

Открытые данные для ИИ

Что дальше?



Опубликовано в 2024 г. Организацией Объединённых Наций по вопросам образования, науки и культуры (ЮНЕСКО), 7 place de Fontenoy, 75352 Paris 07 SP, France
и Региональным офисом ЮНЕСКО в Алматы.

© UNESCO, 2024
ISBN: 978-92-3-400074-1



Данная публикация предлагается в открытом доступе под лицензией Attribution-Share Alike 3.0 IGO (CC-BY-SA 3.0 IGO) (<http://creativecommons.org/licenses/by-sa/3.0/igo>). Используя содержание данной публикации, пользователи соглашаются с правилами пользования Репозитория открытого доступа ЮНЕСКО (<https://www.unesco.org/ru/open-access/cc-sa>).

Название на языке оригинала: «Open data for AI: what now?». Опубликовано в 2024 г. Организацией Объединённых Наций по вопросам образования, науки и культуры и Региональным офисом ЮНЕСКО в Алматы.

Использованные названия и представление материалов в данной публикации не являются выражением со стороны ЮНЕСКО какого-либо мнения относительно правового статуса какой-либо страны, территории, города или района или их соответствующих органов управления, равно как и линий разграничения или границ.

Ответственность за взгляды и мнения, высказанные в данной публикации, несут авторы. Их точка зрения может не совпадать с официальной позицией ЮНЕСКО и не накладывает на Организацию никаких обязательств.

Изображения на Shutterstock не подпадают под действие лицензии CC-BY-SA и не могут быть использованы или воспроизведены без предварительного разрешения правообладателей.

Публикация подготовлена под руководством Зёнке Зише (Soenke Ziesche).

Графический дизайн, дизайн обложек и вёрстка: Cromatik Ltd.

ePub: Cromatik Ltd.

Перевод на русский язык и вёрстка русской версии: Александр Кочетков.

Отпечатано в Казахстане.

КРАТКИЙ ОБЗОР

ОТКРЫТЫЕ ДАННЫЕ — ОБЯЗАТЕЛЬНОЕ УСЛОВИЕ ДЛЯ РАБОТЫ И РАЗВИТИЯ ИИ

Для борьбы с эпидемией проводился общемировой обмен своевременными, актуальными и высококачественными открытыми данными. Это совместное и эффективное начинание может послужить образцом для построения лучшего и более устойчивого будущего для всех.

Наличие необработанных исходных данных является первым важным шагом на пути их обработки и преобразования в полезную информацию. Однако требуемые данные должны обладать определёнными характеристиками, такими как точность, актуальность и достоверность. Кроме того, критически важными аспектами обработки данных являются их удобство поиска, доступность, функциональная совместимость и пригодность для повторного использования (концепция FAIR, от англ. Findable, Accessible, Interoperable, Reusable) кем угодно для любых целей.

Цель настоящих руководящих принципов — донести до государств-членов понимание ценности открытых данных, а также описать порядок их курирования и открытия. В документе указаны конкретные шаги, которые государства-члены могут предпринять для открытия своих данных в три этапа: [подготовка](#), [открытие](#), [осуществление последующих мероприятий](#).

Настоящие руководящие принципы разработаны в ключе Рекомендации ЮНЕСКО об этических аспектах искусственного интеллекта, в которой, помимо прочего, содержится призыв к открытию данных для целей искусственного интеллекта. Эта публикация призвана также сыграть важную роль в поддержке Рекомендаций ЮНЕСКО по открытой науке путём содействия совместному использованию данных, повышению воспроизводимости и прозрачности, развития совместимости данных и стандартов, поддержки сохранения данных и долгосрочного доступа к ним.

120
стран не приняли
политические меры
в отношении
открытых
данных



*«Мысли о войне возникают в умах людей,
поэтому в сознании людей следует
укоренять идею защиты мира»*

Открытые данные для ИИ

Что дальше?

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ	5
БЛАГОДАРНОСТИ	6
РЕЗЮМЕ	7
1. ДАННЫЕ В БОРЬБЕ С ГЛОБАЛЬНОЙ ПАНДЕМИЕЙ — ПРИМЕР ИЗ ПРАКТИКИ	11
1.1 ВВЕДЕНИЕ	12
1.2 ТРУДНОСТИ, СВЯЗАННЫЕ С ИИ	13
1.3 ТРУДНОСТИ, СВЯЗАННЫЕ С ДАННЫМИ	15
1.4 ПУТИ ПРЕОДОЛЕНИЯ	16
1.5 ОБЗОР НАКОПЛЕННОГО ОПЫТА В ОБЛАСТИ ОТКРЫТЫХ ДАННЫХ	17
2. КОНЦЕПЦИЯ И ПЕРСПЕКТИВЫ	21
2.1 ВВЕДЕНИЕ	22
2.2 ОТКРЫТЫЕ ДАННЫЕ	25
2.2.1 ОБЩИЕ СВЕДЕНИЯ	25
2.2.2 ОПРЕДЕЛЕНИЕ, ПРИНЦИПЫ, СХЕМА «ПЯТИ ЗВЁЗД» И ЛИЦЕНЗИИ	27
2.2.3 КЛАССИФИКАЦИЯ ОТКРЫТЫХ ДАННЫХ	29
2.2.4 АРГУМЕНТЫ «ЗА» И «ПРОТИВ» ОТКРЫТЫХ ДАННЫХ	31
2.2.5 СУЩЕСТВУЮЩИЕ ОБЩИЕ ИНИЦИАТИВЫ И РЕСУРСЫ ПО ОТКРЫТЫМ ДАННЫМ	32
2.2.5.1 ИНИЦИАТИВЫ ПО ОТКРЫТЫМ ДАННЫМ, СВЯЗАННЫЕ С ЦУР	34
2.2.5.2 ГОСУДАРСТВЕННЫЕ ИНИЦИАТИВЫ ПО ОТКРЫТЫМ ДАННЫМ	35
2.2.5.3 ИНИЦИАТИВЫ ПО ОТКРЫТЫМ НАУЧНЫМ ДАННЫМ	35
2.2.5.4 ЧАСТНЫЕ ИНИЦИАТИВЫ ПО ОТКРЫТЫМ ДАННЫМ	35
2.2.5.5 ИНИЦИАТИВЫ ПО ОТКРЫТЫМ ДАННЫМ В СВЯЗИ С COVID-19	36
2.2.6 ИНДИКАТОРЫ И «БАРОМЕТРЫ» В ОБЛАСТИ ОТКРЫТЫХ ДАННЫХ	38
2.3 ИИ И ОТКРЫТЫЕ ДАННЫЕ	39
2.3.1 ОБЩИЕ СВЕДЕНИЯ	39
2.3.2 СИНЕРГИЯ МЕЖДУ СФЕРАМИ ИИ И ОТКРЫТЫХ ДАННЫХ В ИНТЕРЕСАХ РЕШЕНИЯ ГЛОБАЛЬНЫХ ПРОБЛЕМ	40
2.3.3 ПОТЕНЦИАЛЬНЫЙ РОСТ СВЯЗАННЫХ С ИИ РИСКОВ ИЗ-ЗА ИСПОЛЬЗОВАНИЯ ОТКРЫТЫХ ДАННЫХ	42
ДРУГИЕ ТЕМЫ	43
3. РУКОВОДЯЩИЕ ПРИНЦИПЫ	47
3.1 ПРЕДИСЛОВИЕ	48
3.2 ПОДГОТОВКА	49
3.3 ОТКРЫТИЕ ДАННЫХ	51
3.4 ОБЕСПЕЧЕНИЕ ПОВТОРНОГО ИСПОЛЬЗОВАНИЯ И УСТОЙЧИВОСТИ	53
ЗАКЛЮЧЕНИЕ	56
4. БИБЛИОГРАФИЯ	57

ПРЕДИСЛОВИЕ

Технологические процессы, производственные циклы и связанные с ними научные исследования генерируют большие объёмы данных в различных форматах и в различном порядке. Эти факторы приводят к трудностям в обработке данных и их использовании в системах генеративного ИИ. По мере роста роли искусственного интеллекта (ИИ), наличие и возможность доступа к данным становятся как никогда важными.

ЮНЕСКО проводит исследования в области искусственного интеллекта, чтобы сформулировать идеи и практические решения для содействия цифровой трансформации и построению инклюзивных обществ знаний. Мы создали эти руководящие принципы, чтобы способствовать развитию общего понимания темы открытости данных, её преимуществ и проблем. Цель — проинформировать читателей, в частности государства-члены ЮНЕСКО, о важности открытых данных для сферы ИИ и помочь государствам-членам в принятии соответствующих мер путём выполнения ряда шагов.

В настоящей публикации даётся определение типа и области применения экосистемы данных и переосмысливается концепция данных «FAIR» (англ. Findable, Accessible, Interoperable, Reusable) применительно к сфере ИИ.

Далее в документе перечислены двенадцать конкретных шагов, которые могут предпринять государства-члены для того, чтобы сделать свои данные открытыми и доступными в формате FAIR, с разбивкой на три этапа: 1) подготовка; 2) открытие данных; и 3) последующие мероприятия по повторному использованию и обеспечению устойчивости. Каждый этап состоит из четырёх шагов.

Я надеюсь, что эта публикация послужит призывом к действию, побуждая государства-члены не только поддерживать открытость качественных данных, но и внедрять культуру открытости данных, чтобы стимулировать применение технологий ИИ.

Тауфик Джеласси

Заместитель Генерального директора по вопросам коммуникации и информации

ЮНЕСКО

БЛАГОДАРНОСТИ

ЮНЕСКО благодарит ведущего автора г-на Зёнке Зише за то, что он взял на себя труд по подготовке данной публикации.

ЮНЕСКО также искренне благодарит основную группу, курировавшую подготовку данного документа, а именно:

профессора Джона Шоу-Тейлора,

профессора Марию Фасли,

г-на Давора Орлича,

г-на Алеша Вершича,

г-на Бхану Неупане.

