИ на благо всех американцев.

Мы предполагаем, что в ближайшие годы федеральные расходы составят сотни миллиардов.

Сейчас не время для абстрактной критики промышленной политики или опасений, что дефицитные расходы встанут на пути прогресса.

В 1956 году президент Дуайт Эйзенхауэр, консервативный в финансовом отношении республиканец, работал с Демократическим Конгрессом, чтобы выделить 10 млрд долларов на строительство системы межштатных автомагистралей. Это 96 млрд долларов в современном мире. Конечно, мы можем сделать аналогичные инвестиции в будущее страны.

Мы гордимся двухпартийной работой NSCAI, вместе обсуждали, вместе учились и достигли консенсуса по критическим вопросам. Мы имеем честь представить свои рекомендации Конгрессу и Президенту.

Перефразируя Уинстона Черчилля, мы находимся в начале той конкуренции, которая будет определять наше процветание, национальную безопасность и благосостояние наших граждан.

В нашем докладе представлены первые шаги, которые Соединённые Штаты должны предпринять для защиты, конкуренции и победы в эпоху искусственного интеллекта.

Письмо исполнительного директора:

Начало Начала

Когда мы начали наше путешествие два года назад, мы почти не знали, что нас ждёт. Мы столкнулись с желанием и надеждой многих друзей и союзников получить от Конгресса право на сохранение преимущества Соединённых Штатов в области искусственного интеллекта (ИИ).

Мы пользовались поддержкой американских министерств и ведомств. Многие из них одалживали нам ресурсы, включая детализацию как гражданского, так и военного персонала, и посвящали бесчисленные часы, чтобы помочь нам понять их миссии и приоритеты.

Члены Конгресса тесно сотрудничали с нами, чтобы ускорить принятие нашим правительством искусственного интеллекта в целях национальной безопасности.

В ходе работы Комиссии мы взаимодействовали с сотнями представителей частного сектора, научных кругов, гражданского общества и правительства, получали информацию в ходе проведения секретных и несекретных брифингов. Мы встречались со всеми, кто думает об ИИ, работает с ИИ и развивает ИИ, кто готов уделить нам время.

Почти все наши партнёры пришли к единому мнению по трём пунктам: убеждённость в том, что ИИ-это чрезвычайно мощная технология, признание настоящей необходимости инвестировать больше средств в инновации ИИ и ответственность за разработку и использование ИИ, руководствуясь демократическими принципами. Мы также поговорили с нашими союзниками старыми и новыми от Нью-Дели до Тель-Авива и Лондона, они выразили готовность и желание работать с Соединёнными Штатами для углубления сотрудничества в ИИ.

Я в долгу перед многими людьми, которые добровольно работали с нами, интернировались вместе с нами, предоставляли экспертные знания и были друзьями Комиссии. Я особенно благодарен самоотверженным

штатным сотрудникам Комиссии, которые во многих случаях уходили с важных должностей, чтобы присоединиться к этой важной миссии.

В последние два года мы столкнулись с широко распространённой надеждой на то, что ИИ может принести невероятные выгоды для экономики, благосостояния и безопасности нашей страны.

Мы также слышали опасения, что ИИ как и любая технология может создать новые проблемы и усугубить существующие. Мы выслушали и отнеслись к этим опасениям серьёзно.

В конечном итоге мы пришли к признанию того, что если Америка примет и инвестирует в ИИ, основанный на наших ценностях, это изменит нашу страну и обеспечит, чтобы Соединённые Штаты и их союзники продолжали формировать мир на благо всего человечества.

Спасибо! Илл Байрактари

Основные положения

Ни одна удобная историческая справка не отражает влияния искусственного интеллекта (ИИ) на национальную безопасность.

ИИ-это не единичный технологический прорыв, как бомбардировщик-невидимка с крылом летучей мыши. Гонка за превосходством ИИ не похожа на космическую гонку на Луну. ИИ даже не сравним с технологией общего назначения, такой как электричество. Однако то, что Томас Эдисон сказал об электричестве, включает в себе будущее ИИ: «Это поле полей ... оно хранит секреты, которые реорганизуют жизнь мира.» Поразительная оценка Эдисона исходила из смирения. Все, что он открыл, было «очень мало по сравнению с теми возможностями, которые появляются.»

Комиссия национальной безопасности по искусственному интеллекту (SCAI) смиренно признает, как много ещё предстоит узнать об ИИ и его будущих приложениях. Тем не менее, сегодня мы знаем об ИИ достаточно, чтобы начать с двух убеждений.

Во-первых, быстро растущая способность компьютерных систем решать проблемы и выполнять задачи, которые в противном случае потребовали бы человеческого интеллекта, а в некоторых случаях и превосходили человеческие возможности, меняет мир.

Технологии искусственного интеллекта являются самыми мощными инструментами в поколениях для расширения знаний, повышения благосостояния и обогащения человеческого опыта. ИИ также является квинтэссенцией технологии «двойного назначения».

Способность машины воспринимать, оценивать и действовать быстрее и точнее человека представляет собой конкурентное преимущество в любой области гражданской или военной. Технологии ИИ станут источником огромной силы для компаний и стран, которые их используют.

Во-вторых, ИИ расширяет окно уязвимости, в которое уже вступили Соединённые Штаты. Впервые после Второй мировой войны технологическое превосходство Америки, основа её экономической и военной мощи оказалось под угрозой.

Китай обладает мощью, талантом и амбициями превзойти Соединённые Штаты в качестве мирового лидера в области ИИ в следующем десятилетии, если нынешние тенденции не изменятся.

Одновременно ИИ усиливает угрозу кибератак и дезинформационных кампаний, которые Россия, Китай и другие используют для проникновения в наше общество, кражи наших данных и вмешательства в нашу демократию.

Ограниченное использование атак с использованием искусственного интеллекта на сегодняшний день представляет собой вершину айсберга.

Между тем глобальные кризисы, примером которых являются пандемия COVID-19 и изменение климата, подчёркивают необходимость расширения нашей концепции национальной безопасности и поиска инновационных решений с использованием искусственного интеллекта.

«Заключительный отчёт NSCAI представляет собой комплексную национальную стратегию по реорганизации правительства, переориентации

нации и объединению наших ближайших союзников и партнёров для защиты и конкуренции в грядущую эпоху ускоренной конкуренции и конфликтов.»

Учитывая эти убеждения, Комиссия приходит к выводу, что Соединённые Штаты должны действовать сейчас, чтобы внедрить системы ИИ и инвестировать значительно больше ресурсов в инновации ИИ, чтобы защитить свою безопасность, способствовать своему процветанию и защитить будущее демократии.

Сегодня правительство не организует и не инвестирует средства для победы в технологическом соревновании против преданного конкурента, а также не готово защищаться от угроз с использованием искусственного интеллекта и быстро внедрять приложения искусственного интеллекта в целях национальной безопасности.

Сейчас не время для постепенного увеличения федерального бюджета на исследования или добавления нескольких новых должностей в Пентагоне для технологов Кремниевой долины.

Это будет дорого стоить и потребует значительного изменения менталитета. Америка нуждается в руководстве Белого дома, действиях членов Кабинета и двухпартийной поддержке Конгресса, чтобы выиграть эру искусственного интеллекта.

В Заключительном докладе NSCAI представлена комплексная национальная стратегия реорганизации правительства, переориентации нации и сплочения наших ближайших союзников и партнёров для защиты и конкуренции в грядущую эпоху ускоренной конкуренции и конфликтов с использованием искусственного интеллекта.

Это двусторонний подход.

Часть I, «Защита Америки в эпоху ИИ», описывает ставки, объясняет, что Соединённые Штаты должны делать для защиты от спектра угроз, связанных с ИИ, и рекомендует, как правительство США может ответственно использовать технологии ИИ для защиты американского народа и наших интересов.

Часть II, «Победа в технологическом соревновании», посвящена критическим элементам конкуренции ИИ и рекомендует меры, которые правительство должно предпринять для продвижения инноваций ИИ для повышения национальной конкурентоспособности и защиты важнейших преимуществ США. Рекомендации разработаны как взаимосвязанные и взаимоукрепляющие действия, которые должны предприниматься совместно.

Защита Америки в эпоху искусственного интеллекта

Расширенные возможности ИИ станут инструментами первой помощи в новую эру конфликта, поскольку стратегические конкуренты разрабатывают концепции ИИ и технологии для военных и других злонамеренных целей, а дешёвые и коммерчески доступные приложения ИИ, начиная от «глубоких подделок» и заканчивая смертоносными беспилотниками, становятся доступными государствам-изгоям, террористам и преступникам. Соединённые Штаты должны подготовиться к защите от этих угроз, быстро и ответственно приняв ИИ для целей национальной безопасности и обороны. Защита от способных к ИИ противников, работающих на машинных скоростях без использования ИИ, – это приглашение к катастрофе.

Люди-операторы не смогут идти в ногу или защищаться от кибератак или дезинформационных атак с использованием искусственного интеллекта, роёв беспилотников или ракетных атак без помощи машин с поддержкой искусственного интеллекта. Специалисты по национальной безопасности должны иметь доступ к лучшим мировым технологиям, чтобы защитить себя, выполнить свои миссии и защитить нас. Комиссия рекомендует правительству принять следующие меры:

Защищайтесь от возникающих угроз с помощью ИИ для свободного и открытого общества Америки.

Цифровая зависимость во всех сферах жизни превращает личную и коммерческую уязвимость в потенциальную слабость национальной безопасности.

Противники используют системы искусственного интеллекта для усиления дезинформационных кампаний и кибератак. Они собирают данные об американцах, чтобы построить профили их убеждений, поведения и биологического состава для целенаправленных попыток манипулировать или принуждать людей.

Эта нарастающая буря иностранного влияния и вмешательства требует организационных и политических реформ для укрепления нашей устойчивости.

Правительство должно создать целевую группу и круглосуточный оперативный центр для борьбы с цифровой дезинформацией. Она должна лучше защищать свои собственные базы данных и уделять приоритетное внимание безопасности данных в скрининге иностранных инвестиций, управлении рисками цепочки поставок и национальном законодательстве о защите данных. Правительство должно использовать киберзащиту с использованием искусственного интеллекта для защиты от кибератак с использованием искусственного интеллекта. И биобезопасность должна стать главным приоритетом в политике национальной безопасности.

Конкурентное военно-техническое преимущество наших вооружённых сил может быть утрачено в течение следующего десятилетия, если они не ускорят внедрение ИИ в своих миссиях. Для этого потребуется сочетать лидерство сверху вниз с инновациями снизу-вверх, чтобы внедрить операционно значимые приложения ИИ. Министерство обороны (МО) должно:

Во-первых, заложить основы для широкой интеграции ИИ к 2025 году. Это включает в себя создание общей цифровой инфраструктуры, развитие цифровой грамотной рабочей силы и внедрение более гибких процессов закупок, бюджета и надзора. Это также требует стратегического отказа от военных систем, которые плохо оснащены для ведения войны

с использованием искусственного интеллекта, и вместо этого инвестирования в возможности следующего поколения.

