ИИ в гуманитарной деятельности: права человека и этика

Майкл Пицци, Мила Романофф и Тим Энгельхардт*

Майкл Пицци — научный сотрудник в рамках инициативы Генерального секретаря ООН «Глобальный пульс» и научный сотрудник по цифровой этике в Институте Джейн Фэмили.

Мила Романофф — специалист по конфиденциальности, руководитель по вопросам управления данными и политики в рамках инициативы Генерального секретаря ООН «Глобальный пульс».

Тим Энгельхардт — специалист по правам человека при Управлении Верховного комиссара ООН по правам человека.

Аннотация

Применение систем, основанных на искусственном интеллекте (ИИ), может трансформировать деятельность гуманитарного сектора, но также является источником исключительного риска с точки зрения прав человека, даже при использовании с самыми благими намерениями. Исходя из результатов исследований и экспертных консультаций, проведенных в разных странах мира в последние годы, авторы работы выделяют основные консенсуальные тезисы относительно того, как гуманитарные организации могут заставить ИИ работать в общечеловеческих интересах — а не против них, — обеспечивая при этом соблюдение прав человека. В частности, в ходе таких консультаций подчеркивалась необходимость создания референсного механизма, основанного на международном праве прав человека и призванного служить фундаментом

* В настоящей статье выражена точка зрения авторов, которая может не совпадать с точкой зрения Организации Объединенных Наций.

для обеспечения комплексного учета общечеловеческих интересов в системах ИИ. Кроме того, вспомогательную роль в устранении пробелов и поднятии стандартов выше минимальных требований международного права прав человека способны играть этические принципы. В настоящей работе обобщаются преимущества подобного механизма по обеспечению прав человека, а также определяются конкретные инструменты и передовые виды практики, которые либо уже существуют и могут быть адаптированы к сфере использования ИИ, либо пока не созданы, но требуются для эффективной работы механизма. По мере того как становятся понятны истинные масштабы кризиса, связанного с пандемией COVID-19, ИИ все в большей степени определяет форму мер реагирования, принимаемых для решения наиболее острых глобальных проблем, особенно в области развития и в гуманитарном секторе. Для того чтобы гарантировать положительное воздействие инструментов ИИ на общечеловеческий прогресс и их вклад в достижение целей в области устойчивого развития, гуманитарные организации должны действовать на упреждение, учитывая все аспекты разработки инструментов, стратегий и механизмов подотчетности, которые призваны защищать права человека.

Ключевые слова: искусственный интеллект, этика ИИ, машинное обучение, права человека, гуманитаризм, гуманитарные организации

Введение

Свирепствующая в настоящее время пандемия COVID-19 наносит огромный ущерб во многих областях. Однако, как недавно отметил Генеральный секретарь Организации Объединенных Наций (ООН), она дала нам возможность узнать, какими должны быть глобальные меры реагирования на будущие кризисы. По его словам, мир «непосредственно может наблюдать, как цифровые технологии помогают противостоять этой угрозе и поддерживать связь между людьми»[1]. Искусственный интеллект (ИИ) играет ключевую роль во многих мерах вмешательства, принимаемых сегодня на основании полученных данных. В последние месяцы правительства и международные организации максимально использовали потенциал систем ИИ в области прогнозирования, адаптации и масштабирования в целях создания моделей, позволяющих спрогнозировать распростране-

1 Генеральная Ассамблея ООН. *Дорожная карта по цифровому сотрудничеству: осуществление рекомендаций Группы высокого уровня по цифровому сотрудничеству. Доклад Генерального секретаря*, док. ООН A/74/821, 29 мая 2020 г. (Дорожная карта Генерального секретаря), п. 6, доступно по адресу: https://undocs.org/ru/A/74/821 (все ссылки на интернет-ресурсы приводятся по состоянию на декабрь 2020 г.).

ние вируса, и даже в целях стимуляции исследований на молекулярном уровне[2]. В широком спектре задач — от отслеживания контактов и других форм эпиднадзора за пандемией до клинических и молекулярных исследований — ИИ и другие меры вмешательства на основании данных показали себя как главные средства замедления распространения заболевания, успешной реализации срочных медицинских исследований и информирования мировой общественности.

Цель настоящей работы — выяснить, как механизм управления, опирающийся на систему защиты прав человека и учитывающий этические принципы, может обеспечить использование ИИ в ходе гуманитарных и миротворческих операций и деятельности в целях развития без нарушения прав человека. В работе подробно рассматривается использование ИИ для достижения целей ООН в области устойчивого развития и для решения других гуманитарных задач. Одной из основных тем работы также являются ущерб или риск, которые могут *случайно* или *неизбежно* возникать при законном, а не злонамеренном использовании ИИ (способы такого использования могут быть весьма разнообразны).

Как отмечает Генеральный секретарь, ИИ уже «повсеместно применяется в различных сферах»[3] и существующий сегодня по всему миру интерес к ИИ будет способствовать дальнейшему расширению области его применения[4]. По мере того как становятся понятны истинные масштабы кризиса, связанного с пандемией COVID-19, ИИ все в большей степени определяет форму мер реагирования, принимаемых для решения наиболее острых глобальных проблем, особенно в областях развития и гуманитарной помощи. Тем не менее при отсутствии контроля активное распространение ИИ обусловливает появление серьезных видов риска в области защиты прав человека. Такие виды риска отличаются сложной природой и многоуровневой структурой и проявляются в основном в узкоспециальных условиях. Однако ряд подобных угроз может возникать в любых секторах и вне зависимости от географии.

Так, эти системы могут быть очень мощными, а по своим аналитическим и предиктивным возможностям они всё сильнее опережают человека. Поэтому их обязательно будут использовать для принятия решений вместо людей, особенно в тех случаях, когда необходимо провести быстрый или масштабный анализ, а оператор-человек часто не замечает риска и потен-

2 См., например, инициативы, подробно описанные в двух недавно опубликованных работах о способах применения ИИ и машинного обучения (МО) при реагировании на COVID-19: Miguel Luengo-Oroz *et al.*, "Artificial Intelligence Cooperation to Support the Global Response to COVID-19", *Nature Machine Intelligence*, Vol. 2, No. 6, 2020; Joseph Bullock *et al.*, "Mapping the Landscape of Artificial Intelligence Applications against COVID-19", *Journal of Artificial Intelligence Research*, Vol. 69, 2020, доступно по адресу: www.jair.org/index.php/jair/article/view/12162.

3 Дорожная карта Генерального секретаря (примечание 1 выше), п. 53.

4 Использование ИИ «может принести к 2022 году около 4 трлн долл. США добавленной стоимости на глобальных рынках, согласно прогнозам, сделанным еще до вспышки пандемии COVID-19, которая, по мнению экспертов, может изменить потребительские предпочтения и открыть новые возможности для автоматизации на основе искусственного интеллекта в промышленности, бизнесе и обществе». Там же, п. 53.

циала нанесения серьезного ущерба отдельным лицам или группам лиц, которые уже находятся в уязвимом положении[5]. Кроме того, искусственный интеллект осложняет обеспечение прозрачности и надзор, поскольку разработчики и операторы часто не могут «заглянуть внутрь» систем ИИ и понять, как и почему те принимают то или иное решение. Эта так называемая проблема «черного ящика» может стать препятствием для эффективной подотчетности в случаях, когда такие системы наносят вред, например когда система ИИ принимает или поддерживает решение, имеющее дискриминационные последствия[6].

Некоторые из видов риска и ущерба, обусловленных ИИ, уже регулируются другими отраслями и сводами права, такими как правовые нормы о конфиденциальности и защите данных[7], но многие являются абсолютно новыми. Этика ИИ, или управление ИИ, — это формирующаяся в настоящее время отрасль, задача которой состоит в борьбе с новыми видами риска, создаваемыми подобными системами. Сегодня основным инструментом в этой отрасли являются «этические кодексы» ИИ, цель которых — определять архитектуру и развертывание систем ИИ. В течение нескольких последних лет десятки организаций, в том числе международные организации, правительства стран, частные корпорации и неправительственные организации (НПО), опубликовали собственные наборы принципов, которых, по их мнению, следует придерживаться для ответственного использования ИИ либо в границах соответствующей организации, либо в более широком масштабе[8].

При том что подобные усилия часто заслуживают похвалы, этические кодексы имеют ограничения в ключевых аспектах: они не объединены какой-либо согласованной на глобальном уровне основой; у них, в отличие от правовых норм, нет юридической силы, и потому их редко соблюдает широкий круг лиц; в них часто отражены ценности, присущие лишь конкретной разработавшей их организации, а не широкому спектру лиц и сто-

5 Lorna McGregor, Daragh Murray and Vivian Ng, "International Human Rights Law as a Framework for Algorithmic Accountability", *International and Comparative Law Quarterly*, Vol. 68, No. 2, 2019, доступно по адресу: https://tinyurl.com/yaflu6ku.

6 См., например: Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation", *Harvard Journal of Law and Technology*, Vol. 31, No. 2, 2018; Rachel Adams and Nora Ni Loideain, "Addressing Indirect Discrimination and Gender Stereotypes in AI Virtual Personal Assistants: The Role of International Human Rights Law", 19 June 2019 (работа, представленная в 2019 г. на ежегодной Кембриджской международной конференции по международному праву «Новые технологии: новые вызовы для демократии и международного права», доступно по адресу: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3392243).

7 См., например: Global Privacy Assembly, "Declaration on Ethics and Data Protection in Artificial Intelligence", Brussels, 23 October 2018, доступно по адресу: http://globalprivacyassembly.org/wp-content/uploads/2019/04/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf; UN Global Pulse and International Association of Privacy Professionals, *Building Ethics into Privacy Frameworks for Big Data and AI*, 2018, доступно по адресу: https://iapp.org/resources/article/building-ethics-into-privacy-frameworks-for-big-data-and-ai/.

8 Обзор см.: Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy and Madhulika Srikumar, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, Berkman Klein Center Research Publication No. 2020-1, 14 February 2020.

рон, на которых системы ИИ могут потенциально оказывать воздействие; также они совсем не обязательно используются разработчиками и пользователями инструментов ИИ на повседневной основе. Кроме того, создатели таких принципов часто дают недостаточно инструкций в отношении того, как следует разрешать конфликт или несоответствие между разными принципами (например, в ситуациях, когда тщательное соблюдение одного принципа приводит к нарушению другого), из-за чего их становится еще сложнее применять на практике. В силу того, что создателями и операторами большинства продуктов, в которых используется ИИ, являются технологические компании, подобная модель управления опирается в основном на внутреннее корпоративное регулирование — что не может не беспокоить, ведь при принятии корпоративных решений не обеспечиваются демократическая представленность всех сторон и подотчетность.

Внедрение и практическое использование таких принципов в деятельности в целях развития и в оказании гуманитарной помощи связано с рядом дополнительных сложностей. За исключением нескольких высококачественных аналитических докладов, посвященных этике ИИ и гуманитарной деятельности, объем руководящих указаний для практических специалистов в этой быстро развивающейся сфере остается скудным[9]. Такая ситуация наблюдается на фоне того, что проектам в области развития или в гуманитарной сфере присущ ряд факторов, которые либо усугубляют традиционные этические проблемы, связанные с ИИ, либо обусловливают совершенно новые.

Надлежащее управление ИИ быстро выходит в разряд глобальных приоритетов. Как четко и неоднократно обозначено в Дорожной карте Генерального секретаря по цифровому сотрудничеству, глобальный подход к ИИ — как во время пандемии COVID-19, так и после нее — должен обеспечивать полное соблюдение прав человека[10]. ООН и другие международные организации уделяют все больше внимания этому вопросу, исходя как из растущего спроса на ИИ и другие решения, основанные на данных, для выполнения глобальных задач, в том числе достижения ЦУР, так и из этических угроз, которые влечет за собой использование этих решений.

9 См.: Faine Greenwood, Caitlin Howarth, Danielle Escudero Poole, Nathaniel A. Raymond and Daniel P. Scarnecchia, *The Signal Code: A Human Rights Approach to Information During Crisis*, Harvard Humanitarian Initiative, 2017, p. 4, где подчеркивается важность основанных на правах человека руководящих указаний для гуманитарных организаций, работающих с большими данными. Тем не менее уже действует ряд механизмов, наиболее примечательным из которых является созданная в апреле 2020 г. Группой по науке о данных и этике данных в гуманитарной деятельности (DSEG) Рамочная основа этичного использования передовых методов изучения данных в гуманитарном секторе (“*Framework for the Ethical Use of Advanced Data Science Methods in the Humanitarian Sector*”), доступно по адресу: https://tinyurl.com/yazcao2o. Предпринимались также попытки подготовить указания для практикующих специалистов в области гуманитарного права, так как оно применяется к автономным системам оружия летального действия, в том числе подготовленный в Институте Ассера проект «Интеграция международного права и этических норм в военные системы ИИ» (DILEMA), доступно по адресу: www.asser.nl/research/human-dignity-and-human-security/designing-international-law-and-ethics-into-military-ai-dilema.

10 Дорожная карта Генерального секретаря (примечание 1 выше), п. 50.

В 2019 году Генеральная Ассамблея[11] и Совет по правам человека ООН (СПЧ ООН)[12] приняли резолюции, призывающие применять международное право прав человека к ИИ и другим новым цифровым технологиям. Резолюция Генеральной Ассамблеи содержала предупреждение о том, что «технологии профилирования, автоматического принятия решений и машинного обучения... в отсутствие надлежащих гарантий могут привести к решениям, которые могут повлиять на осуществление прав человека»[13].

Эту задачу необходимо решить оперативно, ведь пока мы пытаемся договориться о том, как применять принципы и механизмы защиты прав человека к ИИ, цифровые технологии продолжают стремительно развиваться. ИИ получает все более широкое применение в международном общественном секторе, а значит, в этой сфере постоянно возникают новые угрозы. Пандемия COVID-19 своевременно напомнила нам об этом. Для того чтобы гарантировать положительное воздействие инструментов ИИ на общечеловеческий прогресс и их вклад в достижение ЦУР, следует действовать на упреждение, учитывая все аспекты разработки инструментов, стратегий и механизмов подотчетности, которые призваны защищать права человека.

Выводы, содержащиеся в настоящей статье, основаны на качественных данных, которые получены в ходе многосторонних консультаций, организованных инициативой Генерального секретаря ООН «Глобальный пульс» или проведенных ею совместно с другими организациями, в обязанности которых входит защита права на неприкосновенность частной жизни и других прав человека, в том числе совместно с Управлением Верховного комиссара ООН по правам человека (Управление ООН по правам человека) и национальными органами по защите данных[14]. Данные для статьи также были собраны в ходе многочисленных бесед и совещаний с большим числом разнопрофильных специалистов в области ИИ и данных, входящих в состав экспертной группы по управлению данными и ИИ инициативы «Глобальный пульс»[15]. К другим источникам качественных данных относятся руководящие указания и доклады экспертов ООН по правам человека; результаты научной работы в области защиты прав человека и этики, а также практические руководства для организаций, осуществляющих дея-

11 Резолюция ГА ООН 73/179, 2018 г.

12 Резолюция СПЧ 42/15, 2019 г.

13 Резолюция ГА ООН 73/179, 2018 г.

14 Консультации были организованы в том числе в форме практических семинаров по разработке рамочных основ этичного использования ИИ в Гане и Уганде, по вопросам ИИ и конфиденциальности в странах Глобального Юга на саммите RightsCon в Тунисе, по подходам к ИИ, основанным на защите прав человека, в Женеве совместно с Управлением ООН по правам человека, а также в форме ряда мероприятий на Форуме по вопросам регулирования интернета в Берлине и в ходе совещания по вопросам этики в контексте развития и гуманитарной деятельности совместно с Международной ассоциацией специалистов в области конфиденциальности и Европейским надзорным органом по защите данных. В упомянутых консультациях, проходивших в период с 2018 по 2020 г., принимали участие эксперты из правительственных органов, международных организаций, организаций гражданского общества, а также представители частного сектора из разных стран мира.

15 См. веб-страницу экспертной группы по управлению данными и ИИ инициативы «Глобальный пульс», доступно по адресу: www. unglobalpulse.org/policy/data-privacy-advisory-group/.

тельность в целях развития, и гуманитарных организаций, которые были опубликованы такими организациями, как Всемирная организация здравоохранения, Управление ООН по координации гуманитарных вопросов (УКГВ)[16], Международный Комитет Красного Креста (МККК)[17], Гарвардская гуманитарная инициатива (HHI)[18], Access Now[19], Article 19[20], Центр по цифровому развитию Агентства США по международному развитию[21] и Группа по науке о данных и этике данных в гуманитарной деятельности (DSEG)[22].