Рецензентами данной публикации выступали:

г-жа Мариэльза Оливейра, директор Департамента партнёрства и мониторинга операционных программ, Сектор коммуникации и информации, ЮНЕСКО,

г-н Жако дю Туа, руководитель Секции универсального доступа к информации, Сектор коммуникации и информации, ЮНЕСКО,

г-жа Дороти Гордон, председатель IFAP, ЮНЕСКО,

г-жа Эльза Эстевес, кафедра ЮНЕСКО по обществам знаний и цифровому управлению в Национальном университете Юга и действительный профессор Национального университета Ла-Платы, Аргентина

г-жа Ирина Кучма, менеджер программы «Открытый доступ» в НКО Stichting eIFL.net.

ЮНЕСКО также благодарит коллег из Сектора коммуникации и информации, Сектора естественных наук, Сектора культуры, Сектора образования, Сектора социальных и гуманитарных наук, Отдела гендерного равенства и Департамента Африки, входящих в состав Целевой группы ЮНЕСКО по ИИ, за поддержку в подготовке этого документа.

РЕЗЮМЕ

Хотя пандемия COVID-19 стала глобальным кризисом, практически затормозившим развитие мира, она также позволила объединить национальные и международные усилия — правительств и директивных органов, академического и научного сообщества и широкой общественности — для решения поставленных ею задач. Для борьбы с пандемией с самого начала были необходимы не только своевременные, актуальные и качественные, но и открытые данные, которые можно было бы использовать для научных исследований, а также для разработки мер вмешательства и политики. Обмен большим объёмом открытых данных, связанных с COVID-19, оказал значительное влияние, в том числе на развитие направлений применения искусственного интеллекта и продемонстрировал потенциал использования данных для решения других глобальных проблем. Действительно, это совместное и эффективное начинание может послужить образцом практики сбора достоверных данных, а также реализации предложения, изложенного в документе Генерального секретаря ООН «Наша общая повестка дня», в котором содержится призыв к модернизации деятельности ООН в соответствии с требованиями перемен, включая данные и стратегическое прогнозирование.

Наличие необработанных исходных данных является первым важным шагом на пути их обработки и преобразования в полезную информацию. Однако требуемые данные должны обладать определёнными характеристиками, такими как точность, актуальность и достоверность. Кроме того, критически важными аспектами обработки данных являются их удобство поиска, доступность, функциональная совместимость и возможность повторного использования (концепция «FAIR», от англ. Findable, Accessible, Interoperable, Reusable) кем угодно для любых целей. Согласно Рекомендации ЮНЕСКО по открытой науке (2021b), к таким данным относятся «данные открытых исследований, включающие, в частности, цифровые и аналоговые данные, как необработанные, так и обработанные, с сопутствующими метаданными, числовые данные, текстовые записи, изображения, аудиозаписи, протоколы, коды анализа и рабочие процессы, которые могут открыто и многократно использоваться, сохраняться и повторно распространяться любым лицом при условии указания источника» (UNESCO, 2021b, p. 9). Помимо того, что открытые данные могут свободно использоваться, изменяться и распространяться любым человеком в любых целях¹, они могут способствовать укреплению доверия, созданию условий для повторного использования и инноваций, а также подотчётности.

В настоящее время с помощью автоматических процессов и средств, включая датчики, собирается огромное количество данных об окружающей среде, промышленности, сельском хозяйстве и здоровье людей. Такие данные могут быть легко доступны, но при этом потенциально слишком объёмны, чтобы человек мог их эффективно обрабатывать или анализировать; тем не менее они могут служить исходными данными для систем искусственного интеллекта. Методы ИИ и научного анализа данных продемонстрировали широкие возможности по анализу больших объёмов данных, что в настоящее время наглядно показывают системы генеративного ИИ, и помогают выявлять ранее неизвестные скрытые закономерности для получения полезной информации в режиме реального

1 <https://opendefinition.org/>

времени. Однако многие современные системы ИИ работают на основе закрытых наборов данных, а данные, отвечающие критериям открытых данных, могли бы принести ещё больше пользы системам ИИ и смягчить потенциальные опасности, связанные с отсутствием справедливости, подотчётности и прозрачности.

Цель настоящих руководящих принципов — донести до государств-членов ценность открытых данных, а также описать порядок их курирования и открытия. Государствам-членам рекомендуется не только поддерживать открытость качественных данных, но и принимать меры по использованию технологий ИИ и содействовать наращиванию потенциала, подготовке и обучению в этой области, включая инклюзивные открытые данные, а также грамотность в области ИИ.

Публикация подготовлена на основе обширного обзора литературы и консультаций с заинтересованными сторонами, с последующим рецензированием. В документе изложены конкретные шаги, которые могут помочь государствам-членам в обеспечения открытия своих данных, разделённые на три этапа: 1) подготовка; 2) открытие данных; и 3) последующие мероприятия по повторному использованию и обеспечению устойчивости; каждый этап состоит из четырёх шагов. Этап подготовки содержит рекомендации государствам-членам по подготовке к открытию данных и включает следующие предлагаемые шаги: разработка нормативно-правовых актов в области открытых данных, сбор и накопление качественных данных, развитие потенциала в области открытых данных и обеспечение готовности данных к применению средствами искусственного интеллекта. Этап открытия данных включает в себя следующие шаги: выбор наборов данных для открытия, юридическое открытие наборов данных, техническое открытие наборов данных и формирование культуры открытых данных. Этап последующих мероприятий по повторному использованию и устойчивому развитию включает в себя следующие шаги: поддержка участия граждан, поддержка международного участия, поддержка полезного применения ИИ и поддержание высокого качества данных.

Настоящие руководящие принципы разработаны в ключе Рекомендации ЮНЕСКО об этических аспектах искусственного интеллекта, в которой, помимо прочего, содержится призыв к использованию открытых данных (UNESCO, 2021d). Если государства-члены ЮНЕСКО будут следовать этим руководящим принципам и открывать свои данные на устойчивой основе, а также создавать потенциал и инициативы в области культуры открытых данных, то смогут использовать те же принципы для решения других задач на глобальном, региональном и национальном уровнях. Несмотря на то, что в настоящем документе представлены конкретные шаги, которые могут помочь государствам-членам в их усилиях по созданию, открытию, предоставлению и использованию данных, использование данных и средств ИИ требует решения проблем и вопросов в целом ряде областей, включая вопросы конфиденциальности, этики и наращивания потенциала. ЮНЕСКО будет оказывать государствам-членам дальнейшую поддержку в реализации этих руководящих принципов, а также готовить специализированные доклады, посвящённые конкретным областям.



1. ДАННЫЕ В БОРЬБЕ С ГЛОБАЛЬНОЙ ПАНДЕМИЕЙ — ПРИМЕР ИЗ ПРАКТИКИ

1.1. ВВЕДЕНИЕ

Пандемию COVID-19 можно охарактеризовать как «наиболее значительный глобальный кризис в области здравоохранения со времён пандемии гриппа 1918 года» (Cascella et al., 2022). К ноябрю 2021 года в мире было зарегистрировано более 251 млн. подтверждённых случаев заболевания. Пять миллионов человек умерли от COVID-19², и значительное число страдает от сохраняющихся симптомов (Raveendran et al., 2021). Практически всё население мира подверглось воздействию, в том числе и длительному, таких мероприятий, как изоляция, что имело и психологические последствия (Odrizola-González et al., 2020; Panchal et al., 2020). Кроме того, пандемия оказала значительное влияние на системы образования (UNESCO, 2021a), а также на экономику³.

Если 30 января 2020 г. ВОЗ объявила COVID-19 чрезвычайной ситуацией в области здравоохранения, вызывающей международную озабоченность, а 11 марта 2020 г. — пандемией, то учёные уже 29 января 2020 г. опубликовали письмо и призвали к открытому обмену данными о последовательности вируса SARS-CoV-2⁴. 16 марта 2020 г. специалисты по обработке данных призвали к созданию инфраструктуры и экосистемы данных, пригодной для борьбы с пандемиями⁵. 30 марта 2020 г. Генеральный директор ЮНЕСКО г-жа Одри Азуле обратилась к правительствам и выступила за научное сотрудничество и включение методов открытой науки в их исследовательские программы⁶. В апреле 2020 г. в рамках программы «Открытая наука» (Open Science) было принято решение об обеспечении открытости данных. В апреле 2020 года было объявлено о начале реализации программы «Открытое обязательство о борьбе с COVID» (Open COVID Pledge), призывающей организации всего мира предоставлять свободный доступ к своим патентам и интеллектуальной собственности, которую подписали многие стороны, в том числе и крупные корпорации⁷.

Пандемия COVID-19 также послужила основой для разработки международного нормативного документа по открытой науке в виде Рекомендации ЮНЕСКО (UNESCO, 2021b). Рекомендация призывает правительства прилагать усилия для развития потенциала в области открытой науки и открытых исследовательских данных, как необработанных, так и обработанных, и сопутствующих метаданных, а также числовых данных, текстовых записей, изображений и аудиозаписей, протоколов, кодов анализа и рабочих процессов.

Для борьбы с пандемией COVID-19 с самого начала были необходимы своевременные, актуальные и качественные данные. Настоящее тематическое исследование показывает, что помимо перечисленных характеристик решающее значение имеет также открытость данных. Во-вторых, обмен действительно большим количеством открытых данных, связанных с COVID-19, имел значительное влияние. Можно провести различие между тем, используются ли имеющиеся данные людьми или для применения в системах ИИ. Ниже описаны некоторые варианты применения технологий ИИ, проблемы, а также выводы для всего движения за открытые данные.