Во-вторых, достичь состояния боевой готовности ИИ к 2025 году. Руководство Пентагона должно действовать сейчас, чтобы провести организационные реформы, разработать инновационные концепции ведения боевых действий, установить цели эффективности ИИ и цифровой готовности и определить совместную сетевую архитектуру ведения боевых действий. Министерство обороны также должно расширить и сфокусировать свой портфель исследований и разработок в области ИИ.

Готовность также потребует содействия интероперабельности ИИ с союзниками и партнёрами.

Управление рисками, связанными с использованием ИИ и атомного оружия.

ИИ обеспечит новые уровни производительности и автономности систем вооружения. Но это также поднимает важные юридические, этические и стратегические вопросы, связанные с применением смертоносной силы.

При условии, что их использование санкционировано командиром или оператором-человеком, должным образом спроектированные и испытанные управляемые и автономные системы вооружения могут использоваться способами, совместимыми с международным гуманитарным правом.

Строгие существующие процедуры обзора и нацеливания вооружений Министерства обороны, включая его специальные протоколы для автономных систем вооружения и приверженность строгим этическим принципам ИИ, способны обеспечить, чтобы Соединённые Штаты выставляли безопасные и надёжные системы оружия с поддержкой ИИ и использовали их законным образом. Хотя в настоящее время нецелесообразно и не в интересах Соединённых Штатов добиваться глобального запрета на использование автономных систем оружия с поддержкой искусственного интеллекта, глобальное бесконтрольное использование таких систем может увеличить

риски непреднамеренной эскалации конфликтов и кризисной нестабильности. Чтобы снизить риски, Соединённые Штаты должны:

(1) чётко и публично подтвердить существующую политику США, согласно которой только люди могут санкционировать применение ядерного оружия, и добиваться аналогичных обязательств от России и Китая;

(2) создать площадки для обсуждения влияния ИИ на кризисную стабильность с конкурентами;

(3) разработать международные стандарты практики разработки, тестирования и использования автономных систем вооружения с поддержкой искусственного интеллекта.

Разведывательное сообщество (ИС) должно внедрить и интегрировать возможности ИИ во все аспекты своей работы от сбора данных до их анализа.

Разведка выиграет от ИИ больше, чем любая другая миссия национальной безопасности. Чтобы извлечь выгоду из ИИ, Управление директора национальной разведки должно расширить возможности и ресурсы своих научно-технических лидеров. Вся ИС должна использовать в своём анализе открытую и общедоступную информацию и уделять приоритетное внимание сбору научно-технической информации.

Для лучшего понимания разведывательным службам необходимо будет разработать инновационные подходы к объединению людей и машин, которые используют ИИ для усиления человеческого суждения.

Масштабируйте цифровые таланты в правительстве.

Агентствам национальной безопасности сейчас нужно больше цифровых экспертов, иначе они останутся неготовыми покупать, создавать и использовать ИИ и связанные с ним технологии.

Дефицит талантов в Минобороны и ИС представляет собой самое большое препятствие для того, чтобы быть уже к 2025 году. Правительству нужны новые трубопроводы талантов, включая Американскую Академию цифровых услуг для обучения нынешних и будущих сотрудников.

Ему нужен гражданский Национальный цифровой резервный корпус, чтобы набирать людей с нужными навыками включая отраслевых экспертов, учёных и недавних выпускников колледжей. И ему нужен Цифровой корпус, смоделированный по образцу Армейского медицинского корпуса, чтобы организовать технологов, уже служащих в правительстве.

Установите обоснованное доверие к системам ИИ.

Если системы ИИ обычно не работают так, как задумано, или непредсказуемы таким образом, что могут иметь значительные негативные последствия, то лидеры не будут их принимать, операторы не будут их использовать, Конгресс не будет их финансировать, а американский народ не будет их поддерживать. Чтобы установить обоснованное доверие, правительство должно сосредоточиться на обеспечении того, чтобы его системы ИИ были надёжными и надёжными, в том числе за счёт инвестиций в исследования и разработки (НИОКР) в области безопасности ИИ и продвижения объединения людей и ИИ в рамках устойчивой инициативы, возглавляемой национальными исследовательскими лабораториями.

Это также должно расширить возможности Министерства обороны по тестированию и оценке по мере роста числа, масштабов и сложности систем с поддержкой искусственного интеллекта.

Ответственные руководители ИИ высшего уровня должны быть назначены по всему правительству для улучшения исполнительного руководства и политического надзора.

Представьте демократическую модель использования ИИ для обеспечения национальной безопасности.

Инструменты ИИ имеют решающее значение для разведки США, национальной безопасности и правоохранительных органов.

Общественное доверие будет зависеть от обоснованной уверенности в том, что использование ИИ правительством будет уважать частную жизнь, гражданские свободы и гражданские права.

Правительство должно заслужить это доверие и обеспечить, чтобы его использование инструментов ИИ было эффективным, законным и законным. Этот императив требует разработки инструментов ИИ для усиления надзора и аудита, повышения общественной прозрачности использования ИИ и создания систем ИИ, которые продвигают цели сохранения конфиденциальности и справедливости. Это также требует обеспечения того, чтобы те, кто пострадал от действий правительства, связанных с ИИ, могли добиваться возмещения ущерба и иметь надлежащую правовую процедуру.

Правительство должно укрепить механизмы надзора и управления и создать целевую группу для оценки меняющихся проблем, связанных с ИИ и частной жизнью, гражданскими свободами и гражданскими правами.

Победа в технологическом конкурсе.

Гонка за исследованиями, разработкой и внедрением ИИ и связанных с ним технологий усиливает технологическую конкуренцию, которая лежит в основе более широкой стратегической конкуренции.

Китай организован, обеспечен ресурсами и полон решимости выиграть этот конкурс.

Соединённые Штаты сохраняют преимущества в критических областях, но нынешние тенденции вызывают беспокойство. В то время как конкурентная реакция осложняется глубокими академическими и коммерческими взаимосвязями, Соединённые Штаты должны сделать все возможное, чтобы сохранить своё инновационное лидерство и позиции в мире.

Правительство США должно принять конкуренцию ИИ и организовать, чтобы выиграть её, организовав преимущества сильных сторон США.

Организуйте с помощью стратегии под руководством Белого дома технологическую конкуренцию.

Соединённые Штаты должны поднять развитие ИИ с технического уровня на стратегический. Новые технологии, с использованием ИИ, сегодня лежат в основе нашего экономического процветания, безопасности и благосостояния.

Белый дом должен создать новый Совет по конкурентоспособности технологий во главе с вице-президентом для обеспечения безопасности, экономики и науки, разработки всеобъемлющей технологической стратегии и контроля за её реализацией.

Соединённые Штаты рискуют проиграть мировую конкуренцию за дефицитный опыт в области искусственного интеллекта, если они не будут развивать больше потенциальных талантов у себя дома и не наберут, и не удержат больше существующих талантов из-за рубежа.

Соединённые Штаты должны действовать агрессивно на обоих фронтах. Конгресс должен принять Закон об образовании в области национальной обороны II, чтобы устранить недостатки во всей американской системе образования от K-12 и переквалификации рабочих мест до инвестирования в тысячи стипендий для студентов и аспирантов в областях, имеющих решающее значение для будущего ИИ. В то же время Конгресс должен проводить всеобъемлющую иммиграционную стратегию для высококвалифицированных иммигрантов, чтобы стимулировать больше талантов в области ИИ учиться, работать и оставаться в Соединённых Штатах с помощью новых стимулов и реформ виз, грин-карт и мобильности рабочих мест.

Ускорьте внедрение инноваций в области ИИ в домашних условиях. Правительство должно сделать крупные новые инвестиции в AIR&D и создать национальную исследовательскую инфраструктуру ИИ, которая демократизирует доступ к ресурсам, питающим развитие ИИ по всей стране.

Правительство должно:

1 – удвоить не оборонное финансирование исследований и разработок в области ИИ ежегодно, чтобы к 2026 году достичь 32 млрд долларов в год,

создать Национальный технологический фонд и утроить число Национальных исследовательских институтов ИИ;

2 – создать Национальную исследовательскую инфраструктуру ИИ, состоящую из облачных вычислительных ресурсов, испытательных стендов, крупномасштабных открытых обучающих данных и открытой сети знаний, которая расширит доступ к ИИ и поддержит эксперименты в новых областях науки и техники;

3 – укрепить коммерческую конкурентоспособность путём создания рынков ИИ и формирования сети региональных инновационных кластеров.

Осуществлять всеобъемлющую политику и режимы в области интеллектуальной собственности (ИС).

Соединённые Штаты должны признать политику в области ИС в качестве приоритета национальной безопасности, имеющего решающее значение для сохранения лидерства Америки в области ИИ и новых технологий. Это особенно важно в свете усилий Китая по усилению и использованию политики в области ИС.

Соединённым Штатам не хватает всеобъемлющей политики в области ИС, необходимой для эры ИИ, и им мешает правовая неопределённость в нынешней доктрине патентоспособности и патентоспособности США.

Правительству США необходим план реформирования политики и режимов в области ИС таким образом, чтобы они отвечали приоритетам национальной безопасности.

Создать устойчивую отечественную базу для проектирования и производства микроэлектроники.

После десятилетий лидерства в индустрии микроэлектроники Соединённые Штаты теперь почти полностью полагаются на иностранные источники для производства передовых полупроводников, которые питают все алгоритмы искусственного интеллекта, критически важные для оборонных систем и всего остального. Проще говоря: американская цепочка поставок передовых чипов находится под угрозой без согласованных действий

правительства. Перестройка отечественного производства чипов обойдётся дорого, но время действовать пришло. Соединённые Штаты должны взять на себя обязательство придерживаться стратегии, направленной на то, чтобы оставаться по крайней мере на два поколения впереди Китая в области современной микроэлектроники, и взять на себя обязательства по финансированию и стимулированию для поддержания многочисленных источников производства ультрасовременной микроэлектроники в Соединённых Штатах.

Защите технологические преимущества Америки.

По мере того как границы технологических преимуществ США сужаются, а иностранные усилия по приобретению американских ноу-хау и технологий двойного назначения возрастают, Соединённые Штаты должны пересмотреть способы наилучшей защиты идей, технологий и компаний, не препятствуя чрезмерно инновациям. Соединённые Штаты должны:

Во-первых, модернизировать экспортный контроль и скрининг иностранных инвестиций для лучшей защиты критически важных технологий двойного назначения в том числе путём создания регулирующего потенциала и полного осуществления недавних законодательных реформ, осуществления скоординированного экспортного контроля за передовым оборудованием для производства полупроводников с союзниками и расширения требований к раскрытию информации для инвесторов из стран-конкурентов.