ИИ в гуманитарной деятельности: возможности

Термин «искусственный интеллект» не обозначает какую-либо конкретную технологию. Это более широкое понятие, охватывающее спектр инструментов или возможностей, которые созданы для симуляции тех или иных способностей человеческого интеллекта. Под ИИ как категорией обычно имеется в виду система, которая автоматизирует аналитический процесс, например идентификацию или классификацию данных; порой системы ИИ способны даже автоматизировать принятие решений. В этой связи некоторые отдают предпочтение термину «система автоматического интеллекта», а не более распространенному обозначению «искусственный интеллект» или «ИИ». Для целей настоящей работы «ИИ» обозначает в первую очередь алгоритмы машинного обучения (МО), которые повсеместно присутствуют в системах ИИ и отличаются способностью выявлять закономерности, делать заключения исходя из этих закономерностей и применять их в совершенно других ситуациях[23]. Модели МО могут быть контролируемыми, то есть требующими от человека загрузки в них набора применимых правил, или неконтролируемыми, то есть способными самостоятельно выводить правила непосредственно из данных и не нуждающимися в человеке

16 См.: OCHA, *Data Responsibility Guidelines: Working Draft*, March 2019, доступно по адресу: https://tinyurl.com/y64pcew7.
17 ICRC, *Handbook on Data Protection in Humanitarian Action*, Geneva, 2017.
18 F. Greenwood *et al.* (примечание 9 выше).
19 Access Now, *Human Rights in the Age of Artificial Intelligence*, 2018, доступно по адресу: www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf.
20 Article 19, *Governance with Teeth: How Human Rights can Strengthen FAT and Ethics Initiatives on Artificial Intelligence*, April 2019, доступно по адресу: www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf.
21 USAID Center for Digital Development, *Reflecting the Past, Shaping the Future: Making AI Work for International Development*, 2018.
22 DSEG (примечание 9 выше).
23 Jack M. Balkin, "2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data", Ohio State Law Journal, Vol. 78, No. 5, 2017, p. 1219 (цитируется в L. McGregor, D. Murray and V. Ng (примечание 5 выше), p. 310). См. также определение искусственного интеллекта, данное Европейским союзом: «Искусственный интеллект — это системы, которые демонстрируют интеллектуальные способности, анализируя окружающий мир и принимая решения — с некоторой степенью самостоятельности — для достижения конкретных целей». European Commission, "A Definition of Artificial Intelligence: Main Capabilities and Scientific Disciplines", 8 April 2019, доступно по адресу: https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines.

для загрузки кодированных правил. По этой причине последний тип моделей часто называют самообучающимися[24]. Глубокое обучение (ГО), в свою очередь, — более мощная разновидность моделей МО, которая использует многослойные искусственные нейросети (элементы которых структурированы по аналогии с нейронами головного мозга человека), чтобы выявлять закономерности и делать прогнозы[25].

Алгоритмические системы могут «выполнять сложные задачи, которые человек либо не способен решить вовсе, либо способен, но с большой потерей скорости; самообучаться для повышения эффективности и проводить сложные аналитические операции для прогнозирования вероятных результатов в будущем»[26]. Сегодня подобные системы выполняют многочисленные функции, среди которых обработка естественного языка, компьютерное зрение, распознавание речи и аудио, предиктивный анализ и выполнение передовых робототехнических операций[27]. Эти и другие методы уже внедряются в деятельность в целях развития и гуманитарную деятельность, что дает возможность использовать инновационные решения во благо. Компьютерное зрение применяется для автоматической идентификации строений на спутниковых снимках, что позволяет оперативно отслеживать миграционные потоки и способствует эффективному распределению помощи при гуманитарных кризисах[28]. Множество проектов, реализуемых в развивающихся странах, используют ИИ для предоставления прогнозных аналитических сведений фермерам, позволяя последним смягчить последствия засухи и других неблагоприятных погодных явлений и получать максимально большой урожай благодаря посеву в оптимальное время[29]. Новаторские инструменты ИИ обеспечивают удаленную диагностику патологических состояний, таких как неполноценное питание, в регионах, где наблюдается нехватка медицинских ресурсов[30]. Перечень способов применения ИИ в гуманитарной сфере растет с каждым днем[31].

24 См.: “Common ML Problems” in Google’s Introduction to Machine Learning Problem Framing course, доступно по адресу: https://developers.google.com/machine-learning/problem-framing/cases.

25 Tao Liu, “An Overview of the Application of AI in Development Practice”, Berkeley MDP, доступно по адресу: https://mdp.berkeley.edu/an-overview-of-the-application-of-ai-in-development-practice/.

26 L. McGregor, D. Murray and V. Ng (примечание 5 выше), p. 310.

27 Проработанные определения каждого из этих терминов см.: Access Now (примечание 19 выше), p. 8.

28 См. веб-сайт проекта PulseSatellite инициативы Генерального секретаря ООН «Глобальный пульс», доступно по адресу: www.unglobalpulse.org/microsite/pulsesatellite/.

29 Примерами могут служить AtlasAI, EzyAgric, Apollo, FarmForce, Tulaa и Fraym.

30 См., например, разработанный компанией Kimetrica инструмент «Методы чрезвычайно быстрого определения ситуации с питанием человека» (MERON) — проект, осуществляемый совместно с ЮНИСЕФ и использующий технологии распознавания лиц для удаленного диагностирования недоедания у детей.

31 Больше примеров применения ИИ в проектах гуманитарного сектора см.: International Telecommunications Union, *United Nations Activities on Artificial Intelligence (AI)*, 2019, доступно по адресу: www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2019-1-PDF-E.pdf; в перечне принятых работ Семинара по роли искусственного интеллекта в гуманитарной поддержке и реагировании на катастрофы (“Artificial Intelligence for Humanitarian Assistance and Disaster Response Workshop”), доступно по адресу: https://www.hadr.ai/previous-years/2020/accepted-papers; а также в перечне проектов: DSEG (примечание 9 выше), Chapter 3.

Расширение применения ИИ в этих и других секторах объясняется рядом факторов. Наиболее важным стимулом при этом является революция данных, в ходе которой происходит экспоненциальное расширение наборов данных, имеющих значение для развития и гуманитарной деятельности[32]. Данные — это топливо, питающее развитие ИИ; для обучения любой модели ИИ нужны наборы релевантных данных. Поиск источников качественных данных традиционно более сложен в странах с развивающейся экономикой, особенно в наименее развитых странах[33], а также в условиях гуманитарных кризисов, когда технологическая инфраструктура, ресурсы и экспертные знания зачастую имеются в очень ограниченном объеме. Однако, согласно недавно опубликованному комплексному аналитическому докладу Группы по науке о данных и этике данных в гуманитарной деятельности, эта ситуация начинает меняться:

> Сегодня по всему миру ведется сбор беспрецедентного количества данных; «гуманитарные» данные генерируются гораздо более широким кругом заинтересованных сторон, нежели ранее; данные становятся более пригодными для машинного считывания и более доступными за счет их размещения на онлайн-порталах. Это создает в секторе среду, благоприятную для инноваций и прогресса, и способствует повышению прозрачности, более обоснованному принятию решений и эффективному оказанию гуманитарных услуг[34].

Основные препятствия для создания ИИ, соблюдающего права человека

Характеристики, превращающие системы ИИ в столь мощные инструменты, одновременно обусловливают и угрозы для прав и свобод тех людей, на которых воздействуют результаты работы таких систем. Именно так зачастую происходит с новыми цифровыми технологиями, поэтому важно точно определить, что именно в ИИ является «новым» или уникальным — и, соответственно, почему ИИ требует особого внимания. Подробный технический анализ инновационных отличительных черт ИИ не входит в цели настоящей работы, но в последующих абзацах обобщены некоторые наиболее часто называемые проблемы, связанные с интеграцией принципов защиты прав человека в функционирование систем ИИ.

Отсутствие прозрачности и объяснимости

Системы ИИ часто малопонятны для лиц, ответственных за принятие решений; это обстоятельство также известно как «проблема черного ящи-

32 UN Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development, *A World That Counts: Mobilising the Data Revolution for Sustainable Development,* 2014.
33 См.: UN Department of Economic and Social Affairs, "Least Developed Countries", доступно по адресу: www.un.org/development/desa/dpad/least-developed-country-category.html.
34 DSEG (примечание 9 выше), p. 3.

ка»[35]. В отличие от обычных алгоритмов, решения, принимаемые в процессе работы моделей МО или ГО, иногда невозможно отследить, а следовательно, и проверить или иным образом объяснить их общественности и лицам, ответственным за мониторинг использования таких моделей (это также называется принципом объяснимости)[36]. Это означает, что системы ИИ могут быть также малопонятны для тех, на кого воздействуют результаты их работы, что ведет к проблемам обеспечения подотчетности в ситуациях, когда системы причиняют ущерб. Низкий уровень прозрачности в работе систем ИИ может лишить людей возможности понять, что их права были нарушены (и почему это произошло), и, следовательно, возможности добиться возмещения ущерба, который повлекли за собой такие нарушения. Кроме того, даже когда возможность понять принципы работы такой системы присутствует, для этого может потребоваться высокий уровень специальных технических знаний, которым обычные люди не обладают[37]. Из-за этого желание получить компенсацию ущерба, причиненного системами ИИ, может остаться неудовлетворенным.

Подотчетность

Отсутствие прозрачности и объяснимости может серьезнейшим образом препятствовать эффективному привлечению к ответственности за ущерб, причиненный автоматически принятыми решениями, как на руководящем, так и на операционном уровнях. Проблема имеет два измерения. Во-первых, люди часто не знают, как и когда ИИ применяется для определения их прав[38]. Как предупреждал бывший Специальный докладчик ООН по вопросу о поощрении и защите права на свободу мнений и их свободное выражение Дэвид Кей, люди редко отдают себе отчет «о масштабах, сфере охвата или даже существовании алгоритмических процессов принятия решений, которые могут оказывать влияние на осуществление их прав на свободу мнений и их свободное выражение». Таким образом, системы ИИ «в принципе не предоставляют» уведомления отдельным лицам об использовании ИИ»[39]. Это особенно актуально в гуманитарных контекстах, когда затронутые лица часто не могут дать осмысленное согласие на сбор

35 Cynthia Rudin and Joanna Radin. "Why Are We Using Black Box Models in AI When We Don't Need To?", *Harvard Data Science Review*, Vol. 1, No. 2, 2019, доступно по адресу: https://doi.org/10.1162/99608f92.5a8a3a3d.

36 См.: Miriam C. Buiten, "Towards Intelligent Regulation of Artificial Intelligence", *European Journal of Risk Regulation*, Vol. 10, No. 1, 2019, доступно по адресу: https://tinyurl.com/y8wqmp9a; Anna Jobin, Marcello Ienca, Effy Vayena, "The Global Landscape of AI Ethics Guidelines", *Nature Machine Intelligence*, Vol. 1, No. 9, 2019, доступно по адресу: www.nature.com/articles/s42256-019-0088-2.pdf.

37 См., например: L. McGregor, D. Murray and V. Ng (примечание 5 выше), p. 319, где объясняется ряд видов риска, вызываемых отсутствием прозрачности и объяснимости: «так как процесс обучения алгоритма не повторяет человеческую логику, это создает проблемы понимания и объяснения соответствующего процесса».

38 *Дэвид Кей*. Доклад Специального докладчика по вопросу о поощрении и защите права на свободу мнений и их свободное выражение, док. ООН A/73/348, 29 августа 2018 г., п. 40.

39 Там же, относительно применения ИИ в информационной онлайн-среде.

и анализ данных (например, когда согласие нужно для получения жизненно важных услуг)[40].

Во-вторых, малопонятность экономики данных и отсутствие механизмов обеспечения ответственности за нарушения прав человека[41] могут затруднить получение людьми информации об ущемлении их прав и поиск способов возмещения причиненного в результате этого ущерба. В связи с этим такая задача может быть затруднена даже для сведущих экспертов или правоприменителей, которые занимаются аудитом этих систем и диагностикой их дефектов. Сложная организационная структура, характерная для большинства проектов в области развития и гуманитарной деятельности, может усугублять такие проблемы[42]. Когда в одном проекте задействована длинная цепочка действующих субъектов (включая организации, предоставляющие финансирование, правительства иностранных государств, международные организации, подрядчиков, частных поставщиков, местные органы власти, партнерские организации гражданского общества и организации, ответственные за сбор данных), кто в конечном счете должен нести ответственность в случае, если система внезапно принимает решение, оказывающее избирательное воздействие (или проводит анализ, результаты которого заставляют изменить ранее принятое решение)?

Непредсказуемость

Отличительной чертой алгоритмов МО и ГО является способность обучаться и развиваться непредсказуемым образом. Другими словами, они могут «постепенно выявлять новые проблемы и разрабатывать новые решения. В зависимости от уровня контроля системы могут выявлять закономерности и делать выводы, которые не могут быть получены людьми, писавшими программы и формулировавшими задачи»[43]. В этом и заключается их важнейшая ценность — алгоритмы МО в некоторых случаях способны анализировать данные, которые их, возможно, не учили анализировать, для решения новых задач или даже действий в совершенно новых условиях. В то же время люди не всегда способны проследить логику функциональных решений такой системы или даже понять их. Поэтому разработчикам и операторам таких систем бывает сложно спрогнозировать — и тем более объяснить — природу и уровень риска, который система или способ ее применения создает в конкретных условиях. Кроме того, адаптивная способность даже самых мощных систем МО имеет свой предел. Многие из них *малоспособны* к логическому обобщению в новых условиях, из-за чего

40 DSEG (примечание 9 выше), p. 7.
41 Isabel Ebert, Thorsten Busch, Florian Wettstein, *Business and Human Rights in the Data Economy: A Mapping and Research Study,* German Institute for Human Rights, Berlin, 2020.
42 Lindsey Andersen, "Artificial Intelligence in International Development: Avoiding Ethical Pitfalls", *Journal of Public and International Affair*s, 2019, доступно по адресу: https://jpia.princeton.edu/news/artificial-intelligence-international-development-avoiding-ethical-pitfalls.
43 Д. Кей (примечание 38 выше), п. 8.

их действия при работе с данными, значительно отличающимися от тех, на которых они были обучены, бывает очень сложно предсказать.

Постепенная утрата конфиденциальности

Способность систем ИИ анализировать и делать заключения на основе огромных объемов конфиденциальных или общедоступных данных может стать причиной серьезных нарушений многих охраняемых аспектов права на неприкосновенность частной жизни. Системы ИИ могут выявлять конфиденциальные данные о местонахождении людей, их политических воззрениях, сексуальных предпочтениях и так далее на основе сведений, которые люди добровольно размещают в интернете (например, тексты и фотографии в социальных сетях) или время от времени фиксируют на своих цифровых устройствах (например, данные спутниковой геолокации или местоположение относительно вышек сотовой связи)[44]. Риск такого развития событий особенно велик в гуманитарных контекстах, когда затронутые системой ИИ люди часто относятся к наиболее ущемленным в правах группам. В результате данные или аналитические сведения, которые в обычных обстоятельствах не являются конфиденциальными, могут становиться таковыми. Так, основная идентификационная информация — фамилия и имя, город и адрес проживания — в большинстве ситуаций может находиться в открытом доступе, но если говорить о беженце, который пытается скрыться от репрессий или преследования у себя на родине, то такая информация, оказавшись в руках недобросовестных лиц, создает угрозу безопасности и жизни этого человека[45]. Кроме того, модели МО, работающие с большими данными, могут стимулировать дальнейший сбор данных, в связи с чем возникают все более серьезные нарушения конфиденциальности и риск деанонимизации. Ко всему прочему, использование ИИ для анализа массивных объемов личных данных также связывают с нарушением и других прав, в том числе права на свободу мнений и их свободное выражение, права на свободу объединения и мирных собраний, а также права на эффективные средства правовой защиты[46].

44 См.: *СПЧ*. Вопрос о реализации во всех странах экономических, социальных и культурных прав: роль новых технологий в реализации экономических, социальных и культурных прав. Доклад Генерального секретаря, док. ООН A/HRC/43/29, 4 марта 2020 г. (доклад по вопросу о реализации во всех странах экономических, социальных и культурных прав), с. 10. См. также: Ana Beduschi, "Research Brief: Human Rights and the Governance of AI", Geneva Academy, February 2020, p. 3: «Из-за все большего усложнения методов, с помощью которых онлайн-платформы и компании отслеживают онлайновое поведение и цифровой след людей, алгоритмы ИИ могут делать заключения о поведении людей, в том числе по поводу их политических и религиозных предпочтений, состояния здоровья и сексуальной ориентации».