2 <https://covid19.who.int/>

3 <https://www.statista.com/topics/6139/covid-19-impact-on-the-global-economy/#dossierKeyfigures>

4 <https://www.covid19dataportal.org/support-data-sharing-covid19>

5 <https://thegovlab.org/static/files/publications/ACallForActionCOVID19.pdf>

6 <https://en.unesco.org/news/unesco-mobilizes-122-countries-promote-open-science-and-reinforced-cooperation-face-covid-19>

7 <https://opencovidpledge.org>



1.2. ТРУДНОСТИ, СВЯЗАННЫЕ С ИИ

Для борьбы с пандемией в различных аспектах были разработаны методы машинного обучения⁸ с использованием технологий искусственного интеллекта. Для города Валенсия (Испания) группой исследователей была разработана модель ИИ на основе данных, которая стала победителем конкурса 500k XPRIZE Pandemic Response Challenge (Lozano et al., 2021). Вычислительная эпидемиологическая модель основана на открытых данных и предназначена для прогнозирования уровня инфицирования COVID-19, а также для назначения планов немедикаментозного вмешательства. Другие системы ИИ применя-

⁸ Под машинным обучением понимается способность компьютерной системы адаптироваться к новым условиям, обнаруживать и экстраполировать закономерности (Russell and Norvig, 2015).

лись к имеющимся данным по COVID-19 для решения таких задач, как диагностика, прогноз течения, локализация и мониторинг заболевания, разработка лекарственных препаратов и вакцин и методов лечения, а также прогнозирование динамики пандемии⁹. Центр оперативно-аналитической информации о пандемиях и эпидемиях ВОЗ намерен использовать технологии ИИ для расширения возможностей прогнозирования и раннего оповещения о будущих эпидемиях и пандемиях¹⁰.

Системы искусственного интеллекта сыграли важную роль в некоторых областях исследований, например, в разработке вакцин¹¹. Однако, несмотря на большое количество диагностических и прогностических инструментов, большинство из них по разным причинам оказались несовершенными¹². Большая часть этих проблем связана с тем, что исходные данные были низкого качества, не стандартизированы, неправильно маркированы и часто поступали из неизвестных источников¹³. Кроме того, данные часто не были инклюзивными, например, не представляли должным образом меньшинства, поэтому были необъективными. Все это в совокупности привело к тому, что модели, обученные на этих данных, не смогли дать точных результатов. Другая проблема заключалась в том, что зачастую ни данные, ни обучающие модели не предоставлялись в общий доступ, поскольку у академических исследователей обычно мало карьерных стимулов для этого.



9 См. обзор: Harrus & Wyndham (2021), Hussain et al. (2020) и Khemasuwan & Colt (2021).

10 <https://pandemichub.who.int/>

11 См. Ong et al. (2020).

12 См. обзор: Heaven (2021), von Borzyskowski et al. (2021), Wynants et al. (2020) и Roberts et al. (2021).

13 См. обзор: OECD (2020) и <https://opendatawatch.com/whats-being-said-resource/data-in-the-time-of-covid-19/>

1.3. ТРУДНОСТИ, СВЯЗАННЫЕ С ДАННЫМИ

Отмечаются пробелы в данных, особенно в развивающихся странах и среди некоторых подверженных риску и уязвимых групп населения (Milan & Treré, 2020)¹⁴, а также проблемы, связанные с отсутствием разбивки данных, особенно по полу¹⁵. Эти проблемы, помимо прочего, связаны с нехваткой финансирования, с которой столкнулись более 60% стран с низким уровнем дохода и уровнем дохода ниже среднего при финансировании сбора данных и статистики по COVID-19¹⁶.

В целом, данные часто не соответствуют так называемым принципам FAIR, не являются доступными, функционально совместимыми и пригодными для повторного использования; аббревиатура также интерпретируется как «Federated, AI-Ready», «объединённые, готовые к применению в ИИ», чтобы проиллюстрировать важность того, что данные могут быть использованы системами искусственного интеллекта¹⁷. В большинстве случаев нет чёткого разграничения, предназначены ли данные для использования в работе искусственного интеллекта для борьбы с пандемией или нет. Другими словами, являются



14 Также см. <https://www.npr.org/sections/goatsandsoda/2021/03/15/977455005/covid-19-data-is-missing-a-lot-of-people-and-raising-a-lot-of-questions?t=1622558115396>.

15 <https://data2x.org/tracking-the-gender-impact-of-covid-19/>

16 <https://paris21.org/news-center/news/press2020-under-covid-19-worrying-stagnation-funding-despite-growing-data-demand>

17 <https://www.go-fair.org/implementation-networks/overview/vodan/>

ли просто открытые данные и открытые данные для ИИ взаимозаменяемыми? Если да, то какие это влечёт последствия?

Ещё одной проблемой является обеспечение баланса между открытием данных по COVID-19 и соблюдением права на неприкосновенность частной жизни¹⁸, поскольку имели место случаи утечки данных пациентов с COVID-19¹⁹.

Пандемия COVID-19 также сопровождалась дезинформацией — явлением, которое ЮНЕСКО (2020) назвала «дезинфомедией». Один из аспектов этого явления заключается в том, что скептики в отношении COVID-19 создавали визуализации данных, якобы показывающие, что ответные меры правительств на пандемию были непропорциональными, а факты и серьёзность пандемии были либо занижены, либо завышены. Такая дезинформация часто распространялась в социальных сетях (Lee et al., 2021).

1.4. ПУТИ ПРЕОДОЛЕНИЯ

С одной стороны, пандемия COVID-19 объединила мир для решения этой глобальной проблемы. Благодаря своевременному открытому обмену данными, сотрудничеству и прогнозированию были достигнуты прорывы, что позволило в рекордно короткие сроки разработать средства диагностики и вакцины. С другой стороны, пандемия усугубила поляризацию не только в отношении открытого обмена данными²⁰, но и в отношении справедливого распределения вакцин, а также недоверия к вакцинации — вопросы, выходящие за рамки настоящего доклада. В целом, изложенные выше положительные аспекты сотрудничества и эффективности должны быть взяты на вооружение и служить образцом для сбора достоверных данных.

Однако следует также отметить, что большинство инициатив по открытым данным носили ситуативный характер и не были надлежащим образом скоординированы, поскольку мир не был готов к пандемии. Поэтому важным уроком является разработка нормативно-правовой базы и моделей управления данными, подкреплённых достаточной инфраструктурой, человеческими ресурсами и институциональным потенциалом для решения описанных выше проблем, связанных с открытыми данными, чтобы лучше подготовиться к пандемиям и другим глобальным вызовам (OECD, 2020). Кроме того, необходимо уточнить взаимосвязь между открытыми данными и ИИ, в том числе определить, какие характеристики открытых данных необходимы для того, чтобы они были «готовы к применению в ИИ».

Как будет показано ниже, этот пример полезен тем, что многие возможности и проблемы, кратко проиллюстрированные в нём, применимы и к открытым данным в целом.

¹⁸ <https://en.unesco.org/covid19/communicationinformationresponse/opensolutions>

¹⁹ <https://www.hrw.org/news/2020/12/15/personal-data-thousands-covid-19-patients-leaked-moscow>

²⁰ См., например: Waltman et al. (2021) о сторонниках совместного использования данных и тех, кто подчёркивает важность патентования.



1.5. ОБЗОР НАКОПЛЕННОГО ОПЫТА В ОБЛАСТИ ОТКРЫТЫХ ДАННЫХ

- ❑ Необходимо разработать принципы управления данными, сотрудничества и обмена данными для научных исследований, а также для государственных учреждений, хранящих или обрабатывающих данные, связанные со здоровьем населения, при обеспечении конфиденциальности данных путём их анонимизации. В частности, это касается сотрудничества между исследователями ИИ и врачами-клиницистами.
- ❑ Государственным служащим, работающим с данными, которые являются или могут стать актуальными в связи с пандемиями, может потребоваться обучение, чтобы осознать важность таких данных, а также необходимость делиться ими.
- ❑ Необходимо собирать и накапливать как можно больше качественных данных. Признаками «качественных» данных являются точность, актуальность и полнота; таким образом, данные должны поступать из различных заслуживающих доверия источников, обязательно этичных, т. е. не должны включать наборы данных с предвзятым и вредоносным содержанием, и должны собираться только с согласия и без нарушения конфиденциальности частной жизни граждан. Данные о людях, в идеале, должны быть дезагрегированы, где это уместно, по уровню дохода, полу, возрасту, расовой и этнической принадлежности, миграционному статусу, инвалидности и географии.

ческому положению (Inter-Agency and Expert Group on Sustainable Development Goal Indicators, 2017). Кроме того, пандемии обычно представляют собой быстро развивающиеся процессы, поэтому важное значение имеет постоянное обновление данных.

- Такие характеристики данных особенно необходимы для улучшения неадекватных диагностических и прогностических инструментов ИИ в будущем. Кроме того, могут быть учтены данные, которые кажутся незначительными, поскольку сильной стороной систем машинного обучения ИИ является обнаружение в данных соответствующих закономерностей и корреляций, которые человек не в состоянии распознать.
- Требуются усилия для преобразования соответствующих данных в машиночитаемый формат, что включает в себя курирование собранных данных, т. е. их очистку и маркировку.
- Необходимо открыть широкий спектр данных, связанных с пандемиями, в соответствии с принципами FAIR, что подразумевает их удобство поиска, доступность, функциональную совместимость и возможность повторного использования; аббревиатура также интерпретируется как «Federated, AI-Ready» — «объединённые, готовые к применению в ИИ».
- Целевая аудитория открытых данных, связанных с пандемией, включает научные и академические круги, директивные органы и частный сектор в целях разработки соответствующих продуктов, а также общественность, и должна быть проинформирована о доступных данных. Общественность играет важную роль в движении за открытые данные и была вовлечена в борьбу с пандемией COVID-19 через инициативы под общим термином «гражданская наука»²¹.
- Инициативы по созданию открытых данных, связанные с пандемиями, должны быть институционализированы, а не формироваться от случая к случаю, и, таким образом, должны быть положены в основу мер обеспечения готовности к новым пандемиям. Такие инициативы также должны носить инклюзивный характер, объединять различных производителей и пользователей данных.
- Кроме того, полезное использование данных о пандемиях для методов машинного обучения ИИ должно регулироваться с целью предотвращения их неправомерного использования для разработки искусственных пандемий, т. е. биологического оружия, с помощью систем ИИ.