Во-вторых, защитить исследовательское предприятие США как национальное достояние-предоставив государственным учреждениям, правоохранительным органам и исследовательским институтам инструменты и ресурсы для проведения детальной оценки рисков и обмена информацией о конкретных угрозах и тактике, координируя усилия по защите исследований с союзниками и партнёрами, укрепляя поддержку кибербезопасности для исследовательских институтов и усиливая визовую проверку для ограничения проблемного сотрудничества в области исследований.

Создайте благоприятный международный технологический заказ. Соединённые Штаты должны работать рука об руку с союзниками и партнёрами в целях содействия использованию новых технологий для укрепления демократических норм и ценностей, координации политики и инвестиций для продвижения глобального внедрения цифровой инфраструктуры и технологий, защиты целостности международных технических стандартов, сотрудничества в продвижении инноваций в области искусственного интеллекта и обмена опытом и ресурсами для защиты от пагубного использования технологий и влияния авторитарных государств в демократических обществах.

Соединённые Штаты должны возглавить Коалицию по новым технологиям для достижения этих целей и создать Многосторонний научно-исследовательский институт ИИ, чтобы укрепить позиции Соединённых Штатов как глобального исследовательского центра по новым технологиям. Государственный департамент должен быть переориентирован, реорганизован и обеспечен ресурсами для руководства дипломатией в новых технологиях.

Побеждайте в конкурсах сопутствующих технологий.

Лидерство в области ИИ необходимо, но недостаточно для общего технологического лидерства США.

ИИ находится в центре созвездия новых технологий, позволяя развивать их. Поэтому Соединённые Штаты должны разработать единый авторитетный список технологий, которые будут лежать в основе национальной конкурентоспособности в 21 веке, и предпринять смелые действия, чтобы катализировать лидерство США в области искусственного интеллекта, микроэлектроники, биотехнологий, квантовых вычислений, 5G, робототехники и автономных систем, аддитивного производства и технологий хранения энергии. С. лидерство в этих технологиях требует инвестиций в конкретные платформы, которые позволят осуществить трансформационные прорывы и создать динамичные внутренние производственные экосистемы в каждой из

них. В то же время правительству необходимо будет постоянно выявлять и расставлять приоритеты в отношении новых технологий на дальнем горизонте.

Вывод

Эта новая эра конкуренции обещает изменить мир, в котором мы живём, и то, как мы живём в нём.

Мы можем либо сформировать грядущее изменение, либо быть сметены им. Теперь мы знаем, что использование ИИ во всех аспектах жизни будет расти, а темпы инноваций будут продолжать ускоряться.

Мы знаем, что противники полны решимости обратить возможности ИИ против нас.

Мы знаем, что Китай полон решимости превзойти НАС в лидерстве ИИ. Мы знаем, что достижения в области искусственного интеллекта строятся сами по себе и дают значительные преимущества первого перехода.

Теперь мы должны действовать. Принципы, которые мы устанавливаем, федеральные инвестиции, которые мы делаем, программы национальной безопасности, которые мы разрабатываем, организации, которые мы перестраиваем, партнёрские отношения, которые мы создаём, коалиции, которые мы создаём, и таланты, которые мы культивируем, определяют стратегический курс Америки.

Соединённые Штаты должны инвестировать все, что требуется, чтобы сохранить своё инновационное лидерство, ответственно использовать ИИ для защиты свободных людей и свободных обществ, а также продвигать границы науки на благо всего человечества.

ИИ собирается реорганизовать мир. Америка должна возглавить атаку.

Задача Комиссии национальной безопасности по искусственному интеллекту (NSCAI) состоит в том, чтобы дать рекомендации Президенту и Конгрессу «продвигать развитие искусственного интеллекта ИИ, машинного обучения и связанных с ними технологий для всестороннего удовлетворения потребностей национальной безопасности и обороны Соединённых Штатов.»

При создании Комиссии Раздел 1051 закона Джона С. Закон Маккейна о разрешении национальной обороны на 2019 финансовый год предписывает NSCAI изучать ИИ через призму национальной конкурентоспособности, средств поддержания технологического преимущества, тенденций в международном сотрудничестве и конкурентоспособности, путей стимулирования инвестиций в фундаментальные и передовые исследования, рабочую силу и подготовку кадров, потенциальных рисков военного использования, этических проблем, установления стандартов данных и стимулирования обмена данными, а также будущей эволюции ИИ.

15 членов комиссии были назначены Конгрессом и Исполнительной властью. Они представляют собой разноплановую группу технологов, руководителей предприятий, академических лидеров и специалистов в области национальной безопасности. Они подошли ко всем расследованиям в двухпартийной манере и достигли консенсуса по Окончательному докладу. В своей деятельности Комиссия руководствуется двумя принципами: необходимостью действий и важностью транспарентности.

Работа Комиссии включает в себя первоначальный доклад в июле 2019 года, промежуточные доклады в ноябре 2019 года и октябре 2020 года, два дополнительных ежеквартальных меморандума, серию специальных документов в ответ на пандемию COVID-19, а теперь и окончательный доклад. Ожидание представления рекомендаций в окончательном отчёте не было вариантом, когда мы начали нашу работу весной 2019 года.

Оценивать широкие последствия для национальной безопасности такой динамичной технологии, как ИИ, в один момент времени все равно что пытаться поймать молнию в бутылку. Учёные продолжают делать прорывы в области ИИ, а коммерческий сектор находит новые способы применения ИИ ускоряющимися темпами.

Конкуренты по всему миру разрабатывают стратегии ИИ и инвестируют ресурсы. Комиссия постоянно давала рекомендации, стремясь соответствовать

скорости развития ИИ и желаниям Исполнительной власти и Конгресса о помощи в принятии решения о том, что делать.

Конгресс уже принял ряд наших рекомендаций в докладе Уильяма М. (Мас) Закон о разрешении национальной обороны Торнберри на 2021 финансовый год, и Исполнительная власть также включила в него рекомендации. И мы постоянно стремимся учиться и обучать широкий круг заинтересованных сторон, чтобы сформировать общее понимание того, как ИИ повлияет на национальную безопасность.

NSCAA стремится к прозрачности. Как Федеральный консультативный комитет, он провёл пять открытых пленарных заседаний общей продолжительностью около 15 часов, транслировал прямые онлайн-трансляции и архивировал записи заседаний на веб-сайте NSCAI. Она ответила на более чем два десятка запросов Закона о свободе информации и опубликовала более 2500 страниц материалов. НСКАИ разместил более 700 страниц проектов материалов для общественного обсуждения и комментариев. За исключением материалов и вопросов, классифицированных по соображениям национальной безопасности, Комиссия стремилась обеспечить полную прозрачность. Мы активно взаимодействуем со средствами массовой информации после каждого пленарного заседания, ежеквартального отчёта и представления Конгрессу.

В десятках отдельных мероприятий мы сотрудничали с неправительственными организациями, федеральными правительственными организациями и международными организациями, чтобы донести наши рекомендации до средств массовой информации и общественности.

Самое главное, мы взяли на себя самые сложные вопросы с ИИ в общественных условиях и дали рекомендации только после консультаций с широким кругом гражданского общества, частного сектора и правительственных групп. Мы попытались выслушать и понять взгляды по всему спектру на глубоко сложные аспекты ИИ. Мы привлекли специалистов по этике, технологов и стратегов национальной безопасности. Мы говорили с воинами и дипломатами. Мы беседовали с учёными и предпринимателями. В

общей сложности члены комиссии и сотрудники приняли участие в сотнях дискуссий. По мере того как члены комиссии вырабатывали консенсус по рекомендациям, мы подходили к вопросам с осторожностью и смирением.

Окончательный Отчёт.

В Заключительном докладе представлены рекомендации NSCAI как стратегия победы в эпоху искусственного интеллекта.

В 16 главах Основного доклада содержатся рекомендации высшего уровня. В прилагаемых Планах действий изложены конкретные шаги, которые департаменты и учреждения могут предпринять для выполнения рекомендаций NSCAI. Комиссия предоставила как можно больше конкретики в том числе предоставив проекты законодательных текстов и исполнительных указов чтобы помочь президенту и Конгрессу быстро перейти от понимания ИИ к действиям на благо американского народа.

Технологии искусственного интеллекта (ИИ) обещают стать самыми мощными инструментами в поколениях для расширения знаний, повышения благосостояния и обогащения человеческого опыта.

Технологии станут основой инновационной экономики и источником огромной силы для стран, которые их используют. ИИ будет подпитывать конкуренцию между правительствами и компаниями, стремящимися внедрить его. И он будет использоваться национальными государствами для реализации своих стратегических амбиций.

США ещё не всерьёз задумались о том, насколько глубоко революция ИИ повлияет на общество, экономику и национальную безопасность. Недавние прорывы в области искусственного интеллекта, такие как победа компьютера над человеком в популярной стратегической игре Go1, шокировали другие страны, но не вызвали такой же реакции в Соединённых Штатах. Несмотря на наше лидерство в области ИИ в частном секторе и университетах, Соединённые Штаты по-прежнему не готовы к наступающей эре. Мы должны признать решительную роль, которую правительство должно будет сыграть в

обеспечении победы Соединённых Штатов в этом инновационном соревновании.

Конгресс и Президент должны будут поддерживать масштаб государственных ресурсов, необходимых для его достижения.

Масштаб технологических возможностей совпадает с моментом стратегической уязвимости. Китай-это конкурент, обладающий мощью, талантом и амбициями, чтобы бросить вызов технологическому лидерству Америки, её военному превосходству и её более широкому положению в мире.

ИИ усиливает угрозу кибератак и дезинформационных кампаний, которые Россия, Китай и другие государственные и негосударственные субъекты используют для проникновения в наше общество, кражи наших данных и вмешательства в нашу демократию.

Ограниченное использование атак с поддержкой искусственного интеллекта на сегодняшний день является верхушкой айсберга.

Между тем глобальные кризисы, примером которых являются глобальная пандемия и изменение климата, расширяют определение национальной безопасности и требуют инновационных технологических решений. ИИ может помочь нам справиться со многими из этих новых проблем.

Нам повезло. Революция ИИ не является стратегическим сюрпризом. Мы испытываем его влияние в нашей повседневной жизни и можем предвидеть, как прогресс исследований воплотится в реальные приложения, прежде чем нам придётся столкнуться со всеми последствиями национальной безопасности.