45 Это отчасти объясняет негативное отношение к технологиям распознавания лиц и другим способам биометрической идентификации. См., например: The Engine Room and Oxfam, *Biometrics in the Humanitarian Sector*, March 2018; Mark Latonero, "Stop Surveillance Humanitarianism", *New York Times*, 11 July 2019; Dragana Kaurin, *Data Protection and Digital Agency for Refugees*, World Refugee Council Research Paper No. 12, May 2019.

46 Доклад по вопросу о реализации во всех странах экономических, социальных и культурных прав (примечание 44 выше), с. 10.

Неравенство, дискриминация и предубеждения

Если данные, на которых обучается модель ИИ, являются неполными, содержат предубеждения или иным образом не соответствуют требованиям, система может генерировать решения, оказывающие избирательное воздействие, или несправедливые решения и результаты[47]. Предубеждения и другие дефекты могут быть внесены в систему на нескольких этапах: при первичном определении границ проблемы (например, при выборе вспомогательной переменной, которая привязана к тем или иным социоэкономическим характеристикам или расовой принадлежности); при сборе данных (например, если ущемленная в правах группа недостаточно представлена в данных, на которых система обучается), а также при подготовке данных[48]. В некоторых случаях предубеждения, существующие в мышлении разработчиков, неосознанно переносятся в код модели. Известен ряд случаев, получивших широкий резонанс, когда в работе систем МО проявлялись расовые или гендерные предубеждения, например инструмент МО компании Amazon для обзора резюме кандидатов, гораздо чаще отвергавший кандидатуры женщин, или ряд инструментов распознавания лиц, которые хуже распознавали лица людей с более темным цветом кожи[49]. В контексте гуманитарной деятельности предотвращение нежелательных предубеждений и дискриминации тесно связано с основным гуманитарным принципом беспристрастности[50], и цена подобной дискриминирующей ошибки может быть особенно высока — например, если она сделана при принятии решения о том, кому направлять критически важную помощь, или даже решения о том, кто будет жить, а кто — нет[51]. На макроуровне алгоритмы (включая ИИ) могут «усиливать существующие неравенства между людьми или их группами, а также усугублять ущемление прав отдельных уязвимых демографических категорий». Так происходит из-за того, что «алгоритмы чаще других аналитических средств могут создавать вредоносные контуры

47 Д. Кей (примечание 38 выше), пп. 37–38.

48 Karen Hao, "This Is How AI Bias Really Happens — and Why It's So Hard to Fix", *MIT Technology Review*, 4 February 2019, доступно по адресу: www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/. Более подробное описание различных типов предубеждений, которые часто присутствуют в наборах данных или моделях обучения, см.: DSEG (примечание 9 выше).

49 K. Hao (примечание 48 выше); Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", *Proceedings of Machine Learning Research*, Vol. 81, 2018; Inioluwa Deborah Raji and Joy Buolamwini, *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*, 2019.

50 «Гуманитарная деятельность должна основываться исключительно на потребности, в ее рамках приоритет следует отдавать наиболее неотложным случаям потрясений безотносительно национальности, расы, гендера, класса, религиозных или политических убеждений». OCHA, "OCHA on Message: Humanitarian Principles", June 2012, доступно по адресу: www.unocha.org/sites/dms/Documents/OOM-humanitarianprinciples_eng_June12.pdf.

51 См., например, приведенную ниже дискуссию по поводу последствий применения автоматических систем оружия для международного гуманитарного права: Noel Sharkey, "The Impact of Gender and Race Bias in AI", *ICRC Humanitarian Law and Policy Blog*, 28 August 2018, доступно по адресу: https://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/.

обратной связи, потенциально тавтологические по своей сути, и действовать неконтролируемым образом в силу самой природы автоматизации алгоритма»[52].

Дефицит контекстуальных знаний на стадии разработки

Очень часто между специалистами, которые разрабатывают систему ИИ, и специалистами, занимающимися ее эксплуатацией, не происходит обмен информацией. Эта проблема становится особенно острой, если систему предстоит применять в гуманитарных контекстах[53]. У разработчиков инструментов может не быть надлежащего уровня знаний об условиях их применения; часто такие инструменты изначально приспособлены для принятия предпринимательских или маркетинговых решений, а не для оказания гуманитарной помощи в развивающихся странах. Использование инструментов, разработанных без учета определенных культурных, социальных и гендерно обусловленных аспектов, может приводить к принятию вводящих в заблуждение решений, которые отрицательно сказываются на жизнях людей. Например, система, задуманная или разработанная в Кремниевой долине, но развернутая в развивающейся стране, может оказаться неспособной учесть уникальные политические и культурные особенности данной страны. Разработчик может не знать о том, что в стране N определенные стигматизированные группы населения недостаточно представлены или вовсе остаются невидимыми в наборах данных, и потому не исправит это предубеждение в обучающей модели. Еще один вариант такой ситуации: разработчик, создающий инструмент, который будет применяться в рамках гуманитарной деятельности, может не знать о том, что представители мигрантских сообществ и внутренне перемещенные лица часто не включаются в данные переписей, демографическую статистику и другие наборы данных[54].

Дефицит специальных знаний и проблемы с реализацией на этапе «последней мили»

Недостаточный уровень специальных знаний или подготовки у лиц, отвечающих за развертывание систем ИИ и прочих инструментов, функционирующих на основании данных, несет ряд видов риска в области защиты прав человека. Это в основном актуально и для общественного сектора, где существует общепризнанная проблема недостаточного умения работать с данными[55]. Такая ситуация находит свое отражение в тенденции к непра-

52 DSEG (примечание 9 выше), p. 29.

53 На основании консультаций, проведенных нами в Женеве.

54 Дискуссию по поводу проблем сбора и анализа данных о популяциях мигрантов см.: Natalia Baal and Laura Ronkainen, *Obtaining Representative Data on IDPs: Challenges and Recommendations*, UNHCR Statistics Technical Series No. 2017/1, 2017, доступно по адресу: www.unhcr.org/598088104.pdf.

55 В стратегии ООН в области данных от 2020 г. уделено много внимания необходимости наращивания потенциала гражданских служащих всех органов ООН в сфере использования данных и новых технологий.

вильному толкованию результатов работы системы, завышенной оценке ее предиктивных возможностей или иначе выраженному чрезмерному доверию к результатам ее работы, в результате которого, например, «решения», принятые системой, получают больший приоритет, нежели человеческие суждения. Эта проблема также обусловливает риск того, что лица, ответственные за принятие решений или разработку политики, будут использовать ИИ в качестве «костыля», создавая видимость объективности и нейтральности сделанного ими выбора за счет обоснования последнего результатами проведенного ИИ анализа. Подобный риск еще больше усугубляется в случае с работой в развивающихся странах или в гуманитарных контекстах, когда недостаток надлежащих технических ресурсов, инфраструктуры или организационного потенциала может затруднить успешную эксплуатацию систем ИИ[56]. Проблемы с реализацией на так называемом этапе «последней мили» могут усугубить риск в области защиты прав человека и другие дефекты, особенно в рамках гуманитарной деятельности. Так, например, нехватка данных — предвиденная или непредвиденная — увеличивает вероятность человеческой ошибки, которая может выражаться в совершенно разных формах: от неспособности осуществлять аудит системы до чрезмерного доверия к результатам ее аналитической деятельности или их неправильной трактовки. Ошибки же, в свою очередь, ведут к неблагоприятному воздействию на людей, например к неспособности оказать критически важную помощь или даже к дискриминации и преследованию.

Дефицит качественных данных

Благонадежность и безопасность ИИ зависит от качества данных. Без готовых наборов качественных данных ИИ нельзя обучить и использовать так, чтобы исключить расширение описанных выше видов риска. Степень доступности и полноты данных, однако, часто бывает обусловлена социальными, экономическими, политическими и другими факторами неравенства[57]. Во многих ситуациях, связанных с деятельностью в целях развития и гуманитарной деятельностью, выполнить сбор качественных данных гораздо сложнее, чем обычно. Вследствие этого возрастает риск получения несправедливых результатов от системы ИИ[58]. При том что стандарты качества данных — давно устоявшееся явление, закрепленное ответственными техническими специалистами в форме принципов и передовых методов обеспечения качества данных[59], достаточно релевантные правовые основы

56 Michael Chui *et al., Notes from the AI Frontier: Modeling the Impact of AI on the World Economy,* McKinsey Global Institute, September 2018.

57 О пробеле в данных (относительно пожилых людей) см.: *СПЧ*. Осуществление всех прав человека пожилых людей, док. ООН A/HRC/42/43, 4 июля 2019 г.; *СПЧ*. Права человека пожилых людей: пробел в данных, док. ООН A/HRC/45/14, 9 июля 2020 г.

58 Jasmine Wright and Andrej Verity, *Artificial Intelligence Principles for Vulnerable Populations in Humanitarian Contexts*, Digital Humanitarian Network, January 2020, p. 15.

59 См., например, соответствующие разделы в OCHA's Data Responsibility Guidelines (примечание 16 выше); ICRC *Handbook on Data Protection in Humanitarian Action* (примечание 17 выше); Principles for Digital Development, доступно по адресу: https://digitalprinciples.org/.

для обеспечения доступности применимых на практике наборов данных по-прежнему отсутствуют. Как отмечает Генеральный секретарь ООН в разработанной им Дорожной карте: «многие существующие цифровые общественные блага [в том числе качественные данные] не являются легкодоступными, поскольку нередко они неравномерно распределены с точки зрения языка, содержания и инфраструктуры, необходимой для получения к ним доступа»[60].

Чрезмерное использование ИИ

Благодаря своим аналитическим и предиктивным способностям системы ИИ могут выглядеть привлекательными «решениями» сложных проблем как для ограниченных в ресурсах специалистов-практиков гуманитарной сферы, так и для лиц, стремящихся найти финансирование для гуманитарных проектов. По этой причине возникает риск чрезмерного использования ИИ, в том числе в условиях, когда имеются более надежные решения[61]. С одной стороны, существует распространенное заблуждение о способностях и ограничениях ИИ, включая технические ограничения его деятельности. В СМИ ИИ обычно изображается в виде всемогущих машин или роботов, которые могут решить широкий спектр аналитических проблем. В реальности же проекты с использованием ИИ обыкновенно являются узкоспециализированными и системы ИИ разрабатываются только для применения в конкретной ситуации и на отдельном наборе данных. В силу этого заблуждения пользователи могут не понимать, что взаимодействуют с системой, где используется ИИ. Кроме того, хотя работа ИИ иногда может заменить человеческий труд или аналитическую деятельность, в большинстве случаев ИИ не может заместить человека при принятии решений по особо деликатным вопросам или по проблемам, без устранения которых могут наступить значительные негативные последствия. Например, доверив системе, основанной на ИИ, вынесение приговоров по уголовным делам, а также принятие решений о предоставлении убежища[62] или лишении родительских прав — то есть в тех ситуациях, когда на кону стоят фундаментальные права и свободы и затронутые люди уже пережили травмы и потрясения — можно усугубить психологический

60 Дорожная карта Генерального секретаря (примечание 1 выше), п. 23.

61 «Автоматические мощности алгоритмов могут служить на благо, но также могут и препятствовать учету человеческих соображений в процессах, оказывающих воздействие на людей. Вот почему использование или чрезмерное использование алгоритмов несет в себе угрозу для популяций, затронутых процессами алгоритмов, так как вклад человека в такие процессы обычно является важным элементом защиты затронутых групп и исправления ошибок, допущенных в их отношении. Зачастую алгоритмы усугубляют существующее неравенство между людьми или их группами, а также ущемление прав отдельных уязвимых демографических категорий. Алгоритмы чаще других аналитических средств могут создавать вредоносные контуры обратной связи, потенциально тавтологические по своей сути, и действовать неконтролируемым образом в силу самой природы автоматизации алгоритма». DSEG (примечание 9 выше), p. 29.

62 Petra Molnar and Lex Gill, *Bots at the Gates*, University of Toronto International Human Rights Program and Citizen Lab, 2018.

ущерб, спровоцировать ограничение самостоятельности людей и даже разрыв социальных связей[63].

Влияние частного сектора

Разработчиками и операторами систем, которые используются в деятельности в целях развития и в гуманитарном секторе, являются в основном частные технологические компании. Такое сотрудничество часто оформляется в виде договоров с независимыми поставщиками или в форме государственно-частного партнерства. Это создает почву для возникновения ситуаций, когда корпоративные интересы могут ставиться выше общественных. К примеру, извлечение прибыли может стать значительным стимулом к тому, чтобы использовать дорогостоящий, «высокотехнологичный» подход, даже если более подходящим для конкретных целей и ситуации является альтернативный «низкотехнологичный» подход, который также доступен[64]. Кроме того, тесное сотрудничество между государствами и предприятиями может наносить ущерб прозрачности и подотчетности, например когда доступ к информации ограничен в соответствии с положениями договора или нормами защиты коммерческой тайны. Столь активное вовлечение коммерческих субъектов также приводит к делегированию им решений по вопросам, представляющим общественный интерес. Так, существует риск того, что гуманитарные организации и государства «будут делегировать частным компаниям решение все более сложных и серьезных задач, связанных с осуществлением цензуры и контроля»[65].

Сохранение и усугубление неравенства

Развертывание сложных систем ИИ для поддержки услуг, оказываемых людям, которые ущемлены в правах и находятся в уязвимом положении, порой может приводить к противоречивым последствиям — усугублению неравенства и дальнейшему ущемлению прав. К основным проблемам, возникающим в этой связи, можно отнести рассмотренные выше влияние предубеждений на данные и применение неподходящих моделей. Однако важно понимать, что сами эти проблемы являются отражением глубоко укоренившихся в структуре общества социально-экономических, гендерных и расовых разрывов — и все более активное применение ИИ влечет за собой риск их углубления. Подобная мысль была высказана в одном из недавно опубликованных исследований ЮНЕСКО, где такой результат связывается с влиянием ИИ на распределение власти: «Охват и власть, даруемые технологиями ИИ, усиливают дисбаланс сил, существующий между

63 DSEG (примечание 9 выше), p. 11.
64 На основании консультаций, проведенных нами в Женеве. См. также: Chinmayi Arun, "AI and the Global South: Designing for Other Worlds", in Markus D. Dubber, Frank Pasquale and Sunit Das (eds), *The Oxford Handbook of Ethics of AI*, Oxford University Press, Oxford, 2020.
65 Д. Кей (примечание 38 выше), п. 44.

людьми, их группами и странами, в том числе так называемый цифровой разрыв внутри стран и между ними»[66]. Отдание контроля за этими технологиями на откуп частным компаниям, как уже было сказано, является одним из главных факторов, способствующих такому развитию событий. Переломить данную тенденцию — сложнейшая задача, для осуществления которой потребуются политическая воля, сотрудничество, открытая работа с различными заинтересованными сторонами, укрепление демократического управления в странах и содействие защите прав человека. Только так можно расширить права и возможности людей и обеспечить их активное участие в определении характеристик технологической и нормативной среды, в которой они живут.

Междисциплинарные соображения

Некоторые из этих проблем являются отличительной особенностью систем ИИ и нехарактерны для других технологий, выработкой стандартов регулирования для которых мы занимались ранее; потому эти проблемы могут требовать новых решений. Стоит также заметить, что некоторые из основных проблем отнюдь не являются новыми. Поэтому иногда у нас есть возможность почерпнуть передовые практические методы управления ИИ из других областей. Например, угрозы для конфиденциальности и безопасности данных и соответствующие стандарты, предназначенные для защиты информации, существуют уже долгое время. Безусловно, по мере развития технологий и генерирования все бóльших объемов данных требуется создавать новые методы их защиты или обновлять старые, отталкиваясь от характера изменений. В силу чувствительного характера собираемых и обрабатываемых данных их безопасность остается одним из основных приоритетов в гуманитарной деятельности.

Более того, многие из проблем, с которыми связано применение ИИ при оказании гуманитарной помощи, уже были решены практическими специалистами в более широкой области «технологий в целях развития»[67], например проблемы с реализацией на этапе «последней мили», упомянутые ранее. Еще одна всегда актуальная проблема выражается в том, что при реализации проектов в целях развития или гуманитарных проектов периодически требуется взвешивать риск сотрудничества с правительствами стран, ситуация с правами человека в которых оставляет желать лучшего. Этот принцип в полной мере применим к таким мощным инструментам, как ИИ. Правительства потенциально могут использовать системы ИИ, создававшиеся в общественно полезных целях — например, систему цифрового отслеживания контактов для сдерживания вспышки заболевания — как

66 UNESCO, *Preliminary Study on the Ethics of Artificial Intelligence*, SHS/COMEST/EXTWG-ETHICS-AI/ 2019/1, 26 February 2019, para. 22.

67 См., например: Principles for Digital Development (примечание 59 выше).