²¹ См. например: <https://citizenscience.org/covid-19/>



2. КОНЦЕПЦИЯ И ПЕРСПЕКТИВЫ



2.1. ВВЕДЕНИЕ

Мир переживает несколько взаимосвязанных кризисов. Помимо продолжающейся пандемии COVID-19 и разрушительных бедствий, ухудшения состояния окружающей среды и утраты биоразнообразия, существуют и другие природные и антропогенные опасности, например, техногенные, в том числе ядерные, аварии, а также угрозы космического происхождения, которые создают усугубляющиеся, каскадные или даже системные риски бедствий. В то же время в мире преобладают неравенство, бедность и голод, которые усугубляются перечисленными выше кризисами и опасностями, в частности, чрезвычайной климатической ситуацией.

Для обеспечения устойчивого развития необходимо проводить постоянное наблюдение за ситуациями в мире, а также измерять и проверять достигнутый прогресс. Для постоянного наблюдения за миром, измерения и проверки прогресса необходимы данные о мире.

Это было показано в рассмотренном выше примере из практики: прогресс в борьбе с пандемией COVID-19 был значительно ускорен благодаря наличию данных. Данные способствовали интерпретации сценариев, разработке и последующей корректировке мер



политики, а также эффективных нефармацевтических, медицинских и вакцинных мероприятий. Кроме того, данные позволили проанализировать государственную политику, что привело к повышению прозрачности и подотчётности.

Если бы соответствующие данные были общедоступны, можно ожидать аналогичного эффекта, что в конечном итоге может способствовать преодолению глобальных кризисов. Общество находится в переломном моменте времени, когда мир объединился для преодоления одного кризиса — пандемии COVID-19, и должен справиться с различными аспектами поляризации, чтобы воспользоваться ещё одной исторической возможностью путём переноса и распространения передового опыта, применявшегося во время пандемии, особенно в сфере обмена открытыми данными, на решение других проблем.

Доступность исходных данных — это только первый шаг. Второй шаг включает в себя очистку данных и отделение необъективных, несогласованных и вредоносных данных, которые нельзя использовать по этическим соображениям. На третьем этапе данные преобразуются в полезную информацию. Кроме того, как было показано в приведённом выше примере из практики, данные могут подаваться в системы искусственного интеллекта, способные выявлять неочевидные закономерности.

Благодаря значительному развитию средств сбора показателей и так называемого Интернета вещей (International Telecommunication Union & CISCO Systems, 2016) измеряется всё больше показателей окружающего мира, процессов в природе и деятельности человека, что создаёт большой объём данных, который слишком велик для человеческого восприятия и анализа без помощи автоматических систем, в том числе систем обработки данных и ИИ.

В то же время многие характеристики мира не измеряются, поскольку либо не разработаны средства сбора этих характеристик, либо они существуют, но не применяются повсеместно, либо существуют другие проблемы, такие как вопросы конфиденциальности данных. В этом заключается трудность — никто не должен выпадать из поля зрения. Кроме того, большая часть собранных данных по ряду причин доступна не всем²².

Представьте себе сценарий, при котором данные по всем 231 показателю ЦУР (целей устойчивого развития), позволяющим отслеживать достижение 169 задач 17 ЦУР, доступны в открытой форме и в режиме реального времени, чтобы каждый мог проверить их в любое время посредством онлайн-платформы²³. Это был бы прекрасный пример доступности актуальных данных.

Очевидно, что для решения глобальных проблем необходимые данные и их источники должны обладать определёнными характеристиками, такими как точность, своевременность, достоверность, полнота и инклюзивность. Однако самих по себе этих характеристик может оказаться недостаточно для решения поставленных задач. Не менее важными аспектами данных являются их удобство поиска (F), доступность (A), функциональная совместимость (I) и пригодность для повторного использования (R) любыми лицами для любых целей. Такие данные соответствующими принципам FAIR, а открытые данные должны отвечать требованиям, позволяющим «свободно использовать, модифицировать и распространять их любыми лицами для любых целей»²⁴.

В центре настоящего доклада находятся открытые данные, поскольку они рассматриваются как необходимое условие для принятия обоснованных планов, решений и мероприятий. Поэтому в докладе утверждается, что государства-члены должны обмениваться данными и информацией, обеспечивая прозрачность и подотчётность, а также возможность для всех использовать эти данные.

Основная цель настоящих руководящих принципов — способствовать всеобщему доступу к информации и знаниям с помощью открытых решений для инклюзивной цифровой трансформации и развития ИИ. В то же время существуют обоснованные опасения по поводу безопасности средств ИИ, которыми нельзя пренебрегать, но которые в силу своей сложности выходят за рамки настоящей публикации (например, Bostrom, 2014; Yampolskiy, 2018). Кроме того, цель документа — повысить осведомлённость о техноло-

22 См. также «Общесистемную дорожную карту по обновлению данных и статистики ООН», разработанную Комитетом главных статистиков системы ООН при участии Комитета высокого уровня по программам: <https://unstats.un.org/unsd/unsystem/documents/Roadmap-Innovating%20UN%20Data%20and%20Statistic.pdf>

23 Хотя с этими целями уже разработаны платформы (<https://unstats-undesapendata.arcgis.com/>, <https://sdg.tracking-progress.org/>, <https://unstats.un.org/sdgs/dataportal>), их применение ограничено значительными пробелами в данных и временными задержками, поскольку страны не имеют возможности собирать показатели в режиме реального времени (например, Sachs et al., 2021).

24 <https://opendefinition.org/>

гиях и инструментах, которые позволяют осуществлять информированное планирование и принятие решений в отношении различных мероприятий, а также осуществлять мониторинг и независимый контроль за ходом работы.

В настоящем докладе утверждается, что главным из множества преимуществ открытых данных является то, что они наглядны и способствуют укреплению доверия и солидарности. Например, решение проблемы изменения климата требует совместных усилий многих заинтересованных сторон, включая граждан всего мира. Необходимым условием этого является то, что все заинтересованные стороны должны не только иметь доступ, но и доверять информации, связанной с изменением климата, которая компрометируется угрозой фальшивых новостей. Тем не менее, если правительства не будут демонстрировать полную прозрачность в отношении информации, связанной с климатом, в своих странах²⁵, то не будут достигнуты ни доверие, ни заинтересованность, ни вовлеченность, ни ответственность сторон. Следовательно, потенциал данных не будет использован в полной мере всеми заинтересованными сторонами, что является настоятельной необходимостью, учитывая огромные проблемы и последствия изменения климата (например, Ratcliffe & Tuzeneu, 2019; Bodor et al., 2020).

Концепция настоящих руководящих принципов заключается в том, чтобы помочь государствам-членам ЮНЕСКО не только в измерении этих аспектов мира, но и предоставлении открытого доступа к данным, которые могут способствовать выработке общедоступных решений (при одновременном соблюдении права человека на неприкосновенность частной жизни), чтобы их можно было обрабатывать и анализировать в интересах общества любым лицом, в том числе с использованием систем искусственного интеллекта.

Методология, использованная при подготовке настоящих руководящих принципов, состояла в проведении тщательного обзора литературы, а также консультаций с заинтересованными сторонами, с последующим экспертным рецензированием.

Остальная часть данной публикации построена следующим образом: в следующем разделе более подробно рассматривается концепция открытых данных. Затем следует раздел об искусственном интеллекте и о том, почему открытые данные особенно важны для работы ИИ. Далее в общих чертах излагаются смежные темы, которые не могут быть рассмотрены в настоящем докладе, но, тем не менее, важны. Во второй части доклада представлены важнейшие действия и шаги в области открытых данных как призыв к действию.

2.2. ОТКРЫТЫЕ ДАННЫЕ

2.2.1. Общие сведения

Технологические достижения последних десятилетий, названные «четвёртой промышленной революцией», привели к беспрецедентному росту объёма и сложности генерируемых данных, а также их сбору в большие массивы (также именуемые «большими данными»). Подсчитано, что в 2020 году цифровая вселенная содержала

²⁵ Дополнительной проблемой является так называемый «эффект перелива», когда действия одной страны оказывают положительное или отрицательное влияние на возможности других стран по достижению ЦУР, особенно связанных с климатом. См. <https://www.unsdsn.org/spillover-effects> и <https://dashboards.sdgindex.org/map/spillovers>.

44 зеттабайта²⁶ данных. Однако на фоне постоянно растущего потока данных возникает ряд проблем, связанных с доступностью достоверных и своевременных данных для целей развития, таких как:

- ❑ данные не собраны;
- ❑ данные собраны, но не распространяются;
- ❑ данные собраны, но разрознены и неизвестны другим сторонам;
- ❑ данные собраны, но они устарели, недостоверны или неточны;
- ❑ данные собраны, но не представлены в функционально совместимых форматах, например, имеются только в бумажном виде;
- ❑ данные собраны, но не помечены соответствующей лицензией.

Огромный объём данных открывает новые возможности, поскольку такие данные, например, могут быть переданы в системы искусственного интеллекта для анализа и дальнейшего использования. Однако даже большие объёмы данных не позволяют составить адекватную картину мира, как хотелось бы, поскольку многие важные данные не собираются, а многие собранные данные недоступны.

Стало очевидным, что данные должны обладать определёнными свойствами, которые делают их полезными для решения глобальных задач, изложенных во введении, такими как их наличие, возможность поиска, доступность, функциональная совместимость, модифицируемость и возможность повторного использования, о чём подробнее будет сказано ниже. Поэтому к заинтересованным сторонам, прежде всего правительствам, а также научному сообществу и частному сектору обращён призыв предоставлять открытые данные. В 2012 г. изобретатель Всемирной паутины Тим Бернерс-Ли в своей статье «Необработанные данные, немедленно!» («Raw data, now!») выразил просьбу предоставлять открытые данные и описал возможности, которые открываются из использования дополнительных данных: «инновации, прозрачность, подотчётность, большая эффективность государственного управления и экономический рост» (Berners-Lee, 2012). Также, бывший Генеральный секретарь ООН Пан Ги Мун признал значимость данных и поручил подготовить доклад «Мир, который имеет значение» («A world that counts»), в котором также подчёркивается ценность открытых данных и рекомендуется наладить партнёрство между международными организациями, правительствами, частными компаниями и организациями гражданского общества для обмена данными и осуществления мониторинга (Independent Expert Advisory Group on a Data Revolution for Sustainable Development, 2014).