Эта комиссия может предупреждать о проблемах национальной безопасности и формулировать выгоды, а не объяснять, почему предыдущие предупреждения были проигнорированы и возможности были упущены. У нас всё ещё есть окно для внесения изменений, чтобы построить более безопасное и лучшее будущее. Темпы инноваций в области ИИ не являются стабильными, они ускоряются. Если Соединённые Штаты не будут действовать, они, скорее всего, потеряют свои лидерские позиции в области ИИ для Китая в следующем

десятилетия и станут более уязвимыми для спектра угроз с использованием ИИ со стороны множества государственных и негосударственных субъектов.

Комиссия приходит к выводу, что Соединённым Штатам необходимо реализовать стратегию защиты и конкуренции в эпоху искусственного интеллекта. Белый дом должен возглавить усилия по реорганизации правительства и переориентации нации.

В настоящем докладе представлены основные элементы стратегии:

Часть I, «Защита Америки в эпоху ИИ» (главы 1-8), описывает, что Соединённые Штаты должны делать для защиты от спектра угроз, связанных с ИИ, со стороны государственных и негосударственных субъектов, и рекомендует, как правительство США может ответственно использовать технологии ИИ для защиты американского народа и наших интересов.

Часть II, «Победа в технологическом соревновании», описывает роль ИИ в более широком технологическом соревновании. Каждая глава посвящена критическому элементу конкуренции и рекомендует меры, которые правительство должно предпринять для продвижения инноваций в области искусственного интеллекта для повышения национальной конкурентоспособности и защиты важнейших преимуществ США.

Почему ИИ Имеет Значение?

В 1901 году Томаса Эдисона попросили предсказать влияние электричества на человечество. Через два десятилетия после разработки электрической лампочки он предвидел универсальную технологию неограниченных возможностей. «Электричество – это поле полей», – сказал он. «Он хранит секреты, которые реорганизуют жизнь мира».

ИИ – это совершенно другой вид технологии общего назначения, но мы стоим на аналогичном этапе и видим столь же широкое воздействие. Быстро растущая способность компьютерных систем решать проблемы и выполнять задачи, которые в противном случае потребовали бы человеческого интеллекта,

трансформирует многие аспекты человеческой жизни и все области науки. Он будет включён практически во все будущие технологии. Вся инновационная база, поддерживающая нашу экономику и безопасность, будет использовать ИИ. Как это «поле полей» используется во благо и во зло будет перестраивать мир.

ИИ уже вездесущ в повседневной жизни, и темпы инноваций ускоряются. Мы считаем само собой разумеющимся, что ИИ уже формирует нашу жизнь как в малом, так и в большом. «Смартфон» имеет множество функций с поддержкой искусственного интеллекта, включая голосовые помощники, теги фотографий, безопасность распознавания лиц, поисковые приложения, рекомендательные и рекламные системы, а также менее очевидные улучшения ИИ в своей операционной системе.

ИИ помогает прогнозировать распространение и эскалацию пандемической вспышки, планировать и оптимизировать распределение товаров и услуг, контролировать транспортные потоки и безопасность, ускорять открытие лекарств и терапевтических средств и автоматизировать рутинные офисные функции.

Осознание темпов изменений имеет решающее значение для понимания мощи ИИ. Применение методов ИИ для решения проблем сжимает временные рамки инноваций и превращает некогда фантастические идеи в реальность по целому ряду дисциплин.

ИИ не может волшебным образом решать проблемы. Поскольку наука переходит из элитной ниши в мейнстрим-инструмент, инженерия будет столь же важна, как и научные прорывы. Ранние последователи в разных секторах извлекли аналогичные уроки: попытка использовать ИИ-это тяжёлая работа даже после того, как наука устоялась. Многие из наиболее важных реальных последствий будут связаны с выяснением того, как использовать существующие алгоритмы и системы искусственного интеллекта, некоторым из которых более десяти лет. Задача интеграции огромна. Использование данных, упрочнение и упаковка лабораторных алгоритмов, чтобы они были готовы к

использованию в полевых условиях, а также адаптация программного обеспечения ИИ к устаревшему оборудованию и жёстким организациям-все это требует времени, усилий и терпения. Интеграция ИИ часто требует преодоления существенных организационных и культурных барьеров и требует лидерства сверху вниз.

Передовые методы глубокого обучения часто являются непомерно дорогими, требующими огромных объёмов данных, вычислительной мощности и специальных знаний. Однако ИИ не будет принадлежать только крупным государствам и крупным технологиям. Многие инструменты машинного обучения, которые питают приложения ИИ, являются общедоступными и пригодными даже для неспециалистов.

Приложения с открытым исходным кодом и инструменты разработки в сочетании с недорогими облачными вычислениями и менее интенсивными подходами к обработке данных расширяют возможности ИИ по всему миру для государственных и негосударственных субъектов.

В современном обществе мы уже гораздо больше полагаемся на машины и автоматизацию, чем можем себе представить. Американские военные, например, десятилетиями использовали автономные системы. Однако по мере совершенствования возможностей ИИ динамика внутри человеко-машинных «команд» будет меняться. В прошлом компьютеры могли выполнять только те задачи, которые укладывались в чётко определённый набор параметров или правил, запрограммированных человеком. По мере того как ИИ станет более способным, компьютеры смогут изучать и выполнять задачи, основанные на параметрах, которые люди явно не программируют, создавая выбор и совершая действия с объёмом и скоростью, которые никогда раньше не были возможны. Во многих областях человеческой деятельности инновации в области ИИ поднимают важные вопросы о том, какой выбор следует делегировать интеллектуальным машинам, в каких обстоятельствах и по каким причинам.

В сфере национальной безопасности эти вопросы приобретают все большее значение по мере интеграции ИИ в оборонные и разведывательные

системы. Во всем нашем обществе мы должны будем решать эти новые сложные проблемы с помощью тонких подходов, интеллектуального любопытства и заботы, которые признают растущее повсеместное распространение ИИ.

Технологии, столь широко распространённые в других сферах жизни общества, окажут равноценное влияние на международную конкуренцию и конфликты. Мы должны принять искусственный интеллект, чтобы изменить способ защиты Америки, сдерживания противников, использования интеллекта для осмысления мира, борьбы и победы в войнах. Мужчины и женщины, которые защищают Соединённые Штаты, должны иметь возможность использовать ИИ и связанные с ним технологии, которые могут помочь им выполнить свои задачи как можно быстрее и безопаснее.

ИИ является квинтэссенцией технологии «двойного назначения» – его можно использовать как в гражданских, так и в военных целях.

Обещание ИИ, что машина может воспринимать, принимать решения и действовать быстрее, в более сложной среде, с большей точностью, чем человек представляет собой конкурентное преимущество в любой области.

Он будет использоваться в военных целях правительствами и негосударственными группами.

Мы можем ожидать масштабного распространения возможностей ИИ. Многие приложения ИИ для обеспечения национальной безопасности потребуют лишь скромных ресурсов и хорошего, но не большого опыта. Алгоритмы искусственного интеллекта часто доступны. Аппаратное обеспечение является «готовым» и в большинстве случаев общедоступным для потребителей (как, например, в случае с графическими процессорами). Возможности «Deepfake» могут быть легко загружены и использованы кем угодно. Инструменты AI-enabled и мутирующие вредоносные программы находятся в руках хакеров. Использование Азербайджаном турецких беспилотных летательных аппаратов и израильских бездельничающих

боеприпасов в боевых действиях против Армении в октябре 2020 года подтвердило распространение автономных военных возможностей. Многие государства наблюдают и учатся на этом опыте. Возрастает вероятность безрассудного или неэтичного использования технологий, основанных на ИИ, государствами-изгоями, преступниками или террористами.

Государственные и негосударственные субъекты, решившие бросить вызов Соединённым Штатам, но избегающие прямой военной конфронтации, будут использовать ИИ для усиления существующих инструментов и разработки новых. Противники используют нашу цифровую открытость посредством ускоренных ИИ информационных операций и кибератак. Adtech станет patsec-tech, поскольку противники признают то, что рекламные и технологические фирмы признавали годами: что машинное обучение является мощным инструментом сбора и анализа данных и таргетирования деятельности. Используя шпионаж и общедоступные данные, противники будут собирать информацию и использовать ИИ для выявления уязвимостей в отдельных людях, обществе и критической инфраструктуре. Они будут моделировать, как лучше всего манипулировать поведением, а затем действовать.

Приложения ИИ помогут военным быстрее и эффективнее готовиться, чувствовать и понимать, принимать решения и действовать. Многочисленные системы вооружения будут использовать одну или несколько технологий искусственного интеллекта. Системы искусственного интеллекта будут генерировать варианты для командиров и создавать боевые сети, соединяющие системы во всех доменах. Он преобразует логистику, закупки, обучение, а также проектирование и разработку нового оборудования.

Внедрение ИИ потребует разработки новых тактических и оперативных концепций. В будущем война будет сталкивать алгоритм с алгоритмом. Источники преимущества на поле боя перейдут от традиционных факторов, таких как размер сил и уровень вооружения,

к таким факторам, как превосходный сбор и усвоение данных, связь, вычислительная мощность, алгоритмы и безопасность системы.

Концепции ИИ и технологии военного назначения.

Россия планирует автоматизировать значительную часть своих военных систем. Она развернула автономные системы в Сирии для испытаний на поле боя.

Китай рассматривает ИИ как способ компенсировать американское обычное военное превосходство путём «прыжка» к новому поколению – технологий. Его военные приняли «интеллектуальную войну» – инвестируя, например, в роящиеся беспилотники, чтобы противостоять США военно-морском деле. Китайские военные лидеры открыто говорят об использовании систем искусственного интеллекта для «разведки, электромагнитного противодействия и скоординированных огневых ударов». Китай тестирует и обучает алгоритмы искусственного интеллекта в военных играх, разработанных на основе реальных сценариев.

Поскольку эти авторитарные государства создают новые военные системы с поддержкой искусственного интеллекта, США обеспокоены тем, что они не будут ограничены теми же строгими испытаниями и этическим кодексом, которыми руководствуются американские военные.

ИИ произведёт революцию в практике интеллекта. Возможно, нет функции национальной безопасности, более подходящей для внедрения ИИ, чем разведка и анализ. Машины будут просеивать массивы данных, собранных из всех источников, находить критическую информацию, переводить языки, объединять наборы данных из разных областей, выявлять корреляции и связи, перенаправлять ресурсы и информировать аналитиков и лиц, принимающих решения. Чтобы защитить американский народ, возможно, самой неотложной и убедительной причиной для ускорения использования ИИ в интересах национальной безопасности является возможность того, что более продвинутый машинный анализ сможет найти и соединить точки перед

следующей атакой, когда один человеческий анализ не сможет увидеть полную картину так ясно.

Защита от противников, способных использовать ИИ, без использования ИИ-это приглашение к катастрофе. ИИ будет сжимать временные рамки принятия решений от минут до секунд, расширять масштабы атак и требовать ответных действий, которые будут обременять пределы человеческого познания. Люди-операторы не смогут защищаться от кибер-или дезинформационных атак с использованием искусственного интеллекта, роёв беспилотников или ракетных атак без помощи машин с поддержкой искусственного интеллекта.