инструмент принудительного надзора[68]. Помимо всех вышеперечисленных проблем, которые широко распространены и потенциально могут приводить к нанесению ущерба, организационные условия, в которых существуют системы и процессы ИИ, являются столь же важным детерминантом связанного с ними риска. Независимо от того, сколь велика заявленная аналитическая или предиктивная мощность системы (будь то простой алгоритм или сложные нейросети), на практике масштаб выгод и риска причинения ущерба при ее использовании будет сильнейшим образом зависеть от степени взаимодействия с ней человека или от степени контроля с его стороны.

Проблемы, описанные выше, существуют отнюдь не только в теории — по всему миру зафиксировано бессчетное количество ситуаций, когда использование передовых систем ИИ приводило к серьезному ущербу. В ряде наиболее резонансных происшествий, связанных с ИИ, операторами системы были правительственные органы или другие субъекты общественного сектора, которые хотели усовершенствовать или рационализировать оказание той или иной общественной услуги. Например, одной из последних тенденций является использование правительствами алгоритмического анализа для определения правомерности заявок на получение социальных пособий или отсеивания недобросовестных заявок[69]. В Австралии, Нидерландах и Соединенных Штатах Америки структурные дефекты таких систем или ненадлежащий контроль человека за их работой привели — кроме прочих проблем — к тому, что множество людей были лишены прав на получение финансовой помощи, жилья или услуг здравоохранения[70]. В августе 2020 года Министерство внутренних дел Соединенного Королевства приняло решение прекратить использование алгоритма принятия решений, задействованного для анализа кандидатов на получение визы, ввиду сообщений о расовых предубеждениях, интегрированных в систему[71].

Об ущербе, причиненном вследствие применения ИИ в гуманитарных контекстах, мы пока знаем относительно немного. По наблюдениям Группы по науке о данных и этике данных в гуманитарной деятельности,

68 См.: UN Human Rights, *UN Human Rights Business and Human Rights in Technology Project (B-Tech)*: Overview and Scope, November 2019, где содержится предупреждение о неотъемлемом риске в области защиты прав человека, который возникает при «продаже продукции или партнерстве с правительствами, стремящимися использовать инновационные технологии для выполнения государственных функций или оказания общественных услуг, что может приводить к возникновению несоразмерно большего риска для уязвимых групп населения, чем для остальной его части».

69 *Филип Олстон.* Доклад Специального докладчика по вопросу о крайней нищете и правах человека, док. ООН A/74/493, 11 октября 2019 г.

70 AI Now Institute, *Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems*, September 2018, доступно по адресу: https://ainowinstitute.org/litigatingalgorithms.pdf; Ф. Олстон (примечание 69 выше). Обратите внимание на то, что даже идеально спроектированная система, в контуре управления которой присутствуют люди, может вести к получению отрицательных результатов, если она не подходит для тех условий, в которых применяется. Например, широко распространенная, глубоко укоренившаяся дискриминация в угнетающей среде может спровоцировать дальнейшее усугубление дискриминации при использовании системы ИИ, даже если сама по себе система не содержит предрассудков и в контуре управления присутствует человек.

71 Henry McDonald. "Home Office to Scrap 'Racist Algorithm' for UK Visa Applicants", *The Guardian*, 4 August 2020.

отраженным в ее докладе, сохраняется «нехватка задокументированных свидетельств» о риске и ущербе, связанных с ИИ, «так как подобные происшествия отслеживаются и освещаются недостаточно активно» и «обычно никто не хочет сообщать об инцидентах»[72]. При том что наличие описанных выше видов риска в других контекстах (таких как социальное обслуживание) детально подтверждено, в гуманитарных контекстах пока приходится говорить только о свидетельствах потенциальной обеспокоенности в связи с биометрическими технологиями и страхах людей, затрагиваемых ими.

Показательным примером может служить недавнее ситуационное исследование психотерапевтического чат-бота «Карим», разработанного и испытанного во взаимодействии с сирийскими беженцами, проживающими в лагере беженцев Заатари. Эксперты, беседовавшие с исследователями Цифровой гуманитарной сети (Digital Humanitarian Network), выразили обеспокоенность по поводу того, что, хотя разработка терапевтического чат-бота с использованием ИИ является передовой практикой, ее результаты отражают недостаточное понимание нужд уязвимых людей в подобных условиях[73]. Помимо лингвистических и логистических препятствий, которые проявились на пилотной стадии проекта, эксперты говорят и о том, что лучше не иметь возможности получить психотерапевтическую помощь вообще, чем получать ее от психотерапевта-машины — такая помощь создавала риск усиления чувства одиночества в долгосрочной перспективе[74]. Вероятно, ситуация с чат-ботом «Карим» — это воплощение того разрыва, который, по утверждениям авторов исследования «Проект гуманитарных технологий» (Humanitarian Technologies Project), существует «между допущениями об использовании технологий в гуманитарных контекстах и реальным опытом их использования и уровнем их эффективности для уязвимых групп людей»[75].

Данные проблемы свидетельствуют о том, что пилотное тестирование инструментов, безопасность которых еще не доказана, на уязвимых группах населения может значительно ущемлять права человека этих групп, если такие инструменты не приспособлены к реальным условиям или у применяющих их лиц нет достаточных специальных знаний[76].

Подходы к управлению ИИ: больше, чем этические принципы

Приведенные ранее примеры демонстрируют, что ИИ может как служить интересам людей, так и наносить таким интересам ущерб, если не приняты

72 DSEG (примечание 9 выше), p. 3.
73 J. Wright и A. Verity (примечание 58 выше), p. 7.
74 Ibid., p. 6.
75 Ibid., p. 9. См. также веб-сайт исследования «Проект гуманитарных технологий», доступно по адресу: http://humanitariantechnologies.net.
76 См.: DSEG (примечание 9 выше), p. 8, где содержится предостережение относительно пилотного тестирования инструментов, безопасность которых еще не доказана, в гуманитарных контекстах.

надлежащие меры предосторожности и не учтен риск. Поэтому технические специалисты, проектирующие такие системы, а также эксперты в области развития и гуманитарной деятельности, занимающиеся развертыванием ИИ, все чаще учитывают необходимость интегрировать права человека и этические принципы в свою работу. Соответственно, разрабатывается все больший объем технических требований и стандартов для обеспечения «безопасности», «защищенности» и «благонадежности» систем ИИ[77]. Однако, для того чтобы гарантировать, что ИИ служит общечеловеческим интересам, недостаточно одного лишь применения технических требований. МакГрегор, Мюррей и Нг в своей работе указывают на необходимость более широкой, всеобъемлющей рамочной основы, которая учитывала бы риск нанесения ущерба на каждом этапе жизненного цикла системы и обеспечивала бы подотчетность в случае возникновения отрицательных последствий[78].

Первые документы, регламентирующие управление ИИ, которые формально разрабатывались именно для достижения этой цели, обычно принимали форму «этических кодексов применения ИИ»[79]. Такие кодексы, как правило, сформированы на базе руководящих принципов, которые поощряются в организации, и имеют сходство с корпоративными положениями о развитии и использовании ИИ. Как следует из названия, подобные кодексы обычно апеллируют к этическим принципам, таким как честность и справедливость, а не гарантируют соблюдение конкретных прав человека[80]. Действительно, права человека — всеобщая и обладающая юридической силой система принципов и договоров, которые должны соблюдать все государства, — совершенно не отражены во многих подобных документах[81]. По словам Филипа Олстона, Специального докладчика ООН по вопросу о крайней нищете и правах человека, многие этические кодексы применения ИИ содержат отсылки к правам человека — в частности, в них фигурирует приверженность соблюдению прав человека в качестве отдельного принципа, — но не охватывают материальные права, закрепленные во Всеобщей декларации прав человека (ВДПЧ) и договорах в области прав человека[82].

77 Peter Cihon, *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*, Future of Humanity Institute, University of Oxford, April 2019.

78 «Сложная природа процесса принятия решений с помощью алгоритмов обусловливает потребность встраивать проекты механизмов обеспечения подотчетности в более масштабную систему, охватывающую весь жизненный цикл алгоритма: от создания концепции и проектирования до развертывания в реальных условиях и применения алгоритмов в принятии решений». L. McGregor, D. Murray, V. Ng (примечание 5 выше), p. 311.

79 Обзор этических кодексов применения ИИ, выпущенных крупными учреждениями, см.: J. Fjeld *et al*. (примечание 8 выше).

80 Ibid.

81 См.: Mark Latonero, *Governing Artificial Intelligence: Upholding Human Rights and Dignity*, Data & Society, 2018, где утверждается, что права человека редко имеют приоритетный статус в рамках национальных стратегий в области ИИ. При этом есть ряд исключений, в том числе Общий регламент ЕС о защите персональных данных (GDPR) и стратегические документы Совета Европы, Глобального партнерства в области ИИ, возглавляемого Канадой и Францией, и Австралийской комиссии по правам человека.

82 См.: Ф. Олстон (примечание 69 выше), где утверждается, что большинство этических кодексов со-

Недостатки подхода, позиционирующего этические принципы как приоритетное средство, становятся все более очевидны. Одним из главных пробелов является отсутствие механизмов обеспечения ответственности для ситуаций, когда этические принципы нарушаются[83]. Большинство этических кодексов не дает ответа на вопрос о том, кто несет материальную ответственность за «неэтичное» использование технологий, как определять размер такого материального ущерба, отслеживать нарушения и привлекать к ответственности за них. Кроме того, неясно, как именно лицо, предполагающее, что ему был причинен ущерб, должно определить, так ли это, равно как и неясно, в каком порядке оно должно добиваться компенсации[84]. В отличие от права прав человека, этические кодексы, как правило, не регламентируют порядок уравновешивания интересов различных групп или лиц, ряд из которых могут получать выгоду от использования систем ИИ в ущерб другим субъектам. Этические кодексы применения ИИ являются важной первой ступенью на пути к созданию механизмов управления, обладающих большей юридической силой, но для обеспечения реального воздействия на ситуацию им необходимо придать более конкретную форму по аналогии с конкретными правами, нарушение которых влечет за собой юридическое наказание.

Права человека в качестве основы

В силу обозначенных выше и прочих причин большинство участников консультаций, организованных инициативой Генерального секретаря ООН «Глобальный пульс» и Управлением ООН по правам человека, сошлись во мнении[85] о том, что права человека должны лежать в основе любого эффективного режима управления ИИ. Международное право прав человека предоставляет легитимную во всем мире комплексную рамочную основу для прогнозирования и предотвращения риска и ущерба, а также возмещения последнего. Как утверждают в своей работе МакГрегор и др., международное право прав человека обеспечивает «организационную структуру для проектирования, разработки и внедрения алгоритмов, а также выявляет факторы, которые государствам и предприятиям следует учитывать, чтобы не ущемлять и не нарушать права человека»[86]. Эта рамочная структура отнюдь не является лишь отдельным и статичным набором

держат отсылки к правам человека, но по существу не обеспечивают их, и что подобные отсылки используются только для придания положениям кодекса легитимности и универсальности.

83 Corinne Cath, Mark Latonero, Vidushi Marda and Roya Pakzad, "Leap of FATE: Human Rights as a Complementary Framework for AI Policy and Practice", в *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January 2020, доступно по адресу: https://doi.org/10.1145/3351095.3375665.

84 Ibid.

85 Под консультациями имеются в виду в том числе совещания и семинары, организованные инициативой Генерального секретаря ООН «Глобальный пульс» и Управлением ООН по правам человека в Женеве, Берлине и Тунисе.

86 L. McGregor, D. Murray and V. Ng (примечание 5 выше), p. 313.

«правил», а, напротив, «способна перенимать альтернативные подходы к подотчетности алгоритмов — в том числе технические решения — и... развиваться самостоятельно и служить основой для других механизмов, так как международное право прав человека постоянно совершенствуется и обновляется, в особенности в аспекте взаимосвязей предпринимательства и прав человека[87].

Аргументы в пользу международного права прав человека (МППЧ) можно подразделить по нескольким не связанным между собой аспектам, что превращает эту рамочную основу в особенно подходящую для смягчения риска и ущерба, причиняемого ИИ. Во-первых, в отличие от этических принципов, МППЧ является универсальным[88]. Оно воплощает единый терминологический аппарат и набор принципов, которые применимы независимо от национальных и культурных особенностей, и тем самым гарантирует, что ИИ будет служить общечеловеческим ценностям в соответствии с положениями ВДПЧ и других регламентирующих документов. Другого набора общих моральных или правовых принципов, обладающего столь же широким признанием, что и ВДПЧ, не существует[89]. В мире, где технологии и данные практически беспрепятственно пересекают границы стран и технологии нельзя эффективно регламентировать в рамках единой юрисдикции, подобная всемирная легитимность имеет критически важное значение.

Во-вторых, международный режим защиты прав человека имеет юридическую силу в отношении государств. В частности, он обязывает их ввести в действие рамочную основу, которая бы «предотвращала нарушения прав человека, устанавливала механизмы мониторинга и надзора, обеспечивала привлечение соответствующих лиц к ответственности и давала бы людям и группам людей, утверждающим, что их права были нарушены, средства защиты таких прав»[90]. На международном уровне внутри режима МППЧ также применяются системы обеспечения подотчетности и защиты прав, в том числе СПЧ и договорные органы, в рамках которых функционируют механизмы подачи жалоб и обзора эффективности работы государств-членов; специальные процедурные органы Совета по правам человека (а именно рабочие группы и специальные докладчики), занимающиеся проведением расследований и подготовкой докладов и заключений[91]; а также все чаще Международный суд ООН, который начал наращивать свое влияние в области прав человека и гуманитарной правоприменительной практики[92]. Кроме того, одну из главных ролей в развитии системы

87 L. McGregor, D. Murray and V. Ng (примечание 5 выше), p. 313.

88 «Права человека считаются всеобщими как в силу того, что они признаются практически всеми странами мира, так и потому, что они применимы ко всем людям вне зависимости от их индивидуальных характеристик». Nathalie A. Smuha, "Beyond a Human Rights-based Approach to AI Governance: Promise, Pitfalls, Plea", *Philosophy and Technology*, 2020 (готовится к публикации).

89 Ibid.

90 L. McGregor, D. Murray and V. Ng (примечание 5 выше), p. 311.

91 Ibid.

92 Lyal S. Sunga, "The International Court of Justice's Growing Contribution to Human Rights and Humanitarian Law," The Hague Institute for Global Justice, The Hague, 18 April 2016.

защиты прав человека взяли на себя региональные механизмы защиты прав человека. Они в том числе обеспечивают людям возможность привлечь к правовой ответственности лиц, виновных в нарушении прав человека[93].

В-третьих, МППЧ сосредоточивает свою аналитическую работу на обладателе прав и носителе обязанностей в конкретных условиях, за счет чего применять принципы к реальным ситуациям становится гораздо проще[94]. Вместо того чтобы стремиться к обеспечению размытых идеалов, таких как честность, право прав человека призывает разработчиков и операторов систем ИИ сосредоточиться на том, кто именно будет затронут применяемой технологией и какие именно основные права человека могут оказаться под угрозой. Это чрезвычайно полезное для практики свойство, которое позволяет перевести внимание с высоких идеалов на конкретные и узко определенные угрозы и случаи нанесения ущерба. Схожим образом многие механизмы подотчетности в области прав человека также позволяют людям защитить свои права, подав жалобы в различные судебные органы. Безусловно, на практике обращение в суды по правам человека и корректное составление жалобы — гораздо более сложная задача, нежели в теории. Однако система защиты прав человека по меньшей мере обеспечивает таким людям «терминологический и процессуальный аппарат для оспаривания правомерности действий учреждений и организаций, обладающих властью», будь то государства или корпорации[95].

В-четвертых, определяя конкретные права, МППЧ также определяет конкретные виды ущерба, которые следует предотвращать, смягчать и возмещать[96]. Тем самым оно задает результаты, на достижение которых государства и другие организации — в том числе гуманитарные и действующие в целях развития — могут направить свои усилия. Так, Комитет ООН по экономическим, социальным и культурным правам разработал стандарты «доступности, адаптивности и приемлемости», которые государства должны обеспечивать в рамках своих программ социальной защиты[97].

Наконец, право прав человека и правоприменительная практика в области прав человека обеспечивают основу для уравновешивания конфликтующих друг с другом прав[98]. Данная основа критически важна

93 UN Human Rights, "Regional Human Rights Mechanisms and Arrangements", доступно по адресу: www.ohchr.org/EN/Countries/NHRI/Pages/Links.aspx.

94 C. Cath *et al*. (примечание 83 выше).

95 Christian van Veen and Corinne Cath, "Artificial Intelligence: What's Human Rights Got to Do With It?", Data & Society, 14 May 2018, доступно по адресу: https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5.