В настоящее время формируется «третья волна» использования открытых данных, характеризующаяся акцентом на эффективном повторном использовании данных²⁷ что также подчёркивается в Рекомендациях ЮНЕСКО по открытой науке (UNESCO, 2021b). Предлагаемый в настоящее время Закон ЕС об управлении данными также поддерживает расширенное повторное использование данных государственного сектора²⁸. Другими при-

26 Один зеттабайт = 10^{21} байт. См. <https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/>

27 Первая волна была сосредоточена на вопросах нормативного и законодательного регулирования, а вторая — на пропаганде открытости данных по умолчанию, без конкретной цели. См. <https://opendatapolicylab.org/images/odpl/third-wave-of-opendata.pdf>

28 <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020PC0767&from=EN>

мерами успешной практики являются, например, Альянс по вопросам больших данных по миграции, который организовал виртуальные семинары по проблемам Западной Африки, в которых приняли участие заинтересованные стороны из правительств, международных организаций, научно-исследовательских кругов, гражданского общества и государственного сектора и определили совместные подходы к инновациям в области данных (Big Data for Migration Alliance, 2021), или Директива ЕС INSPIRE, которая привела к созданию единых моделей данных и спецификаций по 34 темам пространственных данных²⁹.

2.2.2. Определение, принципы, схема «пяти звёзд» и лицензии

Если в Рекомендации по открытой науке определение открытых данных не даётся, но уточняется их содержание и условия их использования (UNESCO, 2021b), то определение открытых данных, данное Фондом открытых знаний, является более конкретным и гласит: «Открытые данные и материалы могут свободно использоваться, изменяться и распространяться любым лицом в любых целях»³⁰. Поскольку основное внимание в настоящих руководящих принципах уделяется открытым государственным данным, которые представляют собой «философию, и всё чаще набор политических мер, способствующих прозрачности, подотчётности и созданию ценности путём обеспечения доступности госу-



29 <https://inspire.ec.europa.eu/inspire-directive/2>

30 <https://opendefinition.org/>

дарственных данных для всех³¹, то имеет смысл утверждать, что правительства не только собирают данные и разрабатывают нормы обмена данными, но и выступают в роли поставщиков данных.

Важным дополнением к этому определению являются шесть принципов Хартии открытых данных, представляющие собой «согласованный на глобальном уровне набор желаемых норм публикации данных»: 1) открытость по умолчанию; 2) актуальность и полнота; 3) доступность и удобство использования; 4) сопоставимость и совместимость; 5) содействие улучшению государственного управления и вовлечению граждан; 6) содействие инклюзивному развитию и инновациям³².

Также уже упоминался ещё один важный набор принципов, называемый FAIR, что расшифровывается как удобство поиска (Findable), доступность (Accessible), функциональная совместимость (Interoperable) и пригодность для повторного использования (Reusable) данных. Руководство по реализации принципов FAIR изложено в так называемых Принципах FAIRификации («FAIRification Framework»³³:

- ❑ для обеспечения возможности как ручного, так и машинного поиска, необходимо присвоить данным глобальные уникальные и постоянные идентификаторы, разработать для них подробные метаданные и обеспечить их размещение на доступном для поиска ресурсе;
- ❑ для обеспечения доступности важно, чтобы данные можно было получить по их идентификатору посредством стандартизованного протокола связи, открытого, свободного и универсально применимого и предусматривающего возможность аутентификации и авторизации;
- ❑ для обеспечения функциональной совместимости важно, чтобы данные были представлены на формальном, доступном, совместно используемом и широко применимом языке и были интегрированы с другими данными посредством квалифицированных ссылок;
- ❑ для возможности повторного использования важно, чтобы данные были хорошо описаны метаданными и соответствовали стандартам сообщества, чтобы была предоставлена чёткая и доступная лицензия на использование данных, а также информация о происхождении данных, методах их сбора и обслуживания.

Набор принципов FAIR был дополнен принципами CARE для управления данными коренных народов. Аббревиатура CARE расшифровывается как коллективная выгода (Collective benefit), полномочия по контролю (Authority to control), ответственность (Responsibility) и этичность (Ethics)³⁴. Использование принципов FAIR и CARE является краеугольным камнем всех действий по управлению данными, направленных на установление стандартов обмена данными и их этичного использования.

Если говорить о техническом открытии данных, то Бернерс-Ли разработал схему «пяти звёзд» для категоризации открытия данных³⁵:

31 <https://www.oecd.org/gov/digital-government/open-government-data.htm>

32 <https://opendatacharter.net/principles/>

33 <https://www.go-fair.org/fair-principles/>

34 <https://www.gida-global.org/care>

35 <https://5stardata.info/en/> и <https://www.w3.org/DesignIssues/LinkedData.html>



- ❑ одна звезда: сделать данные доступными в Интернете в любом формате;
- ❑ две звезды: предоставлять данные в структурированном формате (например, в формате MS Excel вместо отсканированной таблицы);
- ❑ три звезды: предоставлять данные в открытом (непроприетарном) формате (например, CSV вместо MS Excel);
- ❑ четыре звезды: использовать URI для обозначения объектов³⁶, чтобы другие пользователи могли указывать на данные;
- ❑ пять звёзд: возможность связать данные с другими данными, чтобы обеспечить контекст.

Следует различать техническое и юридическое открытие данных. С юридической точки зрения, необходимо указывать, как можно (повторно) использовать данные и требуется ли указание авторства. Существует целый ряд лицензий на открытые данные, ярким примером которых являются лицензии Creative Commons³⁷. Касательно важного вопроса о том, можно ли и в какой степени использовать открытые данные для обучения систем искусственного интеллекта, ведутся постоянные дебаты³⁸.

2.2.3. Классификация открытых данных

Помимо приведённой выше схемы «пяти звёзд», открытые данные можно классифицировать и по другим критериям, которые проиллюстрированы ниже в контексте открытых данных, связанных с COVID-19.

³⁶ URI (Universal Resource Identifier) означает универсальный идентификатор ресурса, позволяющий единообразно идентифицировать физический или логический ресурс в мире. Примерами ресурсов могут служить объекты реального мира, включая людей, места и концепции, или информационные ресурсы, включая веб-страницы, документы и книги.

³⁷ <https://creativecommons.org/about/>

³⁸ <https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/>

- Типы данных: преобладают числовые данные (с геометками), однако для исследований COVID-19 важны и другие типы данных, которыми желательно делиться, например, изображения (компьютерная томография и рентген) и звуки (дыхание и кашель) (Shuja et al., 2021)³⁹.
- Источники данных: основными источниками открытых данных, связанных с COVID-19, являются открытые правительственные данные, а также научно-исследовательские данные. Хотя в случае с пандемией COVID-19 этого не было, в будущем желательно, чтобы и частный сектор открывал соответствующие данные.
- Использование данных: предполагаемое использование данных зависит от контекста и спецификаций конкретного проекта, которые могут быть самыми разными. Для пандемии COVID-19 требовались данные, которые позволяют использовать их в следующих целях: 1) чтобы понять, как распространяется коронавирус, и на основе полученной информации разработать меры общественного здравоохранения по его сдерживанию; 2) для борьбы с вирусом, чтобы обеспечить более эффективную разработку вакцины.
- Продукты данных: различные форматы могут определять способ совместного использования данных, например, репозитории (банки) данных, информационные панели, обеспечивающие взаимодействие, или визуализация данных.



³⁹ В рамках других инициатив предлагался открытый доступ к научным статьям, что не является предметом данного исследования. См., например: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov> и <https://www.semanticscholar.org/cord19>

- Целевая аудитория: субъекты, которые, как ожидается, будут использовать данные, и которые, согласно анализу, проведённому ОЭСР и GovLab (2021), распределяются следующим образом: общественность (71%), гражданское общество (68%), государственные органы (41%), научно-исследовательские/академические круги (38%) и частный сектор (22%).

2.2.4. Аргументы «за» и «против» открытых данных

Существуют аргументы «за» и «против» открытых данных. Например, одним из аргументов в пользу открытых данных⁴⁰ является то, что в любом случае не может существовать авторское право на фактические данные и что каждый должен иметь право доступа к ним. Кроме того, характерной чертой демократического общества должна быть прозрачность деятельности государственных органов через открытые данные, что также способствует укреплению доверия к государству. Также, возможность повторного использования, перекомпоновки и комбинирования данных, а также получения из них новых (научных) знаний позволяет вовлекать граждан и создавать новые инновационные услуги и продукты, способствуя тем самым созданию социальной, экономической ценности и пользы для окружающей среды⁴¹. Не только граждане могут получать новые знания из открытых данных, но и системы искусственного интеллекта, которые, как изложено здесь, могут способствовать преодолению мировых кризисов.



Противники открытого доступа к данным утверждают, например, что открытые данные могут нарушить неприкосновенность частной жизни затрагиваемых лиц, которые вправе контролировать раскрытие активно или пассивно собираемых данных о них. В некоторых случаях открытые данные также могут нарушать авторские права и/или права на интеллектуальную собственность, если данные собираются и используются без согласия, указания авторства или выплаты компенсации. Кроме того, во многих случаях сбор, очистка и распространение данных является трудоёмким и дорогостоящим процессом, который необходимо оплачивать, что не отражается в случае открытого доступа к этим

40 См., например: <https://okfn.org/opendata/why-open-data/>, <https://opendatahandbook.org/guide/en/why-open-data/>

41 См., например: <http://opendatatoolkit.worldbank.org/en/essentials.html#uses>

данным. Кроме того, в наборах данных могут присутствовать систематические ошибки данных, возникающие на различных этапах, в том числе на этапах составления отчётов человеком, отбора данных, их маркировки и классификации, вплоть до интерпретации результатов моделирования. Существует также опасение, что данные могут быть использованы со злым умыслом, особенно учитывая, что многие новые технологии, включая ИИ, могут иметь двойное назначение. В этой связи Бостром (Bostrom, 2011, p. 3) дал следующее определение «опасности данных»: «Конкретные данные, такие как генетическая последовательность смертельно опасного патогена или схема создания термоядерного оружия, в случае их распространения создают риски».