Лучший человек-оператор не может защититься от множества машин, совершающих тысячи манёвров в секунду, потенциально движущихся на гиперзвуковых скоростях и управляемых ИИ в разных областях.

Люди не могут быть везде одновременно, но программное обеспечение может.

Убедительная логика диктует быстрое, но осторожное и ответственное внедрение ИИ. Правительство должно принять ИИ, следуя принципу легендарного баскетбольного тренера Джона Вудена: «Будь быстрым, но не спеши». Как и другие «критические для безопасности» приложения ИИ, военные и разведывательные функции требуют обдумывания и осторожности, прежде чем они будут разработаны и внедрены. Некоторые современные системы искусственного интеллекта узки и хрупки. Все они требуют тщательного тестирования, гарантий и понимания того, как они могут работать иначе в реальном мире, чем на испытательном стенде. Автономные системы вооружения с поддержкой искусственного интеллекта могли бы быть более точными и, как следствие, уменьшить непреднамеренные жертвы среди гражданского населения. Но они также поднимают важные этические вопросы о роли человеческого суждения в применении смертоносной силы. При неправильном проектировании или использовании они также могут увеличить риск военной эскалации.

Складывается консенсус относительно принципов ответственного использования ИИ в оборонном и разведывательном сообществах. Если машина, работающая на ИИ, не работает так, как задумано, предсказуемо и руководствуется чёткими принципами, то операторы не будут её использовать, организации не примут её, а американский народ не поддержит.

«Лучший человек-оператор не может защититься от множества машин, совершающих сотни манёвров в секунду, потенциально движущихся с гиперзвуковой скоростью и управляемых ИИ в разных областях. Люди не могут быть везде одновременно, но программное обеспечение может.»

Спешка была бы контрпродуктивной и опасной, если бы она заставила американцев потерять уверенность в преимуществах, которые может дать ИИ. Риск, однако, неизбежен. Неспособность использовать ИИ для решения реальных задач национальной безопасности рискует поставить Соединённые Штаты в невыгодное положение, сделать американских военнослужащих более уязвимыми и неразумно тратить деньги налогоплательщиков на устаревшее и неэффективное оборудование.

Отсрочка внедрения ИИ перенесёт весь риск на следующее поколение американцев, которым придётся защищаться и, возможно, бороться с противником 21-го века с помощью инструментов 20-го века.

Правительство США по-прежнему действует со скоростью человека, а не машины. Внедрение ИИ требует глубоких изменений в деловой практике национальной безопасности, организационной культуре и менталитете от тактического до стратегического уровня – от поля боя до Пентагона. Правительство отстаёт от коммерческого уровня техники в большинстве категорий ИИ, включая базовую автоматизацию бизнеса.

Он страдает от технических недостатков, которые варьируются от нехватки цифровой рабочей силы до неадекватной политики приобретения, недостаточной сетевой архитектуры и слабой практики обработки данных.

Бюрократия препятствует улучшению партнёрских отношений с лидерами ИИ в частном секторе, которые могли бы помочь. Правительство

должно стать лучшим клиентом и лучшим партнёром. Инновации в области национальной безопасности в отсутствие такого импульса, как крупная война или террористический акт, потребуют сильного руководства.

В дополнение к узким областям применения ИИ в области национальной безопасности и обороны, ИИ является точкой опоры более широкой технологической конкуренции в мире.

ИИ будет использоваться для продвижения всех аспектов национальной мощи от здравоохранения до производства продуктов питания и экологической устойчивости. Успешное внедрение ИИ в смежных областях и технологиях будет стимулировать экономику, формировать общество и определять, какие государства оказывают влияние и осуществляют власть в мире.

Многие страны имеют национальные стратегии ИИ, но только Соединённые Штаты и Китай имеют ресурсы, коммерческую мощь, кадровый резерв и инновационную экосистему, чтобы возглавить мир в области ИИ. В некоторых областях исследований и приложений Китай уже является сверстником ИИ, и в некоторых областях он более технически продвинут. В течение следующего десятилетия Китай может превзойти Соединённые Штаты в качестве мировой сверхдержавы в области искусственного интеллекта.

На равных условиях Соединённые Штаты способны опередить любого конкурента. Однако сегодня существует фундаментальная разница в подходах США и Китая к инновациям в области ИИ, которая ставит под угрозу американское лидерство в области ИИ.

На протяжении десятилетий американская инновационная модель была предметом зависти всего мира. Открытый обмен идеями, свободные рынки и ограниченное участие правительства в поддержке фундаментальных исследований являются столпами американского пути инноваций и отражают американские ценности. В Америке технологические фирмы конкурируют за долю рынка. Они не являются инструментами государственной власти.

Исследователи сотрудничают в открытой исследовательской среде, соревнуясь со своими коллегами, чтобы сделать прорывы в области искусственного интеллекта, не обращая внимания на границы. Международный поток венчурного капитала и коммерции, связанной с искусственным интеллектом, поощряется по мере того, как фирмы конкурируют за прибыль и следующую крупную идею.

Большая часть прогресса в области ИИ в Соединённых Штатах должна оставаться за частным сектором и университетами. Мы не должны терять инновационную культуру, которая строится снизу-вверх и пропитана менталитетом гаражного стартапа. Однако полностью распределённый подход не является выигрышной стратегией в этом стратегическом соревновании. Даже крупные технологические компании не могут конкурировать с ресурсами Китая или делать большие инвестиции, которые понадобятся США, чтобы оставаться впереди. Понадобится гибридный подход, объединяющий усилия правительства и частного сектора, чтобы выиграть технологическую конкуренцию.

Китай организован, обеспечен ресурсами и полон решимости победить в технологическом соревновании. ИИ занимает центральное место в глобальной экспансии Китая, его экономической и военной мощи и внутренней стабильности. У него есть фора для выполнения национального плана ИИ в рамках более масштабных планов по лидерству в мире в нескольких критических и новых технологических областях.

Начиная с 2017 года Китай установил цели, задачи и стратегии ИИ, привязанные к конкретным временным рамкам, с ресурсами, подкреплёнными преданным лидерством, чтобы возглавить мир в области ИИ к 2030 году. Китай выполняет централизованно направленный систематический план по извлечению знаний ИИ из-за рубежа посредством шпионажа, набора талантов, передачи технологий и инвестиций. У Китая есть амбициозные планы по созданию и обучению нового поколения инженеров ИИ в новых центрах ИИ. Он поддерживает фирмы «национальные чемпионы» (в том числе Huawei,

Baidu, Alibaba, Tencent, iFlytek и SenseTime), чтобы вести разработку технологий искусственного интеллекта внутри страны, продвигать государственные приоритеты, которые питают военные программы и программы безопасности под рубрикой военно-гражданского слияния, и захватывать рынки за рубежом. КНР финансирует масштабные проекты цифровой инфраструктуры на нескольких континентах. Китай разработал стратегию интеллектуальной собственности (ИС) и пытается установить глобальные технические стандарты для развития ИИ. И его законы делают практически невозможным для компаний в Китае защитить свои данные от властей.

Достижения в области искусственного интеллекта способствуют широкой конкуренции платформ в области электронной коммерции, поисковых систем, социальных сетей и многого другого.

Страны, компании и исследователи, победившие в конкурсе ИИ в области вычислительной техники, данных, талантов и коммерциализации смогут выиграть гораздо более масштабную игру. По сути, все больше и больше данных, питаемых большей базой потребителей/участников, создают лучшие алгоритмы, которые дают лучшие результаты, что, в свою очередь, приводит к большему количеству пользователей, большему количеству данных и лучшей производительности до тех пор, пока, в конечном счёте, они не станут доминирующими платформами. Если китайские фирмы выиграют эти соревнования, это не только поставит в невыгодное положение американские коммерческие фирмы, но и создаст цифровую основу для геополитического вызова Соединённым Штатам и их союзникам. Доминирование платформы за рубежом позволяет Китаю собирать данные своих пользователей и позволяет Китаю расширять аспекты своей внутренней системы контроля.

Где бы Китай ни контролировал цифровую инфраструктуру, платформы социальных сетей и электронную коммерцию, он будет обладать большими рычагами и властью для принуждения, пропаганды и формирования мира в соответствии со своими целями.

Конкуренция ИИ осложняется глубокими взаимосвязями. Соединённые Штаты и Китай не действуют параллельно, как это делали СССР и американцы в космической гонке, с разобщёнными научно-исследовательскими и опытно-конструкторскими предприятиями и минимальными коммерческими контактами.

Исследовательские экосистемы в Китае и Соединённых Штатах тесно связаны через совместные исследовательские проекты, циркуляцию талантов (особенно из Китая в Соединённые Штаты) и коммерческие связи, которые включают цепочки поставок, рынки и совместные исследовательские предприятия. Было бы контрпродуктивно разрывать технологические связи с Китаем, которые приносят пользу фундаментальным исследованиям и американским компаниям. Однако Соединённые Штаты должны защищать целостность открытых исследований, предотвращать кражу американской интеллектуальной собственности и использовать целенаправленные инструменты, такие как экспортный контроль и скрининг инвестиций, для защиты технологических отраслей, имеющих решающее значение для национальной безопасности.

Соединённые Штаты сохраняют преимущества в критических областях, но тенденции вызывают беспокойство. Лучшие в мире научные таланты сегодня с большей вероятностью останутся дома или мигрируют куда-то ещё. Лидерство США в микроэлектронике, аппаратном обеспечении, на котором работает весь ИИ уменьшилось, и для передовых чипов оно зависит от иностранных цепочек поставок и производителей в Азии, которые уязвимы для принуждения или нарушения. В то время как многие инструменты машинного обучения широко доступны, а затраты на вычислительную единицу снизились, вычислительная мощность и доступ к данным, необходимые для передовых прорывных исследований в области глубокого обучения, затрудняют конкуренцию университетским исследователям и небольшим компаниям. География инноваций по-прежнему сосредоточена лишь в некоторых частях страны.

Правительство США должно принять практический подход к национальной технологической конкурентоспособности. Содействие развитию разнообразной и устойчивой экосистемы НИОКР и коммерческого сектора является обязанностью правительства.

Расширение каналов талантов для привлечения лучших специалистов мира и удвоение усилий по обучению американцев – это выбор государственной политики. Необходимо разумно и настойчиво защищать критически важную интеллектуальную собственность ИИ, пресекать систематическую кампанию незаконной передачи знаний, проводимую конкурентами. Защита преимуществ аппаратного обеспечения и создание устойчивости в цепочках поставок требуют законодательства и федеральных стимулов. Объединение единомышленников-союзников и партнёров для создания международной коалиции, обеспечивающей демократическое видение ИИ, которое будет формировать цифровое будущее, требует дипломатии под руководством США.