96 L. McGregor, D. Murray and V. Ng (примечание 5 выше).

97 См.: Доклад по вопросу о реализации во всех странах экономических, социальных и культурных прав (примечание 44 выше); "Standards of Accessibility, Adaptability, and Acceptability", *Social Protection and Human Rights*, доступно по адресу: https://socialprotection-humanrights.org/framework/principles/standards-of-accessibility-adaptability-and-acceptability/.

98 См.: Karen Yeung, Andrew Howes and Ganna Pogrebna, "AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing", в Markus D. Dubber, Frank Pasquale and Sunit Das (eds), *The Oxford Handbook of Ethics of AI*, Oxford University Press, Oxford, 2020, где отмечается, что МППЧ предоставляет «структурированную рамочную основу для решения конфликтов, возникающих в результате пересечения конкурирующих прав и коллективных

в ситуациях, когда необходимо решить, применять ли технологический инструмент, если это повлечет за собой возникновение как выгод, так и риска. В таких случаях право прав человека регламентирует, как и когда могут ограничиваться те или иные основные права — посредством применения принципов законности, легитимности, необходимости и пропорциональности к предложенному вмешательству ИИ[99]. В этом смысле МППЧ также позволяет установить границы допустимого, то есть действия, которые выходят за эти рамки[100]. Такая основа будет особенно полезна для гуманитарных организаций, которым требуется понять, необходимо ли полностью отказаться (и если да, то в каких ситуациях) от использования той или иной способности ИИ, например технологии распознавания лиц.

В большинстве случаев применения ИИ в гуманитарной деятельности потребность в уравновешивающей рамочной основе вполне очевидна. Уравновешивающий подход был интегрирован в процедуру оценки риска, ущерба и выгод, созданную в рамках инициативы Генерального секретаря ООН «Глобальный пульс». Эта процедура устроена так, чтобы побуждать операторов проектов ИИ или анализа данных не просто учитывать риск для конфиденциальности и вероятность, масштаб и тяжесть/существенность потенциального ущерба, но и соотносить эти риск и ущерб с прогнозируемыми выгодами от реализации проекта. Практика применения МППЧ помогает регламентировать использование мощных инструментов ИИ в таких контекстах, определяя, что такое использование допустимо только в установленных законом случаях, для достижения законной цели и при

интересов в отдельных случаях», а этические кодексы применения ИИ содержат «крайне недостаточно указаний по поводу того, как разрешать подобные конфликты».

99 Ограничение того или иного права (если такое ограничение допустимо в принципе) должно быть обусловлено необходимостью достижения законной цели и носить пропорциональный этой цели характер. Такие ограничения должны быть наименее принудительным вариантом из доступных и не должны применяться или приводиться в качестве аргумента таким образом, который бы противоречил сути соответствующего права. Они должны быть закреплены в общедоступном законодательном акте, четко определяющем обстоятельства, в которых могут вводиться те или иные ограничения прав. См.: Доклад по вопросу о реализации во всех странах экономических, социальных и культурных прав (примечание 44 выше), с. 10–11. См. также: N. A. Smuha (примечание 88 выше), где отмечается, что схожие формулировки для уравновешивания конкурирующих между собой прав содержатся в Хартии ЕС по правам человека, Европейской конвенции по правам человека и в статье 29 ВДПЧ.

100 Catelijne Muller, *The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law*, Ad Hoc Committee on Artificial Intelligence, Strasbourg, 24 June 2020, para. 75, доступно по адресу: https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da. McGregor *et al.* определяют ограничения исходя из «запрета на наложение произвольных прав как ключевого принципа, подкрепляющего МППЧ и релевантного для всех решений, которые потенциально могут противоречить тем или иным правам». L. McGregor, D. Murray and V. Ng (примечание 5 выше), p. 337. Более подробную информацию о соотношении «произвольного» и «необходимого и пропорционального» см.: *Управление ООН по правам человека*. Право на неприкосновенность личной жизни в цифровой век: Доклад Управления Верховного комиссара ООН по правам человека, док. ООН A/HRC/27/37, 30 июня 2014 г., п. 21 и последующие; *Управление ООН по правам человека*. Право на неприкосновенность личной жизни в цифровой век: Доклад Верховного комиссара ООН по правам человека, док. ООН A/39/29, 3 августа 2018 г., п. 10.

условии, что оно необходимо и пропорционально такой цели[101]. Добиваясь равновесия, лица, ответственные за принятие решений, могут обратиться к накопленной за десятилетия совокупности практики применения МППЧ, чтобы выяснить, как следует разрешать трения между конфликтующими правами или правами разных лиц[102]. Среди других примеров инструментов и руководств[103], в которых содержится уравновешивающая основа, можно назвать Международные принципы применения прав человека к надзору за коммуникациями[104]и Директивную записку УКГВ об оценке воздействия данных[105].

Пробелы в осуществлении МППЧ: подотчетность частного сектора

Наиболее значительным ограничением МППЧ является то, что его юридическое действие распространяется только на государства. Поэтому отдельные лица могут обращаться с жалобами на нарушение прав человека только по вертикали — против государства, а не в горизонтальной плоскости — против других граждан, организаций или, что важно, компаний[106]. Это, по-видимому, является препятствием для обеспечения подотчетно-

101 МППЧ «предоставляет четкую рамочную основу для уравновешивания конкурирующих интересов при разработке технологий: согласно его проверенным и надежным правовым установкам ограничение прав человека (такие как право на неприкосновенность частной жизни и право на недискриминацию) допустимо только в установленных законом случаях, для достижения законной цели и при условии, что оно необходимо и пропорционально такой цели. Каждый термин имеет четкое определение, что позволяет объективно оценивать связанные с ним действия и обеспечивает подотчетность». Alison Berthet, "Why Do Emerging AI Guidelines Emphasize 'Ethics' over Human Rights?" *OpenGlobalRights*, 10 July 2019, доступно по адресу: www.openglobalrights.org/why-do-emerging-ai-guidelines-emphasize-ethics-over-human-rights.

102 «Более того, чтобы добиться этого, правоприменители могут исходить из опыта ранее реализованных мер по уравновешиванию, обеспечивая таким образом предсказуемость и правовую ясность. Десятилетия институционально оформленного обеспечения соблюдения прав человека позволили наработать большую базу правоприменительной практики, от которой правоприменители могут отталкиваться при работе над вопросами воздействия систем ИИ на людей и общество, а также над возникающими в результате этого воздействия трениями, — будь то конфликт прав, принципов или интересов». N. A. Smuha (примечание 88 выше).

103 Более подробные инструкции по созданию процедуры оценки воздействия, ориентированной на защиту прав человека, см.: *Управление ООН по правам человека*. Руководящие принципы предпринимательской деятельности в аспекте прав человека, Нью-Йорк и Женева, 2011 г., доступно по адресу: https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_RU.pdf.

104 С Принципами можно ознакомиться по адресу: www.eff.org/files/necessaryandproportionatefinal.pdf. Справочную информацию и правовой анализ см.: Electronic Frontier Foundation and Article 19, *Necessary and Proportionate: International Principles on the Application of Human Rights to Communication Surveillance*, May 2014, доступно по адресу: www.ohchr.org/Documents/Issues/Privacy/ElectronicFrontierFoundation.pdf.

105 ICRC, Privacy International, UN Global Pulse and OCHA Centre for Humanitarian Data, "Guidance Note: Data Impact Assessments", Guidance Note Series No. 5, July 2020, доступно по адресу: https://centre.humdata.org/wp-content/uploads/2020/07/guidance_note_data_impact_assessments.pdf. Больше примеров процедур оценки воздействия, предназначенных для гуманитарных контекстов, см. в данной Директивной записке.

106 John H. Knox, "Horizontal Human Rights Law", *American Journal of International Law*, Vol. 102, No. 1, 2008, p. 1.

сти ИИ, так как частный сектор играет ведущую роль в разработке ИИ и создает бóльшую часть инноваций в этой области. Безусловно, в соответствии с МППЧ государства обязаны интегрировать стандарты защиты прав человека в национальное законодательство, которое регулирует деятельность частного сектора. Но, как показывает практика, так происходит далеко не всегда, и даже если государства *действительно* интегрируют правовые акты в области прав человека в национальное законодательство, у них нет возможности обеспечить их соблюдение за пределами соответствующей юрисдикции. При этом многие крупные технологические компании одновременно работают в разных странах, в том числе в тех, где меры по защите прав человека обеспечены в меньшей или недостаточной степени.

Тем не менее право прав человека имеет моральное и символическое влияние, которое может определять ход общественных дискуссий, способствовать конструктивной критике и служить средством оказания давления на компании; кроме того, все чаще признается, что у компаний есть обязанности в области прав человека, которые должны исполняться независимо от способности и готовности государств исполнять их собственные обязанности в области прав человека[107]. Существует ряд механизмов и рычагов давления, которые стимулируют компании к выполнению соответствующих требований.

Руководящие принципы предпринимательской деятельности в аспекте прав человека ООН начинают выполнять роль международной нормы поведения для субъектов предпринимательства, исключающего нарушение прав человека[108]. Эти принципы закрепляют концепцию обязанности субъектов предпринимательства соблюдать права человека в ходе предпринимательской деятельности. В них также содержится призыв к компаниям проводить комплексные внутренние проверки соблюдения прав человека, благодаря которым можно выявить и устранить или смягчить отрицательное воздействие на права человека, возникающее при поставке, разработке и использовании продукции, выпускаемой такой компанией[109]. Все больше органов по защите прав человека заявляют о том, что к алгоритмической обработке данных, ИИ и другим новым цифровым тех-

107 См.: I. Ebert, T. Busch and F. Wettstein (примечание 41 выше). См. также: C. van Veen and C. Cath (примечание 95 выше), где утверждается, что «права человека как терминологический аппарат и правовая основа сами по себе являются источником власти, поскольку права человека отличаются значительной моральной легитимностью, а репутационные издержки для организации, которую сочтут виновной в нарушении прав человека, могут оказаться очень высокими». О ситуации с алгоритмическими системами см.: Council of Europe, *Recommendation CM/Rec(2020)1 of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems*, 8 April 2020.

108 Руководящие принципы предпринимательской деятельности в аспекте прав человека ООН (примечание 103 выше). Глава I Руководящих принципов определяет порядок регулирования компаний со стороны государства.

109 Там же, глава II. См. также: UN Human Rights, *Key Characteristics of Business Respect for Human Rights*, B-Tech Foundational Paper, доступно по адресу: www.ohchr.org/Documents/Issues/Business/B-Tech/key-characteristics-business-respect.pdf.

нологиям применяются те же обязанности[110] — наиболее свежим примером подобных заявлений может служить доклад Верховного комиссара ООН по правам человека об использовании таких технологий, как распознавание лиц, в контексте мирных протестов[111]. Кроме того, Управление ООН по правам человека в настоящее время работает над созданием исчерпывающего руководства по применению Руководящих принципов предпринимательской деятельности в аспекте прав человека ООН к разработке и использованию цифровых технологий[112]. Растет и число ведущих компаний сферы ИИ, которые начинают применять Руководящие принципы предпринимательской деятельности в аспекте прав человека ООН к своим продуктам, основанным на ИИ (в их числе — Element AI, Microsoft и Telefonica)[113].

Второй повод для критики подхода к ИИ, основанного на защите прав человека, — это утверждение о том, что выделение правам человека приоритета на каждом этапе цикла развертывания технологий затруднит внедрение инноваций. В некоторой степени оно правдиво — внимание к правам человека может иногда становиться причиной задержки или даже отказа от внедрения продукта, использование которого создаст риск. Тем не менее такая расстановка приоритетов также позволяет на более поздних этапах избежать бóльших потенциальных затрат на урегулирование последствий нарушения прав человека[114]. Кроме того, ценность подхода, основанного на защите прав человека, заключается в том, что он обеспечивает не только их соблюдение, но и их интеграцию в концепцию проекта, а также в процесс разработки и развертывания. Отдание приоритета правам человека на каждом этапе процесса разработки должно, таким образом, уменьшить число случаев, когда внедрять тот или иной продукт оказывается слишком рискованно.

110 См.: Council of Europe, *Addressing the Impacts of Algorithms on Human Rights: Draft Recommendation*, MSI-AUT(2018)06rev3, 2018 («Субъекты частного сектора, задействованные при проектировании, разработке, продаже, развертывании, эксплуатации и обслуживании алгоритмических систем, будь то в общественном или частном секторе, должны проводить должные экспертизы на предмет соответствия правам человека. Такие субъекты несут ответственность за соблюдение признанных на международном уровне прав человека и основных свобод своих заказчиков и других сторон, которых затрагивают их действия. Эта ответственность существует независимо от способности и готовности государства исполнять свои собственные обязанности в области прав человека»). См. также: Д. Кей (примечание 38 выше).

111 *СПЧ*. Воздействие новых технологий на поощрение и защиту прав человека в контексте собраний, включая мирные протесты: доклад Верховного комиссара ООН по правам человека, док. ООН A/HRC/44/24, 24 июня 2020 г.

112 UN Human Rights, *The UN Guiding Principles in the Age of Technology*, B-Tech Foundational Paper, доступно по адресу: www.ohchr.org/Documents/Issues/Business/B-Tech/introduction-ungp-age-technology.pdf.

113 Среди примеров можно назвать инструмент Microsoft для оценки воздействия в области прав человека (ОВПЧ) и аналогичный инструмент Google, используемый при распознавании лиц знаменитостей; см. также: Element AI, *Supporting Rights-Respecting AI*, 2019; Telefonica, "Our Commitments: Human Rights," доступно по адресу: www.telefonica.com/en/web/responsible-business/human-rights.

114 L. McGregor, D. Murray and V. Ng (примечание 5 выше).

Роль этики

Если пределы допустимого при управлении ИИ должны определяться правами человека, то этические принципы могут служить основным средством соблюдения принципа ответственности при таком управлении. Даже самые активные поборники применения подхода к ИИ, основанного на защите прав человека, не отрицают значимость этики как вспомогательного инструмента для подкрепления и дополнения прав человека. В контексте ИИ под «этикой» обычно понимают так называемые принципы FAccT: справедливость ("Fairness"), отчетность ("ACCountability") и прозрачность ("Transparency"), которые иногда также обозначаются FATE, где E — это этика ("Ethics")[115]. Некоторые убеждены в том, что подход FAccT противоречит принципу строгости права и предполагает отказ от использования строго определенных прав в пользу более широкого суждения о том, какое воздействие на общество будет оказывать та или иная система[116]. В этом смысле этика часто рассматривается как структура, которую гораздо легче адаптировать к условиям технологической эволюции и реалиям современного мира, нежели принципы МППЧ, разработанные десятки лет назад — задолго до начала распространения систем ИИ и МО.

Несмотря на то что между подходами, основанными на правах человека и этике, существуют важные различия, в ходе наших консультаций выяснилось, что противопоставление «права человека — этика», доминирующее как принцип в политике относительно ИИ, в некотором смысле может быть ложной дихотомией[117]. Стоит отметить, что по существу предназначение прав человека и этики общее. Как лаконично определено организацией Access Now, при неэтичном использовании ИИ с большой долей вероятности нарушаются и права человека (и наоборот)[118]. При этом активисты в области прав человека совершенно обоснованно выражают беспокойство по поводу такого явления, как «вымывание этики»[119]. Оно выражается в том, что создатели технологий — нередко частные компании — осуществляют саморегулирование с опорой на размытые этические кодексы, соблюдение которых невозможно эффективно обеспечить. Технические эксперты, в свою очередь, часто скептически относятся к перспективе адаптации «строгих» законов в области прав человека к новым функциям и риску нанесения ущерба при применении ИИ и МО. При том

115 Microsoft выпустила ряд публикаций, освещающих ее работу над FATE. См.: "FATE: Fairness, Accountability, Transparency, and Ethics in AI", доступно по адресу: www.microsoft.com/en-us/research/group/fate#!publications.

116 C. Cath *et al.* (примечание 83 выше).

117 Полезную справочную информацию о преимуществах и недостатках рамочных основ этики и прав человека в области ИИ см.: Business for Social Responsibility (BSR) and World Economic Forum (WEF), *Responsible Use of Technology*, August 2019, p. 7 (где утверждается, что применение этических принципов и прав человека должно «создавать синергию»).

118 Access Now (примечание 19 выше).

119 Ben Wagner, "Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping?", в Emre Bayamlioglu, Irina Baraliuc, Liisa Janssens and Mireille Hildebrandt (eds), *Being Profiled: Cogitas Ergo Sum. 10 Years of Profiling the European Citizen*, Amsterdam University Press, Amsterdam, 2018.

что в обоих случаях подобное беспокойство может быть небезосновательным, эти два подхода на самом деле могут дополнять друг друга, а не конфликтовать между собой.