2.2.5. Существующие общие инициативы и ресурсы по открытым данным

За последние несколько лет был создан ряд инициатив и ресурсов в области открытых данных, которые далее кратко представлены в хронологическом порядке и по темам.

В 2012 году Тим Бернерс-Ли и Найджел Шедболт основали Институт открытых данных, целью которого является создание экосистемы открытых и достоверных данных⁴².

В 2013 году Департамент ООН по экономическим и социальным вопросам (ДЭСВ) опубликовал «Руководство по открытым правительственным данным для вовлечения граждан»⁴³.

Также в 2013 г. лидеры стран «Большой восьмёрки» одобрили Хартию открытых данных «Большой восьмёрки», пропагандирующую использование данных для стимулирования прозрачности, инноваций и подотчётности. В 2015 г. в кулуарах Генеральной Ассамблеи ООН была представлена уточнённая и более инклюзивная версия Международной хартии открытых данных (Open Data Charter, 2015).

Кроме того, в 2013 году был учреждён Альянс по исследовательским данным, который представляет собой общественную инициативу, выдвинутую Европейской комиссией, Национальным научным фондом США и Национальным институтом стандартов и технологий США, а также Департаментом инноваций правительства Австралии⁴⁴.

В 2014 году была основана глобальная сеть «Открытые данные в интересах развития» (OD4D). Её задачей стала популяризация экосистем открытых данных в развивающихся странах с целью поддержания социальных изменений⁴⁵.

В 2015 году Рабочая группа по противодействию коррупции (ACWG) «Большой двадцатки» опубликовала антикоррупционные принципы открытых данных, отражающие роль открытых данных как инструмента обеспечения прозрачности, подотчётности и доступа к информации⁴⁶.

В 2016 г. в журнале Scientific Data была опубликована статья, в которой объясняется важность находимости, доступности, функциональной совместимости и повторного исполь-

42 <https://theodi.org/>

43 <https://publicadministration.un.org/en/ogd>

44 <https://www.rd-alliance.org/about-rda>

45 <https://www.od4d.net/>

46 <http://www.g20.utoronto.ca/2015/G20-Anti-Corruption-Open-Data-Principles.pdf>



зования данных (Wilkinson et al., 2016). В том же году лидеры стран «Большой двадцатки» на своём саммите одобрили эти так называемые принципы FAIR, которые подробно описаны выше.

В 2018 году Статистическая комиссия Департамента ООН по экономическим и социальным вопросам учредила «Рабочую группу по открытым данным», занимающуюся разработкой принципов, рекомендаций и поддержкой внедрения экосистем открытых данных в государствах-членах⁴⁷.

В 2020 году Генеральный секретарь ООН представил «Дорожную карту по цифровому сотрудничеству» для решения проблем, связанных с новыми технологиями, на пути к созданию более безопасного и справедливого цифрового мира (UN General Assembly, 2020).

Также в 2020 году Координационный совет руководителей системы ООН (2020 г.) одобрил «Общесистемную дорожную карту по обновлению данных и статистики ООН», цель которой — «содействовать улучшению мира посредством своевременных, заслуживающих доверия данных и статистики для всех».

В 2021 году 193 государства-члена ЮНЕСКО приняли «Рекомендацию об этических аспектах искусственного интеллекта», которая стала важной вехой не только потому, что это был первый глобальный нормативный документ по этическим аспектам искусственного интеллекта, но и потому, что документ призывал государства-члены содействовать развитию идеи открытых данных (UNESCO, 2021d).

47 <https://unstats.un.org/open-data/>

➤ 2.2.5.1. Инициативы по открытым данным, связанные с ЦУР

Система глобальных показателей достижения целей в области устойчивого развития охватывает 17 целей, 169 связанных с ними задач и 231 показатель, реализация и мониторинг которых в значительной степени зависят от данных. Проблема заключается в том, что по состоянию на февраль 2022 г.⁴⁸ применительно к 95 показателям многие необходимые данные не собираются странами на регулярной основе. Кроме того, Сакс и др. (Sachs et al., 2021) указали на значительные пробелы в данных, особенно по ЦУР 4 (Качественное образование), 5 (Гендерное равенство), 12 (Ответственное потребление и производство), 13 (Борьба с изменением климата) и 14 (Сохранение морских экосистем), и призвали к дальнейшим инвестициям в укрепление статистического потенциала в соответствующих странах. Также утверждается, что достижению ЦУР помимо таких характеристик данных, как доступность, точность, своевременность и дезагрегированность⁴⁹, также способствует и открытость данных (например, Petrov et al., 2016). В связи с этим Статистическая комиссия ДЭСВ ООН создала «Центр открытых данных по ЦУР»⁵⁰. Кроме того, Сеть решений ООН в области устойчивого развития поддерживает платформу для мониторинга достижения ЦУР⁵¹.



48 <https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/>

49 Дезагрегация данных, хотя и не является предметом рассмотрения в настоящем документе, имеет большое значение для достижения ЦУР. Межведомственная экспертная группа по показателям Целей устойчивого развития (2017 г.) призывает к тому, чтобы показатели ЦУР «были дезагрегированы, где это уместно, по доходам, полу, возрасту, расовой и этнической принадлежности, миграционному статусу, инвалидности и географическому положению, или другим характеристикам».

50 <https://unstats-undesa.opendata.arcgis.com/> and also <https://sdg.tracking-progress.org/>

51 <https://www.unsdn.org/sdg-index-and-monitoring>

➤ 2.2.5.2. Государственные инициативы по открытым данным

Государственные органы производят большое количество данных, которыми они в прошлом практически не делились, и которые тем более невозможно было получить для использования в функционально совместимом формате. Уже существует множество инициатив по предоставлению открытого доступа к данным, выдвинутых национальными правительствами, межправительственными организациями, субнациональными региональными структурами, а также местными органами управления⁵². Партнёрство «Открытое правительство» — это инициатива, поддерживающая страны, включая местные органы власти, в обеспечении прозрачности, включая предоставление открытых данных⁵³. Существуют также региональные инициативы, такие как openAFRICA — «крупнейшая в Африке платформа открытых данных, созданная добровольцами»⁵⁴, Азиатское партнёрство по открытым данным⁵⁵, Латиноамериканская инициатива по открытым данным⁵⁶ и официальный портал европейских данных⁵⁷. В 2021 году завершился второй этап инициативы «Большой двадцатки» по устранению пробелов в данных (International Monetary Fund and Financial Stability Board, 2021).

➤ 2.2.5.3. Инициативы по открытым научным данным

В результате научной деятельности образуются большие объёмы данных. В духе научного подхода и в интересах прогресса эти данные должны быть доступны для совместного использования, особенно если они получены за счёт государственного финансирования. При этом не стоит забывать о проблеме возможной опасности данных, о которой говорилось выше. Понимание важности открытых данных в науке и инициативы в этой области не новы. Например, Комитет по данным Международного совета по науке был создан ещё в 1966 году⁵⁸. Несмотря на то, что принято соглашение об обмене научными данными и работают различные платформы⁵⁹, получить доступ к данным трудно из-за их большого объёма, а также неадекватного управления ими. Эта проблема решается в рамках Руководящих принципов FAIR — инициативы по улучшению управления данными и их рационального использования с упором на автоматизацию поиска данных (Wilkinson et al., 2016). Ещё одной важной вехой стала Рекомендация ЮНЕСКО по открытой науке, принятая Генеральной конференцией ЮНЕСКО в ноябре 2021 года (UNESCO, 2021b).

➤ 2.2.5.4. Частные инициативы по открытым данным

Корпорации также собирают и владеют большими объёмами данных, как правило, с единственной целью увеличить свои доходы. Однако значительная часть этих данных имеет ценность для достижения устойчивого развития. Раньше компании отказывались делиться своими данными, однако теперь появилась бесприоритетная возможность делиться данными: несколько корпораций присоединились к инициативе ООН «Глобаль-

52 См. (обновляемый) перечень: https://en.wikipedia.org/wiki/List_of_open_government_data_sites

53 <https://www.opengovpartnership.org/>

54 <https://africaopendata.org/>

55 <https://www.linkedin.com/company/aodp/>

56 <https://idatosabiertos.org/en/>

57 <https://data.europa.eu/en>

58 <https://codata.org/>

59 См., например: <https://opendatascience.com/>

ный пульс» (UN Global Pulse) под названием «Филантропия данных» и обязались делиться своими данными на благо общества⁶⁰. Многие другие корпорации работают с Учебным и научно-исследовательским институтом ООН в рамках инициативы Программы ООН по применению спутниковой информации в оперативных целях, предоставляя спутниковые снимки в гуманитарных целях⁶¹. Хотя в настоящее время эти данные можно использовать для реализации и мониторинга целей устойчивого развития, у корпораций также есть целый ряд стимулов для предоставления доступа к своим данным, таких как взаимность обмена, исследовательская деятельность, подбор персонала и доступ к новым идеям, улучшение репутации и связей с общественностью, увеличение дохода, соблюдение нормативных требований, социальная ответственность и корпоративная благотворительность (Klein & Verhulst, 2017). В предлагаемом в настоящее время Законе ЕС об управлении данными вводится аналогичный механизм, называемый «альтруизмом в области данных», призванный поощрять и давать возможность частным лицам и компаниям делиться данными на благо общества⁶².

➤ 2.2.5.5. Инициативы по открытым данным в связи с COVID-19

В ответ на призывы к предоставлению открытых данных, связанных с COVID-19, были выдвинуты различные инициативы по обмену данными, начиная от открытых данных государственных органов и заканчивая частными проектами⁶³. Ниже приводятся некоторые примеры, иллюстрирующие разнообразие инициатив.