Конкуренция ИИ потребует лидерства Белого дома. Важнейшие элементы стратегии слишком сложны для руководства каким-либо одним ведомством или ведомством, поскольку они пересекаются с национальной безопасностью, экономической и технологической политикой. Только сильное исполнительное руководство из Белого дома может управлять политикой, вынуждать к компромиссам и мобилизовать страну на необходимые инвестиции.

Широкое внедрение ИИ правительствами по всему миру оказывает влияние не только на международный порядок между государствами, но и на политический порядок внутри них. Ставки на будущее ИИ тесно связаны с непрекращающейся борьбой между авторитарными и демократическими политическими системами, и идеологиями.

Сама технология не обладает идеологией, но то, как она разработана, где она используется и какие законы управляют её использованием, отражают приоритеты и ценности тех, кто её разрабатывает и использует. Вскоре

большинство или все правительства получают больше возможностей для наблюдения и анализа с помощью искусственного интеллекта. По мере распространения технологии основное различие между государствами будет заключаться не столько в качестве или сложности технологии, сколько в том, как она используется, с какой целью и по каким правилам.

Авторитарные режимы будут продолжать использовать ИИ в распознавании лиц, биометрии, прогнозной аналитике и слиянии данных в качестве инструментов наблюдения, влияния и политического контроля.

Использование Китаем технологий искусственного интеллекта для подавления своего уйгурского меньшинства и мониторинга всех своих граждан предвещает, как авторитарные режимы будут использовать системы искусственного интеллекта для облегчения цензуры, отслеживания физических движений и цифровой активности своих граждан и подавления инакомыслия.³²³ Глобальное распространение этих цифровых систем создаёт перспективу более широкого внедрения авторитарного управления.

Но либеральные демократии также используют ИИ для целей внутренней безопасности и общественной безопасности. Более половины развитых демократий мира используют системы наблюдения с поддержкой искусственного интеллекта. Такие технологии имеют законные общественные цели и совместимы с верховенством права. Тем не менее, в государствах, склоняющихся к нелиберальной практике, использование цифровых инструментов таким образом, чтобы подорвать верховенство закона, может склонить чашу весов в сторону дальнейшего демократического отката. Сохранение индивидуальных свобод требует постоянной бдительности. Ответственная демократия должна обеспечить, чтобы использование ИИ правительством было ограничено разумными ограничениями в соответствии с правами и свободами, которые определяют свободное и открытое общество.

Правительство США должно разработать и внедрить технологии с поддержкой ИИ с достаточной прозрачностью, строгим надзором и

подотчётностью для защиты от неправильного использования. Просто заявить о несогласии США с авторитарным использованием ИИ недостаточно. Соединённые Штаты также должны продемонстрировать, как демократия должна использовать ИИ для защиты безопасности своих граждан, поддерживая либерально-демократические ценности.

Существует настоятельная необходимость в использовании ИИ в целях национальной безопасности, например, против иностранных и внутренних террористов, действующих в пределах наших границ. Существует также непреходящая необходимость обеспечить, чтобы приложения безопасности ИИ соответствовали основным ценностям индивидуальной свободы и равной защиты в соответствии с законом.

Соединённые Штаты должны возглавить коалицию демократий. Поскольку мы гарантируем, что ИИ разрабатывается и используется таким образом, который безопасен для демократии внутри страны, мы также должны продвигать глобальные нормы, чтобы сделать его использование безопасным для демократии за рубежом. В то время как способность правительства США влиять на практику управления другими государствами ограничена, сильной стороной внешней политики США в отношении ИИ должно быть поощрение прав человека и противодействие техно авторитарным тенденциям.

Соединённые Штаты могут использовать дипломатию и использовать свои глобальные партнёрские отношения для пропаганды создания технических стандартов и норм защиты частной жизни в международных органах, а также могут работать с государствами-единомышленниками, чтобы обеспечить другим странам альтернативу использованию китайских технологий и методов социального контроля и доступа к технологиям, защищающим демократические ценности, такие как конфиденциальность. Мы не стремимся к фрагментированному цифровому миру. Мы хотим, чтобы Соединённые Штаты и их союзники существовали в мире с разнообразным набором вариантов в цифровой инфраструктуре, электронной коммерции и социальных сетях, которые не будут уязвимы для авторитарного принуждения

и которые поддерживают свободу слова, права личности, неприкосновенность частной жизни и терпимость к различным взглядам.

Вывод

США находится в начале новой эры конкуренции. Теперь мы знаем, что использование ИИ во всех аспектах жизни будет расти, а темпы инноваций ускорятся. В тоже время противники полны решимости обратить возможности ИИ против нас. Мы знаем, что конкурент полон решимости превзойти нас в лидерстве ИИ. Ещё мы знаем, что ИИ ускоряет прорывы в самых разных областях и тот, кто первым переведёт разработки ИИ в прикладные программы, получит преимущество. Теперь мы должны действовать. Принципы, которые мы устанавливаем, федеральные инвестиции, которые мы делаем, программы национальной безопасности, которые мы разрабатываем, организации, которые мы перестраиваем, партнёрские отношения, которые мы создаём, коалиции, которые мы создаём, и таланты, которые мы культивируем, определяют стратегический курс Америки. Соединённые Штаты должны инвестировать всё, что требуется, чтобы сохранить своё инновационное лидерство, ответственно использовать ИИ для защиты свободных людей и свободных обществ, а также продвигать границы науки на благо всего человечества. Искусственный интеллект (ИИ) – это не отдельная часть аппаратного или программного обеспечения, а скорее совокупность технологий, зависящих от взаимосвязанных элементов, которые можно представить в виде стека.

Для рассмотрения такой широкой темы законодательный мандат Комиссии предусматривал указания относительно того, как охватить её работу, включив в неё технологии, которые решают задачи, требующие человекоподобного восприятия, познания, планирования, обучения, коммуникации или физических действий и технологий, которые могут учиться и действовать автономно, будь то в форме программных агентов или воплощённых роботов.

Успешная разработка и внедрение технологий ИИ зависит от ряда взаимосвязанных элементов, которые можно представить в виде стека.

ИИ требует таланта, данных, оборудования, алгоритмов, приложений и интеграции. Мы рассматриваем талант как наиболее существенное требование, потому что он управляет созданием и управлением всеми другими элементами. Данные имеют решающее значение для большинства систем искусственного интеллекта.³ Маркированные и кураторские данные позволяют использовать большую часть текущего машинного обучения (ML) для создания новых приложений и повышения производительности существующих приложений искусственного интеллекта.

Базовое оборудование обеспечивает вычислительную мощность для анализа постоянно растущих пулов данных и запуска приложений.

Этот аппаратный уровень включает облачные вычисления и хранилища, поддерживаемые сетью и коммуникационной магистралью, инструментальной для подключения интеллектуальных датчиков и устройств на границе сети.

Алгоритмы-это математические операции, которые говорят системе, как ориентироваться в данных, чтобы дать ответы на конкретные вопросы. Приложение делает ответы полезными для конкретных задач.

Интеграция этих элементов имеет решающее значение для создания успешной сквозной системы искусственного интеллекта. Это требует значительных инженерных талантов и инвестиций для интеграции существующих потоков данных, трубопроводов принятия решений, устаревшего оборудования, тестовых проектов и т.д. Эта задача интеграции может быть сложной и исторически недооценивалась.

Технологии и приложения искусственного интеллекта, такие как распознавание образов, мл, компьютерное зрение, понимание естественного языка и распознавание речи, развивались в течение многих десятилетий.

В первые годы существования ИИ, период, который Агентство перспективных исследовательских проектов министерства обороны (DARPA) называет «первой волной», исследователи исследовали множество подходов,

включая символическую логику, экспертные системы и планирование. Некоторые из наиболее эффективных результатов были основаны на «ручном знании», определяемом людьми, а затем используемом машиной для рассуждения и взаимодействия.

За последние 10 лет мы стали свидетелями «второй волны» ИИ, движимой крупномасштабным статистическим мл, который позволяет инженерам создавать модели, которые могут быть обучены конкретным проблемным областям, если им даны примеры данных или смоделированы взаимодействия. Обучаясь на основе данных, эти системы предназначены для решения конкретных задач и достижения конкретных целей с компетенциями, которые в некоторых отношениях параллельны когнитивным процессам человека: восприятию, рассуждению, обучению, общению, принятию решений и действию.

Сегодня большинство полевых крупномасштабных систем ИИ используют элементы подходов ИИ как первой, так и второй волны.

Сегодня мы достигли точки перегиба. Глобальная цифровая трансформация привела к огромному количеству данных. Статистические алгоритмы мл, особенно глубокие нейронные сети, созрели как решатели проблем хотя и с ограничениями.⁶ Мощные и сетевые вычисления, которые подпитывают возможности мл, стали широко доступными. Конвергенция этих факторов теперь ставит эту способную технологию в руки как технических, так и нетехнических.

Фундаментальный «вопрос уже не в том, как работает эта технология, а в том, что она может сделать для вас.»

Хотя нынешняя технология все ещё имеет значительные ограничения, она хорошо подходит для определённых случаев использования. Мы вступили в эпоху развёрнутого ИИ.

ИИ теперь вездесущ, встроен в устройства, которые мы используем и с которыми взаимодействуем ежедневно, например, в наших смартфонах, встроены беспроводные маршрутизаторы для автомобилей. Мы обычно

полагаемся на приложения, обогащённые искусственным интеллектом, будь то поиск нового ресторана, навигация по трафику, выбор фильма или обслуживание клиентов по телефону или онлайн.

Современное применение ИИ в ключевых областях.

Предсказание		Современное применение ИИ в ключевых областях	Понимание естественного языка (NLU)	
Прогнозирование событий и анализ закономерностей повлияли почти на все отрасли (например, финансы, сельское хозяйство и транспорт).	Примеры: Профилактическое Техническое Обслуживание Транспортных Средств Точное земледелие		Машины способны обрабатывать, анализировать, понимать и имитировать человеческий язык, как устный, так и письменный.	Пример: Более богатое взаимодействие Человека и Компьютера
Планирование и оптимизация			Компьютерное зрение	
Определение необходимых шагов для выполнения ряда задач может сэкономить время и деньги, а также повысить безопасность.	Пример: Транспортное планирование в городах		Восприятие и изучение визуальных задач из окружающего мира с помощью камер и датчиков.	Пример: Мониторинг животноводства
Моделирование и имитация			Роботизированная Автоматизация технологических процессов (RPA)	
Моделирование нашего физического, экономического и	Примеры: Исследование COVID-19		Программное обеспечение робототехники поможет организациям	Пример: Платформы RPA

социального мира для поддержки изучения, оптимизации и тестирования операций с помощью моделирования без вмешательства или прерывания текущих процессов.	Отслеживание Космического мусора		автоматизировать утомительные и повторяющиеся задачи.	