Например, для создания правовых норм в области прав человека достаточно специфичных для регулирования новых цифровых технологий, может потребоваться довольно много времени, а если речь идет об интеграции таких норм в национальное законодательство, то срок увеличивается дополнительно. В подобной ситуации, когда правовые нормы не предлагают ясных и готовых ответов на вопросы разработчиков и операторов ИИ, этические принципы помогут заполнить существующие пробелы[120]; данная функция, однако, может выполняться и за счет толкования положений существующих нормативных актов в области прав человека или с помощью прецедентного права. Кроме того, этические нормы могут поднимать планку минимально допустимых стандартов выше, чем это определено в рамочных основах по защите прав человека, или помогать интегрировать принципы, которые недостаточно закреплены в праве прав человека[121]. Так, например, организация, разрабатывающая инструменты ИИ, может обязаться гарантировать надзор человека за любыми решениями, принимаемыми с участием ИИ, — такой принцип в явной форме не закреплен ни в одном договоре в области прав человека, но, безусловно, будет укреплять права человека (и способствовать их осуществлению)[122]. Другие организации, стремящиеся обеспечить равное распределение экономических или материальных выгод от использования ИИ, в своей работе с ИИ могут руководствоваться этическими принципами справедливого распределения благ[123] или солидарности [124].

При развертывании ИИ в рамках деятельности в целях развития или гуманитарной деятельности цель состоит не просто в том, чтобы предотвратить принятие мер со стороны регулятивных органов или снизить риск начала судебного разбирательства за счет соблюдения требований и принципов. Деятельность в целях развития или гуманитарная деятельность в целом не предполагает большого числа потенциальных ситуаций, когда в отношении ее субъектов можно применить меры принуждения к соблюдению нормативных актов или меры надзора. Это логично, ведь задача ее субъектов обычно заключается в том, чтобы существенно улучшить качество жизни и повысить благополучие членов целевых сообществ. Если ИИ не позволяет обеспечить соблюдение прав лиц, затрагиваемых его

120 На основании консультаций, проведенных нами в Женеве. См. также: Josh Cowls and Luciano Floridi, "Prolegomena to a White Paper on an Ethical Framework for a Good AI Society", June 2018, доступно по адресу: https://papers.ssrn.com/ abstract=3198732.

121 Там же, где утверждается, что этические принципы и права человека могут взаимоусиливать друг друга, а этика может охватывать более широкий спектр ситуаций, нежели права человека. См. также: BSR and WEF (примечание 117 выше).

122 Access Now (примечание 19 выше), p. 17.

123 BSR and WEF (примечание 117 выше).

124 Miguel Luengo-Oroz, "Solidarity Should Be a Core Ethical Principle of AI", *Nature Machine Intelligence*, Vol. 1, No. 11, 2019.

воздействием, это может существенно затруднить решение данной первостепенной задачи гуманитарных организаций и организаций в области развития. В силу перечисленных выше причин субъекты деятельности в целях развития или гуманитарной деятельности все активнее работают над тем, чтобы ИИ создавался с учетом прав и этических норм[125].

Принципы и инструменты

Регулирующая основа, опирающаяся на права человека, будет эффективна только в том случае, если интегрировать ее в повседневную практическую деятельность организации. Для этого необходимо создать инструменты и механизмы разработки и использования систем ИИ на каждом этапе жизненного цикла продукта — и для каждого способа его применения. В этом разделе представлен ряд таких инструментов, которые часто упоминались участниками наших консультаций и бесед как полезные или необходимые.

В своей стратегии в отношении новых технологий Генеральный секретарь ООН отметил обязательства Организации как по «наращиванию внутреннего потенциала ООН и расширению использования в ней новых технологий», так и по «поддержанию диалога о механизмах регулирования и сотрудничества»[126]. Группа высокого уровня Генерального секретаря по цифровому сотрудничеству дала схожие рекомендации, призвав активизировать сотрудничество в цифровой сфере для создания стандартов и принципов прозрачности, объяснимости и подотчетности при разработке и использовании систем ИИ[127]. Кроме того, внутри ООН и других международных организаций был предпринят ряд подготовительных действий по разработке этических принципов и практических инструментов[128].

125 См., например, веб-страницу «Проекты» (Projects) на сайте инициативы Генерального секретаря ООН «Глобальный пульс», доступно по адресу: www.unglobalpulse.org/projects/.

126 *ООН*. Стратегия Генерального секретаря в отношении новых технологий, сентябрь 2018 г., доступно по адресу: https://www.un.org/en/newtechnologies/images/pdf/SGs-Strategy-on-New-Technologies-RU.pdf.

127 High-Level Panel on Digital Cooperation, *The Age of Digital Interdependence: Report of the UN Secretary-General's High-Level Panel on Digital Cooperation*, June 2019. (High-Level Panel Report), доступно по адресу: https://www.un.org/en/pdfs/DigitalCooperation-report-for%20web.pdf.

128 ЮНЕСКО опубликовала предварительный набор принципов в области ИИ в 2019 г. и в настоящее время готовит проект инструмента стандартизации для этики применения ИИ. Уточненный вариант первой версии проекта рекомендации был представлен в сентябре 2020 г. Другие учреждения, в том числе Организация экономического сотрудничества и развития (ОЭСР) и Европейская комиссия, также опубликовали свои собственные наборы принципов. OECD, *Recommendation of the Council on Artificial Intelligence*, 21 May 2019; European Commission, *Ethics Guidelines for Trustworthy AI*, 8 April 2019, доступно по адресу: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Комитет министров Совета Европы утвердил Рекомендацию CM/Rec(2020)1 (примечание 107 выше). Совет Европы также изучает возможность принятия правовой основы для разработки, проектирования и применения ИИ на основании стандартов Совета Европы в области прав человека, демократии и верховенства права; см.: Council of Europe, "CAHAI — Ad Hoc Committee on Artificial Intelligence", доступно по адресу: www.coe.int/en/web/artificial-intelligence/cahai.

Внутренние принципы применения ИИ

Подготовка набора принципов применения ИИ, основанного на правах человека и дополненного этическими нормами, может быть полезна для направления работы организации в данной сфере — и, в конечном счете, для внедрения прав человека в ее практическую деятельность. Цель такого «кодекса» должна заключаться в том, чтобы предоставить каждому члену группы указания, позволяющие обеспечить постоянное внимание к потребностям и правам человека на каждом этапе жизненного цикла ИИ. Что еще более важно, принципы также могут быть положены в основу любых инструментов или механизмов обеспечения соответствия, которые будут впоследствии разработаны организацией, в том числе процедур оценки риска и аудита, а также технических стандартов. Принципы должны иметь достаточно общий характер для того, чтобы их можно было применять как указания в новых обстоятельствах, например при появлении ранее не предусмотренного технологического новшества, но в то же время отличаться конкретностью, достаточной для их интеграции в повседневную практическую деятельность организации.

Генеральный секретарь ООН рекомендует разрабатывать технологии ИИ, «которые заслуживали бы доверия, основывались на правах человека, являлись бы безопасными и устойчивыми и способствовали укреплению мира»[129]. Руководящие принципы любой организации должны зиждиться на этих четырех опорах, но в остальном — в зависимости от сути и контекста деятельности организации — возможны значительные вариации в их содержании. В рамках наших консультаций высказывалось предположение о том, что эффективный набор принципов должен основываться на принципах защиты прав человека — истолкованных в контексте ИИ или приспособленных к нему, — а также на вспомогательных этических принципах, которые обеспечивают гибкость, достаточную для разрешения новых сложностей, возникающих по мере развития технологий.

Определение полного набора принципов не входит в цели настоящей статьи, однако все больше специалистов соглашаются с тем, что некоторые задачи требуют отдельного внимания. Три таких задачи — недопущение дискриминации, прозрачность и объяснимость, а также подотчетность —

129 Дорожная карта Генерального секретаря (примечание 1 выше), п. 88. См. также: High-Level Panel Report (примечание 127 выше), Recommendation 3C, pp. 38–39 («Разработка автономных интеллектуальных систем должна осуществляться таким образом, чтобы существовала возможность объяснить их решения, а люди несли ответственность за их использование. При помощи систем аудитов и сертификации следует проверять соответствие автономных интеллектуальных систем техническим и этическим стандартам, которые необходимо разработать, используя многосторонний подход и принцип множества заинтересованных сторон. Решение вопросов жизни и смерти нельзя делегировать машинам... Необходимо улучшить цифровое сотрудничество, при этом всевозможные заинтересованные стороны должны тщательно продумать структуру, а также применение... таких принципов, как прозрачность и непредвзятость в автономных интеллектуальных системах в различных социальных условиях»).

будут подробно проанализированы ниже. Часто упоминаются и другие принципы: антропоцентричное устройство, человеческий контроль или надзор, всеохватность и разнообразие, конфиденциальность, техническая надежность, солидарность, устойчивость, демократичность, надлежащее управление, осведомленность и грамотность, принципы философии убунту и запрет автономных систем оружия летального действия. На рисунке 1 приведена таблица принципов, которые чаще всего включаются в руководства по этике в сфере ИИ, созданная по данным анализа, проведенного в 2019 году сотрудником инициативы Генерального секретаря ООН «Глобальный пульс» Рене Клаусеном-Нильсеном.

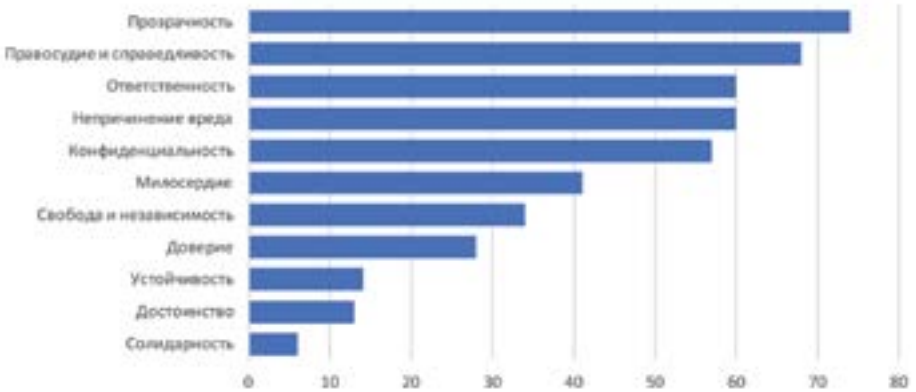

Рисунок 1. Этические принципы, зафиксированные в существующих руководствах по применению ИИ. Анализ выполнен Рене Клаусеном-Нильсеном (инициатива Генерального секретаря ООН «Глобальный пульс») на основании материалов A. Jobin, M. Ienca, and E. Vayena, примечание 36 выше.

Безусловно, само по себе принятие корпоративного этического кодекса не гарантирует отдание приоритета правам человека при разработке организацией инструментов ИИ. Для того чтобы оказывать реальное воздействие, этические принципы должны быть внедрены в ее практическую деятельность. Фундаментальным этапом такого внедрения является принятие стратегического обязательства по соблюдению прав человека на уровне высшего руководства организации. Кроме того, выполнение этого обязательства должно сопровождаться и направляться надлежащими руководящими и надзорными органами и процессами. На следующем этапе внедрения можно предусмотреть трансформацию принципов в технические стандарты, позволяющие проводить контроль качества и аудиторские проверки. К примеру, некоторые эксперты предлагают вводить технические стандарты для обеспечения прозрачности алгоритмов или использовать правила, с помощью которых можно автоматически выявлять результаты алгоритмической обработки данных, потенциально способные обуслов-

ливать несправедливость[130]. Кроме того, этический кодекс следует составлять так, чтобы он обеспечивал информационную основу и способствовал созданию конкретных инструментов и процедур для смягчения риска в области прав человека на каждом этапе жизненного цикла ИИ. Например, кодекс можно внедрить в инструменты осуществления должных экспертиз на предмет соответствия правам человека.

Каждый из вышеперечисленных принципов действительно может иметь критически важное значение, но в рамках наших консультаций мы сосредоточились на трех связанных между собой этических принципах, которые прочно закреплены в МППЧ и требуют дальнейшей проработки и более осторожной реализации: недопущение дискриминации, прозрачность и объяснимость, а также подотчетность. Организациям, применяющим ИИ при оказании гуманитарной помощи, следует разработать стратегии и механизмы, которые обеспечат неизбирательность воздействия систем ИИ; возможность понять и объяснить их решения, как минимум в степени, которая соответствует возникающему риску; а также подотчетность и определение сторон, ответственных за ущерб, нанесенный вследствие работы таких систем. Это особенно важно в тех случаях, когда ИИ применяется в работе по поддержке уязвимых групп населения. Проблемы управления ИИ не исчерпываются этими тремя аспектами, но их можно использовать как отправную точку для дискуссий по поводу того, что отличает ИИ от других технологий и почему он создает уникальные вызовы в области прав человека[131].

Недопущение дискриминации

Один из ключевых принципов, соблюдение которых должны обеспечивать гуманитарные организации, — это недопущение дискриминации. В устройстве систем ИИ обычно отражены текущее соотношение и динамика сил, поэтому при их развертывании возникает риск появления новых факторов неравенства и отношений зависимости — или усугубления уже существующих. Поэтому в качестве базового соображения следует отметить, что любое решение о разработке или развертывании системы ИИ в гуманитарном контексте должно приниматься с учетом комплексной картины потенциального функционирования этой системы в целевой среде и с учетом потенциального воздействия на жизни людей, особенно тех, кто находится в уязвимом положении.

В ходе наших консультаций и научной работы было предложено несколько решений. Прежде всего, для обеспечения неизбирательного использования систем ИИ совершенно необходимо учитывать разнообразие и гарантировать всеохватность. Этот принцип должен распростра-

130 См.: A. Beduschi (примечание 44 выше), где приводится аргументация в пользу технических стандартов, в которых «учтены нормы и принципы прав человека».

131 Подробное подразделение прав и принципов, фигурирующих в ВДПЧ и актуальных для различных аспектов использования систем ИИ, см.: Access Now (примечание 19 выше).

няться на все аспекты разработки и использования ИИ — от донесения различных точек зрения до групп, разрабатывающих и развертывающих системы ИИ, до обеспечения представленности целевых популяций в данных, на базе которых проводится обучение систем. Конструктивные комплексные консультации с представителями затронутых групп чрезвычайно важны для предотвращения ограничительного и дискриминирующего воздействия развернутых решений, в которых используется ИИ.

Во-вторых, остро необходимы наращивание потенциала и обмен знаниями. Специалисты-практики, с которыми мы консультировались, подчеркивают потребность в добросовестном органе-посреднике, который бы координировал обмен знаниями на общемировом уровне и предоставлял бы практические рекомендации относительно решения проблем, связанных с предубеждениями. Такой орган мог бы собирать информацию о передовых практических методах управления ИИ в рамках деятельности в целях развития и гуманитарной деятельности и выявлять области, где эксперименты с использованием ИИ следовало бы запретить. Подобное посредническое учреждение могло бы служить поставщиком новых знаний для организаций, использующих ИИ, которые не способны подвергнуть свои системы ИИ критическому анализу или не располагают необходимыми для этого ресурсами. Многим организациям не к кому обратиться за помощью в диагностике потенциальных проблем дискриминации, которую можно провести на объединенных в пакеты данных путем их критического анализа на предмет предубеждений.

В-третьих, с учетом того, что риск нежелательного дискриминирующего воздействия невозможно свести к нулю, может быть сделан вывод о чрезвычайной рискованности или неочевидности отведения системам ИИ центральной роли в некоторых областях (например, в подготовке окончательных заключений). К таким областям относятся уголовное правосудие, социальное обеспечение и обработка заявок от беженцев или просителей убежища, где ряд пилотных проектов и ситуационных исследований уже выявили факты возникновения проблематичных дискриминационных последствий, которые напрямую влияют на жизни людей. В ходе наших консультаций организациям в таких случаях было предложено прибегнуть к использованию ограничительных запретов и мораториев[132].

Прозрачность и объяснимость

Без прозрачности и объяснимости систем ИИ невозможно обеспечить их подотчетность. При этом абсолютная прозрачность многих систем МО

132 Растет число юрисдикций, в которых запрещены технологии распознавания лиц или запрещено их использование в уголовном правосудии. Тем не менее некоторые организации пока не готовы принять такие ограничения. См.: Chris Klöver and Alexander Fanta, "No Red Lines: Industry Defuses Ethics Guidelines for Artificial Intelligence", trans. Kristina Penner, *Algorithm Watch*, 9 April 2019, доступно по адресу: https://algorithmwatch.org/en/industry-defuses-ethics-guidelines-for-artificial-intelligence/ (здесь один из источников видит причину в отсутствии четких ограничений в этических руководствах ЕС относительно давления на промышленность).