- Информационная панель по COVID-19 Центра системных наук и инженерии при Университете Джона Хопкинса, показывает количество случаев заболевания, смертей и вакцинированных по странам и является источником информации для общественности и директивных органов⁶⁴.
- Платформа Worldometer предоставляет глобальную динамично обновляемую статистику по COVID-19, получаемую путём анализа, проверки и агрегации данных из более чем 5 тыс. источников в режиме реального времени⁶⁵.
- Европейская платформа данных по COVID-19 содействует обмену и анализу данных с целью ускорения исследований коронавирусов⁶⁶.
- Репозиторий открытых данных GitHub COVID-19 стремится стать крупнейшей эпидемиологической базой данных по COVID-⁶⁷.

60 <https://www.unglobalpulse.org/2011/09/data-philanthropy-public-private-sector-data-sharing-for-global-resilience/>

61 <https://www.unitar.org/sustainable-development-goals/united-nations-satellite-centre-UNOSAT>

62 <https://www.consilium.europa.eu/en/press/press-releases/2021/10/01/eu-looks-to-make-data-sharing-easier-council-agrees-position-on-data-governance-act>

63 См. обзор: Alamo et al. (2020).

64 <https://coronavirus.jhu.edu/map.html>

65 <https://www.worldometers.info/coronavirus/>

66 <https://www.covid19dataportal.org/>

67 <https://github.com/GoogleCloudPlatform/covid-19-open-data>

- На платформе GISAID размещены данные более 450 тыс. вирусных геномов⁶⁸, однако платформу критикуют за то, что пользователи не могут повторно публиковать данные геномов без разрешения (Van Noorden, 2021).
- Nextstrain — проект с открытым исходным кодом для отслеживания эволюции патогена в режиме реального времени⁶⁹.
- Global.health — хранилище данных и платформа визуализации, обеспечивающая открытый доступ к анонимизированным эпидемиологическим данным в построчном формате в режиме реального времени⁷⁰.
- Vivli — платформа для обмена анонимными данными об участниках завершённых клинических испытаний⁷¹.
- Coronavirus Watch — платформа, предоставляющая информацию о новых тенденциях, связанных с COVID-19 во всём мире, в виде карт, визуализации статуса заболевания и информационных панелей⁷².
- Универсальный портал ресурсов по COVID-19 (CURE) поддерживается ЮНЕСКО и представляет собой портал ресурсов с открытым доступом, публикующий последние статьи, панели данных, а также образовательные и учебные материалы по COVID-19⁷³.



68 <https://www.gisaid.org/>

69 <https://nextstrain.org/>

70 <https://global.health/>

71 <https://search.vivli.org/>

72 <https://www.unesco.org/en/articles/corona-virus-media-watch-launched-unescos-international-research-centre-artificial-intelligence>

73 <https://www.goap.info/cure>

2.2.6. Индикаторы и «барометры» в области открытых данных

Кроме того, существует несколько индексов и «барометров» для оценки и ранжирования стран в зависимости от их усилий в области открытых данных, например, такие как:

- ❑ Open Data Barometer (Барометр открытых данных)⁷⁴;
- ❑ Global Data Barometer (Глобальный барометр данных)⁷⁵;
- ❑ Open Data Inventory / Watch (Инвентаризация/обзор открытых данных)⁷⁶;
- ❑ Global Open Data Index (Глобальный индекс открытых данных)⁷⁷;
- ❑ OECD Open Government Data / OURdata Index (Открытые правительственные данные ОЭСР/ Индекс OURdata)⁷⁸;
- ❑ Open Data Maturity in Europe (Индекс зрелости открытых данных в Европе)⁷⁹.



⁷⁴ <https://opendatabarometer.org>

⁷⁵ <https://globaldatabarometer.org/>

⁷⁶ <https://opendatawatch.com/>

⁷⁷ <http://index.okfn.org/>

⁷⁸ <https://www.oecd.org/gov/digital-government/open-government-data.htm>

⁷⁹ <https://data.europa.eu/en/impact-studies/open-data-maturity>

2.3. ИИ И ОТКРЫТЫЕ ДАННЫЕ

2.3.1. Общие сведения

Хотя открытые данные могут использоваться во многих областях, они особенно важны для области искусственного интеллекта, о чём речь пойдёт далее в этой части.

Существуют различные определения ИИ в связи с тем, что само понятие «интеллект» не является однозначным. Легг и Хаттер (Legg and Hutter, 2007, p. 12) дают следующее общее определение: «Интеллект измеряет способность агента достигать целей в широком диапазоне условий». Если под «агентом» в этом определении понимать человека или животное, то это обычный интеллект, а если «агент» — машина, то это искусственный интеллект. Область ИИ можно разделить на взаимодействующие поддисциплины, среди которых Рассел и Норвиг (Russell and Norvig, 2015) перечисляют следующие: обработка естественного языка, представление знаний, автоматизированное формирование рассуждений, машинное обучение, компьютерное зрение и робототехника.

Бурный рост технологий ИИ в последнее время в значительной степени обусловлен масштабными достижениями в области машинного (и особенно глубокого) обучения, которое требует больших массивов данных, хотя некоторые другие поддисциплины ИИ не требуют такого большого объёма данных. Генеративный ИИ — это зонтичный термин для обозначения очень успешно показывающих себя в настоящее время алгоритмов машинного обучения, которые привлекают большое внимание прессы и способны генерировать искусственные цифровые материалы — текст, изображения, аудио и видео, — на основе больших объёмов обучающих данных. Генеративный ИИ подразумевает не просто обработку существующих данных, а создание алгоритмов или моделей, способных генерировать новые материалы, например изображения, видео, музыку или текст. Глубокое обучение — это тип машинного обучения, в котором используются нейронные сети для понимания закономерностей и генерации новых данных. Сценарии применения генеративного ИИ включают создание реалистичных изображений или видео, генерацию ответов на естественном языке, проектирование новых продуктов и создание музыки. Существуют опасения по поводу возможного неправомерного использования генеративного ИИ, например, для создания ложных новостей или «дипфейков» (англ. deepfake). В связи с этим необходимо тщательно проанализировать этические последствия использования генеративного ИИ и разработать правила его ответственного применения. Не последнюю роль играет и то, что качество созданного ИИ контента всё чаще таково, что человек не может отличить, кем он был создан — машиной или человеком, что вызывает опасения и споры.

Таким образом, прогресс в машинном обучении, с одной стороны, строится на развитии инновационных технологий, а с другой — на постоянно растущем изобилии доступных данных, тогда как их недостаток в предыдущие десятилетия сдерживал развитие ИИ.

Однако, говоря о технологиях ИИ и данных, необходимо отметить ряд сложностей, существующих в настоящее время:

- ❑ поток данных настолько велик, что многие данные невозможно проанализировать;
- ❑ существуют инструменты ИИ, для работы которых не хватает данных, поскольку соответствующие данные не были собраны, или не были опубликованы, или разрознены, или устарели, недостоверны или неточны, или находятся в непригодных для обработки форматах, или не размечены;

- более того, в некоторых областях имеются ограниченные, но критически важные данные, но не хватает инструментов ИИ, способных извлекать полезную информацию из таких небольших объёмов данных.

Утверждается, что организации работают тем быстрее, чем быстрее обрабатываются их данные. Однако если скорость обработки данных затруднена из-за проблем с их доступностью, пропускной способностью и задержками, обусловленных устареванием средств обработки данных, то организации могут не достичь желаемых результатов и утратить конкурентное преимущество. Более того, огромный объём потенциально доступных необработанных данных не принесёт пользы, если эти данные не будут преобразованы из ряда различных форматов в пригодную для использования информацию, на основе которой будут приниматься правильные решения.

ИИ способен анализировать большие (или, если нет другого варианта, малые) объёмы данных, выявлять ранее неизвестные или скрытые закономерности и предоставлять полезную информацию в режиме реального времени. Это касается и полезной и корректирующей информации для реализации Повестки дня в области устойчивого развития на период до 2030 года, а также для поиска решений других глобальных проблем. Об этом говорится в Дорожной карте Генерального секретаря ООН по цифровому сотрудничеству, которая призывает укреплять цифровой потенциал для расширения цифрового сотрудничества и инклюзивности (Цель 4), а также поддерживать глобальное сотрудничество в области ИИ (Цель 6) (UN General Assembly, 2020).

2.3.2. Синергия между сферами ИИ и открытых данных в интересах решения глобальных проблем

Работа многих систем ИИ опирается на закрытые данные. Однако признано, что данные, отвечающие критериям открытых данных, могли бы принести дополнительную пользу системам ИИ и снизить такие потенциальные риски, как отсутствие объективности, подотчётности и прозрачности (Davies, 2019). Это понимание отражено в Рекомендации ЮНЕСКО по открытой науке (UNESCO, 2021b), Рекомендации ЮНЕСКО об этических аспектах искусственного интеллекта (UNESCO, 2021d), а также является темой нынешней «третьей волны» открытых данных, поскольку все они поощряют эффективное повторное использование данных.

Это также отражено в рекомендации, содержащейся в докладе Администрации президента США, а именно Национального научно-технического совета Комитета по технологиям, под названием «Подготовка к будущему искусственного интеллекта», где говорится, что «Федеральные агентства должны уделять приоритетное внимание открытым данным для обучения и стандартам открытых данных в области ИИ. Правительство должно уделять особое внимание открытию наборов данных, позволяющих использовать ИИ для решения социальных проблем. Возможные шаги могут включать разработку инициативы "Открытые данные для ИИ" с целью открытия значительного количества наборов государственных данных для ускорения исследований в области ИИ и активизации использования стандартов открытых данных и передовых методов в работе государственных органов, в научных кругах и частном секторе» (Executive Office of the President of the United States of America – National Science and Technology Council, Committee on Technology, 2016, p. 14).

Вместе с тем было признано, что одного обмена существующими данными в доступной, функционально совместимой и пригодной для повторного использования форме

недостаточно, необходимо решать и другие перечисленные ранее проблемы, такие как отсутствие данных, их устаревание, недостоверность, неточность или отсутствие маркировки. Вызывает озабоченность, например, отсутствие данных о маргинализированных группах или неформальных секторах экономики, что, по мнению Access Partnership (2018), является политической проблемой, поскольку системы искусственного интеллекта не будут иметь представления об этих аспектах жизни общества.