Прогнозировать будущее ИИ сложно. Пять лет назад мало кто мог предсказать недавние прорывы в понимании естественного языка, которые привели к созданию систем, способных генерировать полный текст, почти неотличимый от человеческой прозы.

При заметном росте инвестиций в глобальную индустрию ИИ за последние пять лет и беспрецедентном количестве долларов, инвестируемых на НИОКР, во всем мире, замедления ИИ не предвидится, только новые горизонты для развёрнутого ИИ.

Границы технологии искусственного интеллекта.

Следующее десятилетие исследований ИИ, скорее всего, будет определяться усилиями по включению существующих знаний, продвижению новых способов обучения и созданию более надёжных, обобщаемых и надёжных систем.

Исследования по развитию человеко-машинного объединения будут находиться на переднем крае, как и совершенствование гибридных методов ИИ, усовершенствованных методов обучения и объяснимого ИИ.

Объединение людей и ИИ. Овладение сотрудничеством и объединением человека и ИИ является основополагающим элементом для будущего применения ИИ.

Синергия между людьми и ИИ обещает большее, чем сумма его частей. Исследователи решают эту проблему, изучая вопросы делегирования полномочий, наблюдаемости, предсказуемости, направленности и доверия. Более глубокое понимание того, как люди будут учиться работать с ИИ, даст понимание для создания эффективных обучающих программ для людей. Прогресс в понимании языка преследуется для создания систем, которые могут суммировать сложные входные данные и вступать в диалог с помощью человеческого общения-важнейшего компонента объединения следующего поколения.

Граница объединения включает в себя необходимость совместной разведки среди когорт агентов, будь то смешанные группы людей и машин или команды координирующих машин.

Новые способы обучения. Новые методы обучения позволяют повысить эффективность как обучения, так и вывода из данных. Это уменьшает зависимость от обширных массивов данных и расширяет возможности систем для решения задач, выходящих за рамки их первоначального объёма, выстраивая пути к контекстуальному обучению и здравому смыслу. Гибридные методы ИИ объединяют различные подходы ИИ, чтобы извлечь выгоду из их взаимодополняющих сильных сторон.

Например, нейро-символические исследования сочетают символические манипуляции с нейронными сетями.

Подходы, основанные на моделях и данных, также могут быть объединены, например, использование знаний физики в рамках статистических систем мл. Исследователи также продвигают методы контролируемого обучения с низким запасом маркированных данных, в то время как другие разработали более эффективные методы маркировки данных.

Одним из таких перспективных подходов является генерация синтетических данных с помощью моделирования. Это позволяет модели видеть условия и сценарии, с которыми она, возможно, не сталкивалась с реальным набором данных, сохраняя при этом отношения между важными переменными в исходных данных и конфиденциальность конфиденциальных данных.

Граничные вычисления. Преодоление барьеров размера, веса и мощности также увеличивает вездесущность ИИ и способствует защите частной жизни. Компании работают над тем, чтобы упаковать больше вычислительной мощности в более жёсткие, специализированные чипы, которые используют меньше энергии для обучения и запуска одних и тех же моделей. Такие чипы позволяют потребительским устройствам запускать сложные модели локально, а не передавать данные извне и ждать, пока модели будут запущены удалённо.

Сохранение данных полностью на устройстве, где модель обучается или запускается, это прогресс, который потенциально может повысить индивидуальную конфиденциальность в системах, работающих на ИИ.

Успехи в рассуждении. По сравнению с людьми, даже нашим самым способным современным системам ИИ не хватает того, что можно было бы назвать «рассуждением здравого смысла».

Предпринимаются усилия по созданию систем, способных обобщать знания и переводить обучение в различные области.

Система ИИ, наделённая здравым смыслом, могла бы эффективно моделировать человеческую способность делать и использовать предположения о физических свойствах, целях, намерениях и поведении людей и объектов и тем самым характеризовать вероятные последствия действия или взаимодействия. Достижения в области категоризации, в создании обобщённых структурированных онтологий и в понимании языка будут стимулировать способность машин учиться, понимая контекст и содержание, и позволят людям находить быстрые решения проблем, на изучение которых исторически потребовались бы годы. Это исследование обещает проложить путь к более

объяснимому ИИ наряду с большей способностью обнаруживать и смягчать предвзятость, что будет иметь важное значение для повышения надёжности этих более общих технологий ИИ.

К Более Общему Искусственному Интеллекту. Решения ИИ на сегодняшний день демонстрируют узкие и глубокие компетенции, но с фундаментальным отличием от возможностей, демонстрируемых людьми.

Люди выполняют задачи, обучаясь без явных контролируемых сигналов, они обобщают навыки, необходимые для одной задачи, и применяют их к другим задачам, накапливают, манипулируют и рассуждают с большим количеством знаний здравого смысла.

Некоторые исследователи используют термин «искусственный общий интеллект» (ИИ) для обозначения цели расширения ИИ за пределы узких, вертикальных клиньев знаний.

Дебаты были сосредоточены на том, могут ли быть конкретные прорывы, которые приведут к более общим, человекоподобным способностям, или же эта область, скорее всего, будет продолжать продвигать более общий ИИ по одному или нескольким измерениям навыков.

Независимо от перспективы, для создания более общих систем искусственного интеллекта потребуется значительный прогресс во всех областях исследований, упомянутых в этом разделе. Если это будет достигнуто, то более общие методы ИИ могут иметь огромные преимущества, но также могут привести к новым рискам, если не будут решены проблемы безопасности. Хотя прорывы ни в коей мере не гарантируются, Соединённые Штаты должны продолжать исследования систем с более человеческими возможностями, сопровождаемые соразмерными инвестициями, чтобы обеспечить безопасность и управляемость этих систем.

Достижения в области ИИ, включая овладение более общими возможностями ИИ в одном или нескольких измерениях, скорее всего, обеспечат новые возможности и приложения.

Некоторые из этих достижений могут привести к переломным моментам или скачкам в возможностях.

Такие достижения могут также привести к возникновению новых проблем и рисков, а также к необходимости разработки новой политики, рекомендаций и технических достижений для обеспечения соответствия систем целям и ценностям, включая безопасность, надёжность и надежность. США должны следить за прогрессом в области ИИ и делать необходимые инвестиции в технологии, а также уделять внимание политике, чтобы обеспечить соответствие систем ИИ и их использования нашим целям и ценностям.

Глядя в будущее с поддержкой искусственного интеллекта. Следуя траекториям исследовательских потоков, описанных выше, мы видим будущее, в котором ИИ даёт человечеству беспрецедентные возможности, открывая возможности в науке, образовании, космической технике, здравоохранении, инфраструктуре, производстве, сельском хозяйстве, развлечениях и бесчисленном множестве других секторов.

Например, прогресс в понимании естественного языка может обеспечить повсеместный перевод в реальном времени для более неясных языков, для которых письменные и устные данные обучения ограничены. Это изменило бы способ нашего общения через географические и культурные барьеры, открыв возможности для бизнеса, дипломатии и свободного обмена идеями.

Прорывы в интеграции мультимодальных, многоисточниковых данных могут позволить в режиме реального времени моделировать ИИ-управляемое моделирование для федеральных мер реагирования на кризисы, включая пандемии и стихийные бедствия.

Каналы беспилотных летательных аппаратов, дополненные картами, макетами зданий и другими визуальными слоями данных, могли бы расширить возможности первых спасателей в понимании аварийных ситуаций, а искусственный ИНТЕЛЛЕКТ мог бы помочь строить планы реагирования, ускорять командование и управление и оптимизировать логистику для целого ряда сценариев реагирования на стихийные бедствия.

ИИ как двигатель прогресса.

Здравоохранение				
Продвижение профилактики заболеваний с помощью искусственного интеллекта. Помощь в раннем выявлении заболеваний и мониторинге. Смягчение последствий инвалидности. Поддержка медицинских работников в уходе за пациентами		Примеры: Предотвращение сердечно-сосудистых заболеваний и инсульта. Биологические сенсоры. Помощь слабовидящим. Роботы телемедицины		
Наука		ИИ как двигатель изобретения	Образование	
Ускорение открытия и эффективного использования материалов	Примеры: Ускорение разработки материалов и повышение их устойчивости		Оптимизация обучения для отдельных людей. Повышение безопасности при обучении	Примеры: Динамическое увеличение учебной программы
	Ускорение открытия лекарств и вакцин			Удалённые цифровые репетиторы, использующие виртуальную и дополненную реальность
Космос			Умные города	
Обеспечение расширенного и гибкого освоения космоса	Пример: Автономные космические аппараты и интеллектуальные среды обитания		Повышение безопасности	Примеры: Подключённые интеллектуальные датчики
		Создание интеллектуальной инфраструктуры	Умные дороги и мосты	
		Оптимизация сложных транспортных узлов	Координация движения людей и товаров	
		Поддержание окружающей	Поддержка низкоуглерод-	

			среды	ных энергетических систем

Примечания

1. Закон Джона С. Маккейна о санкционировании национальной обороны на 2019 финансовый год включает следующее определение для руководства работой Комиссии:

а. Любая искусственная система, которая выполняет задачи в различных и непредсказуемых обстоятельствах без значительного контроля со стороны человека или которая может учиться на опыте и улучшать производительность при воздействии наборов данных.

б. Искусственная система, разработанная в компьютерном программном обеспечении, физическом оборудовании или другом контексте, которая решает задачи, требующие человекоподобного восприятия, познания, планирования, обучения, общения или физических действий.

в. Искусственная система, созданная для того, чтобы думать или действовать как человек, включая когнитивные архитектуры и нейронные сети.

г. Набор методов, включая машинное обучение, который предназначен для аппроксимации когнитивной задачи.

д. Искусственная система, предназначенная для рациональных действий, включая интеллектуальный программный агент или воплощённый робот, который достигает целей с помощью восприятия, планирования, рассуждения, обучения, общения, принятия решений и действий. См. Pub. L. 115-232, 132 Stat. 1636, 1965 (2018).

2. Andrew W. Moore, et al., The AI Stack: A Blueprint for Developing and Deploying Artificial Intelligence, Proc. SPIE 10635, Ground/Air Multisensor Interoperability, Integration and Networking for Persistent ISR IX, 106350C (4 мая 2018 г.), <https://doi.org/10.1117/12.2309483>; См. Также Dave Martinez, et al., Artificial Intelligence: Short History, Present Developments, and Future Outlook, MIT Lincoln Laboratory at 27 (Jan. 2019), <https://www.ll.mit.edu/media/9526>.

3. Обратите внимание, что основанный на модели ИИ требует данных для ручного построения модели(моделей). Как правило, это требует меньше

данных, чем статистическое машинное обучение, но больше человеческих усилий.