и ГО недостижима[133]. Модели, предусматривающие неконтролируемое обучение, способны классифицировать, сортировать или ранжировать данные на основании набора правил или закономерностей, которые они выявляют самостоятельно. Поэтому создатели модели не всегда могут сказать, как и почему результаты анализа получились именно такими[134]. Это означает, что для использования подобных технологий организациям понадобится тщательно оценить, возможно ли применять непрозрачные или не поддающиеся объяснению системы таким образом, чтобы они способствовали соблюдению прав человека, а не их нарушению (и если да, то как этого добиться).

Прозрачность бывает как минимум двух типов, каждый из которых критически важен для обеспечения подотчетности. Первый тип — техническая прозрачность, то есть прозрачность моделей, алгоритмов и наборов данных, которые формируют ту или иную систему ИИ. Второй — организационная прозрачность, которая связана с доступностью информации по следующим вопросам: используется ли система ИИ для той или иной цели? какой тип системы или ее способностей используется? кто профинансировал создание системы и ввод ее в эксплуатацию? кто ее создавал? кто принимал конкретные решения, определившие ее устройство? кто определял, где она будет использоваться? какие результаты были получены и как они использовались?[135] Несмотря на то что два типа прозрачности связаны между собой, каждый из них требует отдельного аппарата механизмов и стратегий, необходимых для обеспечения прозрачности и объяснимости системы.

133 «Несмотря на то что в настоящее время нет возможности сделать работу систем с использованием МО полностью объяснимой, разработчики в любом случае могут предоставить ценную информацию о том, как функционирует система. Публикуйте простые для понимания справочные материалы на местном языке. Проводите встречи с членами сообщества, чтобы объяснить, как работает инструмент, и дать им возможность задать вопросы и оставить обратную связь. Постарайтесь принять во внимание уровень грамотности и особенности более широкой информационной экосистемы. Эффективный процесс общественного просвещения должен использовать привычные для сообщества способы получения и распространения информации, будь то печатные материалы, радио, очное общение или другие каналы». L. Andersen (примечание 42 выше).

134 См.: “Common ML Problems” (примечание 24 выше).

135 См. Доклад по вопросу о реализации во всех странах экономических, социальных и культурных прав (примечание 44 выше), п. 52, где утверждается, что разрыв в знаниях и понимании между общественностью и лицами, ответственными за принятие решений, может стать «[особенно острой проблемой] в контексте автоматизированных процессов принятия решений, основанных на искусственном интеллекте»; что «подробная и общедоступная информация важна для принятия обоснованных решений и получения согласия затрагиваемых сторон»; и что «полезными могут также быть нормативные положения, требующие от компаний раскрывать информацию в тех случаях, когда системы искусственного интеллекта используются таким образом, что это влияет на осуществление прав человека, и предоставлять результаты соответствующих оценок воздействия на права человека». См. также: L. McGregor, D. Murray and V. Ng (примечание 5 выше), где утверждается, что прозрачность подразумевает наличие информации о том, как и зачем был создан алгоритм, о логике или общей структуре модели, о допущениях, положенных в основу процесса разработки; о способах отслеживания эффективности; об изменениях в самом алгоритме с течением времени; о факторах, имеющих значение для функционирования алгоритма, и об уровне вовлеченности людей.

Для учета и обеспечения принципа прозрачности участники наших консультаций и научной работы предложили в качестве основополагающего принципа концепцию присутствия человека в контуре управления. Присутствие человека в контуре управления — это практический метод, подразумевающий наличие человека, ответственного за принятие решений, в структуре процедуры принятия любого решения с использованием ИИ[136]. Это означает, что даже в тех случаях, когда для достижения максимальной эффективности и надежности прогнозирования применяется ГО, ответственность за практическое применение прогнозов и по возможности за аудит системы, которая их сгенерировала, несет человек[137]. Другими словами, конечная ответственность за принятие решений лежит на человеке, даже если он при этом в значительной степени опирается на результаты анализа, проведенного с помощью алгоритма[138]. При этом присутствие человека в контуре управления имеет смысл только тогда, когда его роль не ограничивается формальным утверждением значимых решений. Кроме того, организациям следует подробно исследовать то, как лица, ответственные за принятие решений, взаимодействуют с системами ИИ, и гарантировать, что первые обладают реальной самостоятельностью в организационном контексте[139].

Подотчетность

Подотчетность систем ИИ дает тем, кто оказался затронут тем или иным действием, возможность потребовать объяснения и обоснования этого действия от его исполнителей и получить соответствующую компенсацию, если это действие привело к нанесению ущерба[140]. Подотчетность может

136 Sam Ransbotham, "Justifying Human Involvement in the AI Decision-Making Loop", *MIT Sloan Management Review*, 23 October 2017, доступно по адресу: https://sloanreview.mit.edu/article/justifying-human-involvement-in-the-ai-decision-making-loop/.

137 См.: L. McGregor, D. Murray and V. Ng (примечание 5 выше), где утверждается, что присутствие человека в контуре управления применяется как мера предосторожности, благодаря которой алгоритмическая система не принимает решение сама, а лишь помогает его принять.

138 «ИИ впечатляет больше всего в тех случаях, когда он способен переработать огромные объемы данных и определить более точные корреляции (выполнить диагностику), оставляя подготовку выводов о причинах и принятие окончательных решений человеку. Такое взаимодействие между человеком и машиной особенно важно для инициатив, оказывающих воздействие на социальный сектор, где значимость этических вопросов велика, а мерилом успеха является способность улучшить жизнь ущемленных в правах людей», Hala Hanna and Vilas Dhar, "How AI Can Promote Social Good", World Economic Forum, 24 September 2019, доступно по адресу: www.weforum.org/agenda/2019/09/artificial-intelligence-can-have-a-positive-social-impact-if-used-ethically/.

139 Один из участников нашего мероприятия в Женеве предложил следующую гипотетическую ситуацию: сотрудник правительственного ведомства использует автоматизированный процесс принятия решений для того, чтобы определить, кого следует лишить родительских прав. Для одного из родителей алгоритм выдает результат «7». Как такой результат повлияет на решение оператора? Будет ли оно зависеть от того, какое у оператора сегодня настроение: хорошее или плохое? Побуждают ли операторов учитывать результат, будь то в институциональном порядке или в рамках межличностного общения (коллеги)? Несут ли они личную ответственность за игнорирование или неучет выводов системы?

140 См.: Edward Rubin, "The Myth of Accountability and the Anti-administrative Impulse", *Michigan Law Review*, Vol. 103, No. 8, 2005.

принимать различные формы[141]. Для технической подотчетности требуется проведение аудиторских проверок самой системы. Социальная подотчетность подразумевает необходимость осведомлять общественность о работе систем ИИ и обеспечивать соответствующий уровень цифровой грамотности людей, который бы позволял им понять воздействие таких систем. Правовая подотчетность подразумевает наличие законодательных и нормативных структур, которые позволяют обеспечить подотчетность лиц, ответственных за отрицательные результаты.

Прежде всего, существует острая потребность в действенных надзорных механизмах для отслеживания и измерения прогресса относительно механизмов подотчетности в различных организациях и условиях. Подобные надзорные механизмы учреждаются на национальном, международном или отраслевом уровне и должны иметь существенный стратегический и технический потенциал, а также реальные возможности по защите прав человека. Еще одной функцией такого механизма или иного специализированного органа может быть проведение сертификации инструментов и систем ИИ или присвоение им знака качества, предусматривающее выдачу сертификатов системам, демонстрирующим высокие результаты в области соблюдения прав человека (по итогам аудита). Такая мера позволит как предупредить потребителей, так и потенциально создать потенциал для формирования партнерств с правительствами, международными организациями, НПО и другими организациями, поддерживающими идею подотчетности и соблюдения прав при применении ИИ[142].

Во-вторых, пока продолжается разработка правовых основ, саморегулирование остается значительным средством стандартизации деятельности частных компаний и других организаций. При этом пользователи и лица, ответственные за разработку политики, могут осуществлять мониторинг деятельности компаний с помощью механизмов подотчетности и следить за тем, чтобы ИТ-отрасль использовала все доступные ей возможности для обеспечения соблюдения прав человека.

В-третьих, ключевым элементом рамочных основ для обеспечения подотчетности ИИ являются эффективные средства правовой защиты. В частности, при отсутствии соответствующих правовых механизмов на национальном уровне возмещения ущерба можно добиться на уровне компании или организации посредством внутренних механизмов подачи жалоб[143]. Сообщения сотрудников о нарушениях также являются важным инструментом выявления случаев злоупотребления и содействия подотчет-

141 См.: UN Human Rights (примечание 68 выше), где определяются новые проблемы подотчетности, связанные с ИИ.

142 См.: High-Level Panel Report (примечание 127 выше), Recommendation 3C, pp. 38–39.

143 Руководящие принципы предпринимательской деятельности в аспекте прав человека ООН (примечание 103 выше), п. 29: «С целью оперативного рассмотрения жалоб и прямого возмещения ущерба предприятиям следует учредить в интересах отдельных лиц и общин, которые могут оказаться жертвами неблагоприятного воздействия, эффективные механизмы рассмотрения жалоб на оперативном уровне или принимать участие в их работе».

ности; необходимо также ввести в действие надлежащие меры предосторожности и создать эффективные информационные каналы, чтобы поощрить и защитить сотрудников, которые сообщают о нарушениях.

Наконец, еще одним критически важным условием подотчетности ИИ является обеспечение надлежащих методов работы с данными. В ходе наших консультаций был выявлен ряд механизмов обеспечения подотчетности при работе с данными, в том числе стандарты качества для достоверных данных и механизмы расширения доступа к качественным данным, такие как обязательный обмен данными.

Инструменты проведения должных экспертиз на предмет соответствия правам человека

Должные экспертизы на предмет соответствия правам человека (ДЭПЧ), проводимые на всем протяжении жизненного цикла систем ИИ, все чаще признаются как незаменимое средство выявления, предотвращения и смягчения риска в области прав человека в контексте разработки и развертывания систем ИИ[144]. Такие экспертизы позволяют определить необходимые меры предосторожности и создания эффективных средств правовой защиты в ситуациях, когда людям наносится ущерб. Благодаря ДЭПЧ центральная роль отводится интересам обладателя прав. Конструктивные консультации с внешними заинтересованными сторонами, в том числе с организациями гражданского общества, а также с представителями потенциально затрагиваемых лиц и групп, которые позволяют избежать включения предубеждений в структуру проекта, являются критически важной составляющей таких должных экспертиз[145].

144 См.: I. Ebert, T. Busch and F. Wettstein (примечание 41 выше). См. также: *Комитет по ликвидации расовой дискриминации*. Общая рекомендация № 36 (2020) о предупреждении расового профилирования со стороны сотрудников правоохранительных органов и борьбе с ним, док. ООН CERD/C/GC/36, 17 декабря 2020 г., п. 66 («С этой целью государствам следует поощрять компании проводить должные экспертизы на предмет соответствия правам человека, которые предусматривают: a) проведение оценок для выявления и анализа любого фактического или потенциального неблагоприятного воздействия на права человека; b) интеграцию этих оценок и принятие надлежащих мер для предупреждения и смягчения выявленного неблагоприятного воздействия на права человека; c) отслеживание эффективности принимаемых ими мер; и d) представление официальной информации о том, какие меры приняты в связи с воздействием на права человека»).

145 См.: Доклад по вопросу о реализации во всех странах экономических, социальных и культурных прав (примечание 44 выше), п. 51. В соответствии с Руководящими принципами предпринимательской деятельности в аспекте прав человека ООН должные экспертизы на предмет соответствия правам человека (ДЭПЧ) — это главные мероприятия, которые следует проводить частным компаниям. Ключевые этапы ДЭПЧ, как определено в Руководящих принципах предпринимательской деятельности в аспекте прав человека ООН, включают 1) выявление ущерба, консультации с заинтересованными сторонами и обеспечение проведения оценки со стороны общественных и частных субъектов (если система будет использоваться государственным органом); 2) предотвращение и смягчение ущерба; и 3) обеспечение прозрачности в том, что касается работы по выявлению и смягчению ущерба. Access Now (примечание 19 выше), pp. 34–35.

Оценки воздействия на права человека

Для того чтобы государства, гуманитарные организации, предприятия и другие субъекты могли выполнить свои обязанности в соответствии с МППЧ, им необходимо определить риск в области прав человека, который исходит от их деятельности. ДЭПЧ обычно опирается на оценку воздействия на права человека (ОВПЧ), по итогам которой выявляются потенциальные и реальные негативные последствия осуществленных и планируемых действий с точки зрения прав человека[146]. При том что ОВПЧ является универсальным инструментом, рекомендованным для всех компаний и секторов в соответствии с Руководящими принципами предпринимательской деятельности в аспекте прав человека ООН, организации все чаще применяют механизм ОВПЧ к ИИ и другим новым цифровым технологиям[147]. Согласно Дорожной карте Генерального секретаря запланирована разработка силами Управления ООН по правам человека общесистемного руководства по ДЭПЧ и оценке воздействия в контексте применения новых технологий[148]. В идеале ОВПЧ должна помогать специалистам-практикам в определении воздействия их мер вмешательства, в которых используется ИИ, с учетом таких факторов, как интенсивность и тип воздействия (непосредственная причина, способствующий фактор или напрямую связанное обстоятельство). Конечной целью ОВПЧ является подготовка ориентиров для принятия решений о целесообразности (и способе) использования того или иного инструмента[149].

Среди других потенциально релевантных инструментов для определения отрицательного воздействия гуманитарной организации на права человека можно назвать оценку последствий обработки данных, в рамках которой проверяется соответствие передовым практическим методам обеспечения конфиденциальности и безопасности данных, а также оценку воздействия алгоритмов, направленную на смягчение уникального риска, возникающего исключительно в связи с работой алгоритмов. В некоторых инструментах сочетаются составляющие разных оценок. Так, например, процедура оценки риска, ущерба и выгод в рамках инициативы «Глобальный

146 Д. Кей (примечание 38 выше), п. 68, где отмечается, что «в ходе разработки и развертывания новых систем искусственного интеллекта, включая развертывание существующих систем на новых глобальных рынках, необходимо готовить оценки воздействия на права человека».

147 Danish Institute for Human Rights, "Human Rights Impact Assessment Guidance and Toolbox", 25 August 2020, доступно по адресу: https://www.humanrights.dk/tools/human-rights-impact-assessment-guidance-toolbox.

148 «Для решения проблем и использования возможностей в области защиты и поощрения прав человека, человеческого достоинства и возможности человека активно воздействовать на мир в эпоху цифровой взаимозависимости Управление Верховного комиссара ООН по правам человека разработает общесистемное руководство по вопросам должной осмотрительности в области прав человека и оценки воздействия при использовании новых технологий, в том числе с учетом информации, полученной от гражданского общества, внешних экспертов и наиболее уязвимых и наиболее пострадавших групп населения». Дорожная карта Генерального секретаря (примечание 1 выше), п. 86.

149 C. Cath *et al.* (примечание 83 выше).

пульс» включает в себя элементы как ОВПЧ, так и оценки последствий обработки данных[150]. С помощью этого инструмента каждый член коллектива, включая технических и нетехнических сотрудников, может оценить и смягчить риск, связанный с разработкой, использованием и развертыванием в конкретных условиях продукта, который функционирует на основе данных. Что немаловажно, упомянутая процедура оценки риска, ущерба и выгод позволяет оценить и выгоды (а не только риск) от использования продукта и таким образом отражает основополагающий принцип уравновешивания интересов, закрепленный в праве прав человека.

Преимущество таких инструментов заключается в том, что их можно адаптировать к новым технологиям. В отличие от регулятивных механизмов и ограничительных запретов, инструменты ДЭПЧ не привязаны к конкретным технологиям или техническим возможностям (например, к технологии распознавания лиц), а разработаны так, чтобы предвосхищать новые технологические возможности и предусматривать пространство для инноваций[151]. Кроме того, грамотно спроектированные инструменты ДЭПЧ учитывают тот факт, что адаптированность процедуры к конкретным условиям является ключевым условием правильной оценки риска в области прав человека. Независимо от того, какой инструмент или комбинация инструментов лучше всего подходят в той или иной ситуации, следует убедиться в том, что в процедуре оценки изначально или в результате обновления учтены виды риска, характерные только для применения ИИ. Кроме того, полезно будет приспособить инструменты к конкретным секторам деятельности в целях развития или гуманитарной деятельности, таким как общественное здравоохранение или реагирование на проблемы беженцев, так как в этих областях вероятно возникновение характерных только для них видов риска.

Важно подчеркнуть, что ОВПЧ должна быть интегрирована в более масштабную процедуру проведения ДЭПЧ. Это позволяет в дальнейшем эффективно смягчать и устранять выявленные риск и последствия в рамках непрерывного процесса. Качество должной экспертизы на предмет соответствия правам человека повышается, когда принцип «знание проблемы и демонстрация решения» (“knowing and showing”) поддерживается организационными мерами управления и действиями руководства, чтобы программное обязательство в области прав человека «[было] органично усвоено, начиная с верхнего руководящего звена предприятия и заканчивая всеми его функциональными подразделениями, которые в противном случае могут действовать, не будучи осведомленными о правах человека или без их учета[152]». Все стороны, задействованные в проекте, должны проводить ДЭПЧ на каждом этапе жизненного цикла продукта. Столь же важно

150 UN Global Pulse, “Risks Harms and Benefits Assessment”, доступно по адресу: www.unglobalpulse.org/policy/risk-assessment/.

151 Element AI (примечание 113 выше), p. 9.

152 Руководящие принципы предпринимательской деятельности в аспекте прав человека ООН (примечание 103 выше), комментарий к принципу 16, с. 17.

обеспечить участие в данном механизме всех уровней организации — от специалистов по работе с данными и инженеров до юристов и руководителей проектов, чтобы при проведении ДЭПЧ учитывались экспертные знания из разных областей.

Пояснительные модели

В дополнение к процедурам оценки организации могут пользоваться пояснительными моделями при появлении новой технологической возможности или способа применения ИИ[153]. Пояснительные модели предназначены для того, чтобы технические сотрудники, лучше понимающие, как работает продукт, доступным языком объяснили это своим коллегам — специалистам нетехнического профиля. Такой подход полезен как для специалистов по работе с данными и инженеров — они могут более тщательно обдумать риск, который заложен в создаваемый ими продукт, так и для нетехнических специалистов, в том числе представителей юридического отдела, отдела по вопросам политики и группы руководителей проекта — им это позволит принять обоснованное решение о том, стоит ли в данной ситуации развертывать соответствующую технологию, и если да, то каким образом. В этом смысле пояснительные модели можно назвать предшественниками описанных выше инструментов оценки риска.

Инструменты должной экспертизы для партнерств

В контексте использования таких инструментов важно сделать следующую оговорку: они могут быть эффективны только в том случае, если применяются к каждому звену цепочки процессов, реализуемых в рамках разработки и развертывания ИИ, в том числе к снабжению. Многие организации, предлагающие новаторские решения в данной области, при создании и развертывании своих продуктов полагаются на партнерства с технологическими компаниями, правительствами и организациями гражданского сектора. Для того чтобы гарантировать соблюдение прав человека и этических стандартов, следует соответствующим образом проверять благонадежность партнерств, которые поддерживают миссии в целях развития и гуманитарные миссии. Проблема секторов деятельности в целях развития и гуманитарной деятельности заключается в том, что большинство инструментов и процессов должной экспертизы (пока) не охватывают надлежащим образом вопросы, связанные с применением ИИ. Для устранения вероятного риска нанесения ущерба в таких процессах и инструментах следует учитывать возникающие технологические сложности. Кроме того, они должны обеспечивать приверженность передовым методам ДЭПЧ, правам человека и этическим стандартам со стороны партнеров, в особенности из частного сектора. Процедура оценки риска, ущерба и выгод, созданная в рамках ини-

153 Участники наших консультаций в Женеве использовали термин «пояснительные модели», несмотря на то, что он пока не является широко распространенным.

циативы Генерального секретаря ООН «Глобальный пульс», может служить образцом такой процедуры[154].

Кроме того, принимая во внимание потенциальный риск, возникающий при использовании систем ИИ недостаточно подготовленными операторами, организации должны следить за тем, чтобы права человека соблюдались всеми партнерами, которые в дальнейшем будут задействованы в реализации проекта. По наблюдению Управления ООН по правам человека, бóльшая часть случаев ущемления прав человека из-за применения ИИ «проявляются при использовании продукта». Это происходит либо намеренно — например, если авторитарное правительство злоупотребляет инструментом и проводит с его помощью незаконный надзор, — либо непреднамеренно, в форме непредвиденной дискриминации или ошибки пользователя. Это означает, что разработчик ИИ не может просто передать инструмент партнеру вместе с инструкцией по его рациональному использованию и в дальнейшем отстраниться от реализации проекта. Пользователь или любая третья сторона, с которой разработчик состоит в отношениях партнерства, должны обязаться проводить тщательную, упреждающую и проверяемую ДЭПЧ на всех этапах жизненного цикла инструмента.

Участие общественности

Взаимодействие с группами населения, затронутыми воздействием инструментов ИИ, является одной из важнейших составляющих эффективной ДЭПЧ. Гуманитарным организациям следует уделять внимание взаимодействию с обладателями прав, затронутыми группами населения, гражданским обществом и другими релевантными заинтересованными сторонами, чтобы получить комплексное и подробное представление о потребностях и правах тех, на кого использование ИИ может оказать воздействие. Для этого требуется инициативная информационная работа, в том числе проведение по мере необходимости публичных консультаций, а также обеспечение доступности удобных каналов коммуникации с затронутыми лицами и сообществами. Согласно рекомендации Специального докладчика Дэвида Кея «до завершения разработки или внедрения нового продукта или службы необходимо провести публичные консультации и заручиться поддержкой общественности в целях проведения конструктивного диалога при активном участии представителей гражданского общества, правозащитников и представителей маргинализованных и недопредставленных конечных пользователей». В некоторых случаях организации могут принять решение опубликовать результаты таких консультаций (вместе с процедурами ОВПЧ)[155].

154 UN Global Pulse (примечание 150 выше). См. также: OCHA, "Guidance Note: Data Responsibility in Public-Private Partnerships", 2020, доступно по адресу: https://centre.humdata.org/guidance-note-data-responsibility-in-public-private-partnerships/.

155 Д. Кей (примечание 38 выше), п. 68.

Аудиторские проверки

Организации, осуществляющие деятельность в целях развития, и гуманитарные организации могут обеспечить внешний и независимый обзор инструментов ИИ — будь то разработанных самостоятельно или предоставленных поставщиками — в форме аудиторских проверок[156]. Возможность проводить такие проверки — необходимое условие прозрачности и подотчетности, а также средство формирования у общественности представления о таких системах и инструмент ее взаимодействия с ними. Поставщики решений из частного сектора традиционно не стремятся сделать свои продукты удобными для аудита, аргументируя это как отсутствием технической возможности, так и вопросами коммерческой тайны. Однако на данный момент предложено большое количество моделей, обеспечивающих разумный компромисс между этими соображениями и основополагающим принципом внешней прозрачности[157]. Обеспечение и стимулирование возможности проводить аудиторские проверки систем ИИ — задача, решить которую способны только государственные регуляторы и разработчики частного сектора, взаимодействуя друг с другом. Субъекты деятельности в целях развития и гуманитарной деятельности могут и должны способствовать обеспечению и закреплению такой возможности[158]. Так, например, доноры или операторы могут сделать возможность проведения аудита требованием, необходимым для принятия заявки на грантовое финансирование к рассмотрению.

Прочие институциональные механизмы

Существует также ряд прочих институциональных механизмов, внедрение которых поможет обеспечить соблюдение прав человека во всех аспектах работы организации. Ранее мы рассматривали один из них — принцип присутствия человека в контуре управления, который предусматривает наличие человека, ответственного за принятие решений, в структуре системы ИИ. Это позволяет добиться того, чтобы ни одно решение, потенциально

156 Д. Кей (примечание 38 выше), п. 55.

157 «Представители частного сектора ставят под сомнение возможность проведения аудиторских проверок в сфере искусственного интеллекта, ссылаясь на необходимость защиты патентованных технологий. В то время как эти сомнения могут быть вполне обоснованны, я разделяю мнение... о том, что отказ поставщика обеспечивать транспарентность в рамках функционирования системы противоречит собственным обязательствам государственного органа в отношении подотчетности, особенно в тех случаях, когда приложение искусственного интеллекта используется учреждением государственного сектора». Там же, п. 55.

158 «При использовании каждого из этих механизмов могут возникать проблемы, особенно в информационной среде, однако компании должны прилагать все усилия в целях создания условий для проведения аудиторских проверок систем искусственного интеллекта. В целях повышения эффективности проверок правительства должны принимать политические решения и законы, которые будут заставлять компании разрабатывать аудируемый код искусственного интеллекта, гарантировать сохранение истории проведения аудиторских проверок и тем самым обеспечивать затронутым лицам возможность получения доступа к более транспарентной информации». Там же, п. 57.

влекущее за собой серьезные последствия, не принималось без контроля и утверждения со стороны человека. Еще одной разновидностью таких механизмов является наблюдательная комиссия по этике, чьи функции аналогичны функциям наблюдательных советов, учреждаемых при научно-исследовательских учреждениях[159]. Комиссия, в состав которой в идеале должны входить сотрудники и технического, и нетехнического профилей, занимается обзором и утверждением любых новых технологических возможностей (а в идеальной ситуации — и любых новых способов их практического применения) до начала развертывания. Для того чтобы стать эффективным превентивным органом, комиссия должна обладать реальными полномочиями по приостановке или отмене проектов, за которой не последует никаких последствий для ее членов. Несмотря на то что наблюдательные комиссии могут использовать представленные выше инструменты ДЭПЧ, обзор проектов с их стороны — это отдельный обзор более высокого уровня, нежели инициативная ДЭПЧ, проводимая на каждом этапе жизненного цикла ИИ. Учреждения также должны рассмотреть перспективу проведения регулярного аудита их практических методов применения ИИ и снабжать своих сотрудников и, при необходимости, общественность резюме отчетов, подготовленных аудиторами. Наконец, в контекстах, в которых риск получения результата, имеющего избирательное воздействие, включает нанесение значительного ущерба основным правам людей, от использования ИИ следует отказаться полностью, в том числе посредством ограничительных запретов.

Наращивание потенциала и обмен знаниями

Задачу по внедрению прав человека и этических принципов в практическую деятельность по разработке мощных и непредсказуемых технологий невозможно решить силами одной, пусть даже крупной, организации. Поэтому существует острая потребность в наращивании потенциала, особенно в государственном секторе и НПО. Это актуально как для организаций, занимающихся развертыванием ИИ, так и для учреждений, осуществляющих соответствующий надзор. У многих органов по защите данных, к примеру, может не хватать ресурсов и возможностей для профессионального и комплексного решения этой задачи[160]. Гуманитарным организациям нередко требуется помощь в применении действующих законов и стратегий к ИИ, а также в выявлении пробелов, которые нужно устранить[161]. Кроме того, персоналу организаций, применяющих ИИ, может потребоваться дополнительная подготовка и образование в аспектах этики и прав человека, связанных с ИИ, и в технических вопросах работы систем, чтобы обеспечить доверие к людям, разрабатывающим и применяющим такие системы (а не только к самой системе).

159 На основании наших консультаций.
160 На основании наших консультаций.
161 Element AI (примечание 113 выше).

Управление ИИ изначально было и остается проблемой международного уровня, поэтому, помимо наращивания потенциала внутри организаций, для эффективного управления ИИ необходимо международное сотрудничество. На международном уровне портал для обмена данными, обеспечить работу которого могут традиционные крупные игроки, такие как ООН, и/или технические организации, такие как Институт инженеров электротехники и электроники (IEEE), мог бы служить как ресурс для моделирования инструментов ДЭПЧ, технических стандартов и других передовых практических методов[162]. Если говорить об уровне стран, то эксперты предлагают правительствам создавать «министерства ИИ» или «центры специальных знаний», которые бы занимались координацией работы, связанной с управлением ИИ, во всех правительственных органах[163]. Создание подобного органа даст каждой стране возможность установить механизмы управления, которые соответствуют ее культурным, политическим и экономическим особенностям.

Наконец, ключевым преимуществом системы прав человека является наличие в ней механизмов отчетности и информационно-пропагандистской деятельности на международном уровне. Организациям следует обратить внимание на международные механизмы защиты прав человека, в том числе релевантные рабочие группы СПЧ и специальных докладчиков, чтобы на их примере исследовать и определять новые виды риска, создаваемые ИИ, и узнавать о передовых методах их смягчения[164].

Заключение

Ряд ситуаций, включая текущую пандемию COVID-19, показал, что ИИ может играть значительную роль в поддержке гуманитарных миссий, если его разработка и развертывание осуществляются с соблюдением принципа всеохватности и прав человека. Чтобы минимизировать риск, возникающий в ходе работы таких систем, и максимизировать выгоды от их использования, следует с самого начала интегрировать в системы ИИ принципы защиты прав человека. В краткосрочной перспективе организации могут предпринять несколько критически важных шагов в данном направлении. Для начала организации, разрабатывающие или развертывающие ИИ в гуманитарных контекстах, могут создать набор принципов, опирающихся на права человека и дополненных этикой, которые будут определять порядок работы с ИИ. В этих принципах должны учитываться конкретные усло-

162 В настоящее время ООН осуществляет ряд программ, которые могут способствовать достижению этой цели, в том числе инициативу ЮНЕСКО по созданию первого инструмента ООН для установления стандартов этики ИИ и планы Генерального секретаря ООН по созданию всемирного консультативного органа по сотрудничеству в области ИИ.

163 Element AI (примечание 113 выше).

164 См.: M. Latonero (примечание 81 выше), где содержится призыв к сотрудникам ООН, расследующим нарушения прав человека, и специальным докладчикам продолжать исследовательскую работу и публичное освещение воздействия систем ИИ на права человека.

вия, в которых работают организации, а потому наборы принципов разных организаций могут отличаться.

Кроме того, чтобы исключить результаты, оказывающие избирательное воздействие, очень важно учитывать разнообразие и гарантировать всеохватность. Начиная с наиболее ранних этапов разработки и вплоть до начала практического применения и последующего наблюдения в реализацию проекта должны быть вовлечены группы разнопрофильных специалистов. Большое значение имеют механизмы, обеспечивающие надлежащую степень технической и организационной прозрачности. Полная техническая прозрачность не всегда достижима, но другие механизмы — в том числе пояснительные модели — дают возможность обучать и осведомлять операторов, затронутые группы населения и другие заинтересованные стороны относительно выгод и риска, связанных с той или иной мерой вмешательства, в которой используется ИИ. Знания дают им возможность внести свой вклад и высказать свое мнение по поводу того, следует ли использовать ИИ в данном случае (и если да, то каким образом), а также конструктивно критиковать способы применения ИИ[165]. Весьма важно обеспечить и наличие механизмов обеспечения подотчетности как для сотрудников, работающих над системами внутри организаций, так и для потенциально затрагиваемых воздействием систем ИИ людей. В более широком контексте необходимо взаимодействие с потенциально затрагиваемыми лицами или группами населения, в том числе посредством публичных консультаций, которое следует стимулировать за счет предоставления удобных каналов связи.

Одно из главных преимуществ построения управления ИИ на правах человека заключается в том, что основные составляющие набора инструментов обеспечения соответствия (в большинстве своем) уже созданы. Практические специалисты в области развития и гуманитарной деятельности должны адаптировать и применять существующие механизмы ДЭПЧ, включая ОВПЧ, оценку воздействия алгоритмов и/или процедуру оценки риска, ущерба и выгод, введенную инициативой Генерального секретаря ООН «Глобальный пульс». Эти инструменты следует использовать на каждом этапе жизненного цикла ИИ — от создания концепции до ее практической реализации[166]. В случаях, когда становится ясно, что в них невозможно учесть новые виды риска, связанные с работой систем ИИ, особенно по мере развития у таких систем более совершенных способностей, инструменты следует оценивать и обновлять[167]. Кроме того, организации могут требовать от технологических партнеров из частного сектора проведения аналогичных ДЭПЧ и воздерживаться от партнерства с поставщиками, соблюдение прав человека которыми невозможно проверить[168]. Взаимодействие с потенциально затрагиваемыми системой лицами

165 Access Now (примечание 19 выше).
166 OCHA (примечание 154 выше).
167 N. A. Smuha (примечание 88 выше).
168 Более подробные указания по поводу ДЭПЧ см. в Руководящих принципах предпринимательской деятельности в аспекте прав человека ООН (примечание 103 выше), принцип 17.

должно быть приоритетом для практических специалистов начиная с наиболее ранних этапов разработки вплоть до начала практического применения и последующего наблюдения. Практические специалисты в области развития и гуманитарной деятельности должны стараться обеспечить возможность проведения аудиторских проверок систем, с которыми они работают. Это нужно, чтобы объяснять решения и процессы, связанные с ИИ, затронутым группам населения и чтобы выявлять и возмещать нанесенный ущерб. Наконец, в рамках любого проекта, основанного на данных, необходимо гарантировать наличие высококачественных данных и использование передовых практических методов обеспечения защиты и конфиденциальности данных.