4. Saleema Amershi, et al., Software Engineering for Machine Learning: A Case Study, ICSE-SEIP '19 Proceedings of the 41st International Conference on Software Engineering at 291-300 (2019), <https://2019.icse-conferences.org/details/icse-2019-Software-Engineering-in-ractice/30/SoftwareEngineering-for-Machine-Learning-A-Case-Study>; D. Sculley, et al., Machine Learning: The High Interest Credit Card of Technical Debt, SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop), <https://ai.google/research/pubs/pub43146>.

5. John Launchbury, A DARPA Perspective on Artificial Intelligence, DARPA, 4-7 (Feb. 2017), <https://www.darpa.mil/attachments/AIFull.pdf>.

6. Ограничения современного статистического машинного обучения, например, включают уязвимость неосознанного обучения и усиления предубеждений в обучающих данных; тот факт, что они часто являются сложными моделями, состоящими из очень большого числа изученных параметров, что делает их непрозрачными и трудными для интерпретации; тот факт, что они обучены решать узкие задачи и не имеют обобщения на другие связанные проблемы (например, когда оперативные данные принципиально изменяют характеристику обучающих данных) и тот факт, что они требуют больших объёмов помеченных обучающих данных.

7. Эндрю Мур, Когда ИИ становится повседневной технологией, Harvard Business Review (7 июня 2019 года), <https://hbr.org/2019/06/when-ai-becomes-an-everyday-technology>.

8. Tom B. Brown, et al., Language Models are Few-Short Learners, arXiv (22 июля 2020 г.), <https://arxiv.org/abs/2005.14165>.

9. Zachary Arnold, et al., Tracking AI Investment: Initial Findings from the Private Markets, Center for Security and Emerging Technology (сентябрь 2020), <https://cset.georgetown.edu/wp-content/uploads/CSET-Tracking-AI-Investment.pdf>.

10. Zachary Arnold, et al., Tracking AI Investment: Initial Findings from the Private Market По данным ЮНЕСКО, глобальные расходы на НИОКР достигли

рекордного уровня почти в 1,7 трлн долларов США. Посмотрите, Сколько Ваша страна инвестирует в НИОКР?, Институт статистики ЮНЕСКО (последний доступ 7 января 2021 года), <http://uis.unesco.org/apps/visualisations/research-and-developmentspending/>.

11. Недавнюю дискуссию экспертов по ИИ смотрите в разделе AI DEBATE 2: Moving AI Forward: An Interdisciplinary Approach, Montreal Artificial Intelligence (Dec. 23, 2020), <https://montrealartificialintelligence.com/aidebate2.html>.

Искусственный интеллект в контексте – Примечания

12. См., Например, Bryan Wilder и др., Learning to Complementate Humans, International Joint Conferences on Artificial Intelligence Organization (2020), <https://doi.org/10.24963/ijcai.2020/212>; Ece Kamar, et al., Combining Human and Machine Intelligence in Large-scale Crowdsourcing, Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012) (4-8 июня 2012 г.), <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/galaxyZoo.pdf>; Рамья Рамакришнан и др., Преодоление слепых зон в реальном мире: использование взаимодополняющих способностей для совместного исполнения, Материалы конференции AAAI по искусственному интеллекту (17 июля 2019 г.), <https://doi.org/10.1609/aaai.v33i01.33016137>; Saleema Amershi, et al., Guidelines for Human-AI Interaction, CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (May 2019), <https://doi.org/10.1145/3290605.3300233>; Эрик Хорвиц, Размышления о вызовах и обещаниях смешанного инициативного взаимодействия, журнал AI Magazine (15 июня 2007 г.), <https://doi.org/10.1609/aimag.v28i2.2036>.

13. Цель этих новых методов состоит в том, чтобы устранить необходимость во многих сложных вычислениях, которые делают традиционное обучение очень медленным. Vincent Dutordoir, Sparse Gaussian Processes with Spherical Harmonic Features, arXiv (30 июня 2020 г.), <https://arxiv.org/abs/2006.16649>.

14. Обсуждение архитектур гибридного интеллекта, сочетающих символические манипуляции с глубоким обучением, см. в книге Gary Marcus,

The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence, arXiv at 14-19 (Feb. 19, 2020), <https://arxiv.org/abs/2002.06177>.

15. См. Neuro-symbolic AI, MIT-IBM Watson AI Lab (последний доступ 16 января 2021 года), <https://mitibmwatsonailab.mit.edu/category/neuro-symbolic-ai/>.

16. Anuj Karpatne, et al., Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling, Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining (ACM SIGKDD) 2018 (Feb. 20, 2018), <https://arxiv.org/pdf/1710.11431.pdf>.

17. Rajat Raina, et al., Self-Teached Learning: Transfer Learning from Unlabeled Data, Proceedings of the 24th International Conference of Machine Learning (июнь 2007), <https://dl.acm.org/doi/abs/10.1145/1273496.1273592>; Dr. Bruce Draper, Learning with Less Labeling, DARPA (последний доступ 19 декабря 2020 года), <https://www.darpa.mil/program/learning-with-less-labeling>.

18. Рахул Диксит и др., Искусственный интеллект и машинное обучение в ситуациях разреженных/неточных данных, IEEE (21 августа 2020 г.), <https://ieeexplore.ieee.org/document/9172612>.

19. См. Cem Dilmegani, The Ultimate Guide to Synthetic Data in 2021, AI Multiple (12 января 2021), <https://research.aimultiple.com/synthetic-data/>.

20. Реальное обещание синтетических данных, MIT News (16 октября 2020 года), <https://news.mit.edu/2020/realpromise-synthetic-data-1016>.

21. 10 Breakthrough Technologies 2020, MIT Technology Review (26 февраля 2020 г.), <https://www.technologyreview.com/10-breakthrough-technologies/2020/#tiny-ai>.

22. Самая большая и полная в мире база знаний Здравоего смысла-Cycorp (последний доступ к ней состоялся 19 декабря 2020 года), <https://www.cyc.com/the-cyc-platform/the-knowledge-base>; Джон Павлус, Здравый смысл приближается к компьютерам, журнал Quanta (30 апреля 2020 года), <https://www.quantamagazine.org/common-sense-comes-to-computers-20200430/>;

Antoine Bosselut, et al., COMET: Commonsense Transformers for Automatic Knowledge Graph Construction, Proceedings of the 57th Annual Meeting of the

Association for Computational Linguistics (2019),
<https://homes.cs.washington.edu/~msap/pdf/bosselut2019comet.pdf>.

23. Эми Блюменталь, Как сделать ИИ заслуживающим доверия, Science Daily (31 августа 2020 года),
<https://www.sciencedaily.com/releases/2020/08/200827105937.htm>.

24. См. Benedict Neo, Top 4 AI companies leading in the race Towards Artificial General Intelligence, Towards Data Science (13 апреля 2020 года),
<https://towardsdatascience.com/four-ai-companies-on-the-bleeding-edge-of-artificial-general-intelligence-b17227a0b64a>; Srishti Deoras, 9 Компаний, Выполняющих Исключительную Работу В AGI, Как И OpenAI, Analytics India Magazine (25 Июля 2019 Года), <https://analyticsindiamag.com/9-companies-doing-exceptional-work-in-agi-just-like-openai/>.

25. В ключевых соображениях Комиссия описывает практику, технологии и оперативную политику разработки и внедрения систем, соответствующих ключевым ценностям. Важно отметить, что агентства должны рассматривать ценности как (1) воплощённые в выборе инженерных компромиссов и (2) явно представленные в целях и функциях полезности системы ИИ. См. Ключевые соображения по ответственному развитию и внедрению искусственного интеллекта, SCAI (июль 2020 г.), <https://www.nscai.gov/wp-content/uploads/2021/01/Key-Considerations-for-Responsible-Development-Fielding-of-AI.pdf>. В дополнение к действиям и инвестициям, необходимым сейчас (см. План действий, связанный с главой 7 настоящего доклада), Ключевые соображения включают политику и практику, которые должны быть обновлены, чтобы отразить новые соображения ИИ по мере развития технологии.

26. Это потребует НИОКР, как указано в Ключевых соображениях и главе 7 настоящего доклада. Это также потребует постоянных инвестиций в системную архитектуру, чтобы ограничить последствия сбоя системы, контролировать производительность ИИ по мере запуска систем, чтобы оценить, работают ли они так, как задумано, и преодолеть Научно-технические пробелы в области

аудита и надзора. Дополнительные сведения см. в главах 7 и 8 настоящего доклада.

27. Как отмечалось в главе 8 и её Плане действий, это потребует, например, постоянной оценки и постоянного руководства по последствиям применения ИИ и новых технологий для защиты частной жизни и гражданских свобод. Запрещённые результаты и руководящие принципы политики должны будут обновляться с течением времени по мере изменения общинных норм и технических возможностей. Кроме того, ключевые соображения указывают на то, что инженерные системы должны будут оценить общую осуществимость и соответствие запрещённым результатам, выраженным в политике, а также продемонстрированную техническую зрелость конкретных технологий ИИ-кандидатов. См. Ключевые соображения по ответственному развитию и внедрению искусственного интеллекта, SCAI (июль 2020 г.), <https://www.nscai.gov/wp-content/uploads/2021/01/Key-Considerations-for-ResponsibleDevelopment-Fielding-of-AI.pdf>.

28. Предполагаемые системы перевода будут использовать обратную связь с системой, активно исправляя ошибки перевода и распознавания, которые делает программное обеспечение, повышая производительность по мере продолжения взаимодействия между переводящими сторонами. Одной из основных целей является быстрое внедрение новых языков, которые ранее не были замечены. Например, проект Carnegie Mellon DIPLOMAT делает возможным интерактивный перевод речи с помощью новой архитектуры под названием Multi Engine Machine Translation (MEMT). DIPLOMAT даёт пользователям возможность вносить коррективы в перевод для поддержки быстрого развития новых языков, которые ранее не были замечены. Роберт Фредерк, Интерактивный перевод речи в проекте DIPLOMAT, Институт языковых технологий Университета Карнеги-Меллона (последний доступ 19 декабря 2020 года), <http://www.cs.cmu.edu/~air/papers/acl97-workshop.pdf>.

29. Моделирование и имитация также могут помочь подготовиться к эффективным ответным мерам в рамках пандемической цепочки поставок,

организованным правительством. Madhav Marathe, High Performance Simulations to Support Real-time COVID19 Response, SIGSIM-PADS '20 (июнь 2020), <https://dl.acm.org/doi/pdf/10.1145/3384441.3395993>.

30. Edgybees (последний доступ 19 декабря 2020 года), <https://edgybees.com/>.

31. Министерство энергетики объявляет о создании Первого консорциума из пяти компаний Министерства энергетики США (18 августа 2020 года). <https://www.energy.gov/articles/departments-energy-announces-first-five-consortium>.