Это можно проиллюстрировать и следующим примером: системы перевода между английским и французским языком сегодня достаточно совершенны, поскольку для обучения этих систем доступны большие массивы данных обоих языков, чего нельзя сказать о множестве других языков. Другими словами, успешное развитие машинного обучения ИИ зависит от доступности данных, что следует рассматривать как возможность для всех государств-членов: чтобы разработать инструменты ИИ для принятия справедливых решений, обеспечения общественного блага и т. д., данные не должны быть предвзятыми или дискриминирующими по отношению к определённым территориям, культурам или группам, поэтому необходимо предоставлять и всесторонне учитывать данные о территориях, культурах или группах⁸⁰.



Существуют инициативы, направленные на решение этих проблем, например, FAIR Forward⁸¹ и Альянс «Открытость во благо» (Open for Good Alliance)⁸². Эти инициативы направлены на повышение качества локализованных данных для обучения ИИ с целью усовершенствования локализованных инструментов ИИ. Например, в рамках инициативы FAIR Forward собрано 1200 часов голосовых записей руандийского языка киньяруанда. Проектом FAIR Forward также изучались индийские языки и отмечены проблемы с поиском для них наборов речевых данных общего назначения хорошего качества и в стандартизованном формате (GIZ, 2020). Хотя такие конкретные инициативы можно только приветство-

⁸⁰ Следует отметить, что, хотя для снижения предвзятости и дискриминативности систем ИИ необходимы более разнообразные данные; при этом само по себе увеличение разнообразия данных не является достаточным решением этой сложной проблемы.

⁸¹ <https://www.giz.de/expertise/html/61982.html>

⁸² <https://www.openforgood.info>

вать, порталов открытых обучающих данных для ИИ мало, а общая стратегия в области данных, которая бы решала обе задачи, — как сделать данные доступными для всех и как сделать так, чтобы небольшие наборы данных давали полезную информацию — отсутствует.

Технологии на базе ИИ всё больше проникают во все сферы жизни общества. Таким образом, связь сферы ИИ с гендерной проблематикой становится всё более значимой в борьбе за гендерное равенство. Имеются опасения, что усилия в области гендерной проблематики и открытых данных носят преимущественно разрозненный характер (GODAN, 2018), прежде всего потому, что группы, обсуждающие проблемы женщин, и группы, обсуждающие открытость данных, редко взаимодействуют. Таким образом, реализация открытости данных может стать важным шагом на пути к гендерному равенству и расширению прав и возможностей женщин. В 2022–2023 годах ЮНЕСКО уже взяла на себя обязательство возглавить инклюзивный и комплексный подход к развитию цифровых компетенций, в том числе в области ИИ, чтобы преодолеть разрыв в цифровых технологиях и знаниях. Концепция ЮНЕСКО об универсальности Интернета (ROAM-X, ПОДУ-К от «права, открытость, доступность, участие [многих заинтересованных сторон] и кросс-секторное взаимодействие») выступает за открытость данных для преодоления гендерного цифрового разрыва.

Таким образом, существует множество возможностей для использования ИИ в целях измерения и достижения устойчивого развития при условии сбора соответствующих данных, а также их открытого распространения. Как уже отмечалось, предвзятость является фактором риска при применении средств ИИ, который необходимо снижать путём предоставления как можно более полных наборов данных. В связи с этим государствам-членам предлагается не только поддерживать открытость качественных данных, но и способствовать использованию технологий ИИ, а также содействовать наращиванию потенциала, подготовке и обучению в этой области, включая повышение грамотности в области открытых данных с учётом требований технологий ИИ.

2.3.3. Потенциальный рост связанных с ИИ рисков из-за использования открытых данных

ИИ — это технология двойного назначения, поэтому она несёт в себе не только возможности, но и риски. Важно рассматривать открытые данные не только как фактор рисков в связи с технологиями ИИ. Наборы данных, которые поступают в системы ИИ, могут представлять угрозу безопасности, поскольку их использование может привести к нежелательным результатам или реакциям системы ИИ. Хотя упомянутые выше проблемы, связанные с предвзятостью и дискриминацией, также являются нежелательными, но в основном непреднамеренными, возможны и злонамеренные действия, направленные на системы ИИ, реализуемые при неправомерном использовании открытых данных. Такие атаки получили название «отравленный колодезь» (poison in the well) или «атаки уклонения» (evasion attacks) и используют различные методы (Lohn, 2021). Один из них заключается, например, в простой замене меток в наборах данных, а другой, более сложный метод, нацелен на коллекции изображений и заключается во внедрение вредоносных образцов, содержащих тщательно продуманные и трудно распознаваемые человеком возмущения, вызывающие сбой в работе систем распознавания ИИ (Biggio & Roli, 2018). Появление открытых наборов данных непреднамеренно приводит к повышению риска таких атак, поскольку увеличивается объём данных, потенциально подверженных атакам, и, с другой стороны, величина объём данных снижает шансы выявить манипуляции с ними.

ДРУГИЕ ТЕМЫ

Тематика открытых данных и ИИ тесно связана с рядом других тем и проблем, которые не могут быть подробно рассмотрены в настоящем докладе, но, тем не менее, являются актуальными и перечислены ниже, чтобы дать о них общее представление:

- ❑ **конфиденциальность данных:** многие ценные государственные данные касаются граждан. Поэтому баланс между раскрытием этих данных и неприкосновенностью частной жизни граждан является предметом постоянных дискуссий (например, Scassa, 2019);
- ❑ **этика данных:** эта тема связана с приватностью и «оценивает практику использования данных, способную оказать негативное воздействие на человека и общество — при сборе, распространении и использовании данных»⁸³;
- ❑ **правовые вопросы:** использование данных, как правило, регулируется законодательством. Например, статья 8 Хартии Европейского союза об основных правах (2012 г.) гласит: «Каждый человек имеет право на защиту персональных данных, касающихся его или её»⁸⁴;



©Brian A. Jackson/Shutterstock.com

83 <https://theodi.org/service/consultancy/data-ethics/> См., например, также: <https://dataethics.eu/>

84 См. также European Parliament and Council (2016).



- **управление данными:** этой теме посвящена, например, деятельность рабочей группы⁸⁵ в рамках Глобального партнёрства по искусственному интеллекту (GPAI), которое представляет собой «многостороннюю инициативу, направленную на преодоление разрыва между теорией и практикой в области искусственного интеллекта»⁸⁶;
- **грамотность в использовании данных:** грамотность в использовании данных является необходимым условием для использования (открытых) данных. Однако есть основания полагать, что в некоторых группах населения уровень грамотности в использовании данных невысок. Поэтому очень желательно наращивать потенциал в этом направлении (например, Montes & Slater, 2019);
- **инфраструктура данных:** другим необходимым условием предоставления и использования открытых данных является наличие аппаратной составляющей инфраструктуры данных, которая неравномерно распределена по всему миру (Dodds & Wells, 2019);
- **гендерный аспект:** как указывалось выше, достижение гендерного равенства связано с различными аспектами жизни общества, а значит, и с (открытыми) данными. Несбалансированность данных о женщинах особенно негативно сказывается на результатах работы систем искусственного интеллекта, которые отражают эту ситуацию в своих выводах⁸⁷;

85 <https://gpai.ai/projects/data-governance/>

86 <https://gpai.ai>

87 <https://aplusalliance.org>

- ❑ **данные о коренных народах:** подход к данным о коренных народах требует дополнительного осмысления, включая аспекты суверенитета, и может быть обобщён в виде принципов CARE: коллективная выгода, полномочия по контролю, ответственность и этичность⁸⁸;
- ❑ **малые данные:** как уже упоминалось, данные о маргинализированных группах часто доступны лишь в ограниченном объёме. Тем не менее важно, чтобы ИИ всё же мог работать с ними, как, например, в рамках подхода «быстрое обучение» (few-shot learning) (Wang et al, 2020);
- ❑ **«семантическая паутина»:** семантическая паутина — это подход, призванный сделать данные Интернета машиночитаемыми. Для подключения открытых данных к этому подходу была разработана и формализована концепция связанных открытых данных (Linked Open Data) (например, Bauer & Kaltenböck, 2011).

88 <https://www.gida-global.org/care>

3. РУКОВОДЯЩИЕ ПРИНЦИПЫ

3.1. ПРЕДИСЛОВИЕ

Цель руководящих принципов состоит в том, чтобы помочь государствам-членам разработать план действий, а также комплекс мер в отношении открытых данных в интересах устойчивого развития. Приведённая выше концептуальная часть касается общей картины, а руководящие принципы позволят специалистам получить более чёткое представление и «погрузиться» в конкретные темы.

Ниже приводятся практические высокоуровневые шаги по открытию данных, основанные на существующих руководящих принципах. Выделяются три этапа — подготовка, открытие данных и последующие мероприятия по повторному использованию и обеспечению устойчивости, — каждый из которых можно разбить на четыре шага.

Важно отметить, что некоторые из этих этапов могут выполняться одновременно, т. е. не обязательно последовательно.



3.2. ПОДГОТОВКА

➤ Разработка стратегии управления данными и их совместного использования

Разработка стратегии управления данными и их совместного использования является важным предварительным условием для открытия данных, поскольку такая стратегия определяет обязательства государства в отношении данных. Институт открытых данных предлагает следующие элементы стратегии открытых данных: определение понятия открытых данных, общая декларация принципов, описание типов данных и ссылки на соответствующие законы, правила или другие нормативные акты⁸⁹. Странам рекомендуется придерживаться принципа «открыто настолько, насколько возможно, закрыто настолько, насколько необходимо» (Landi et al., 2020). Если данные не могут быть открыты по соображениям юридического характера, неприкосновенности частной жизни или другим причинам, например, в случае персональных или охраняемых данных, это должно быть чётко оговорено. Стратегия управления и обмена данными должна также соответствовать принципам FAIR и CARE применительно к управлению данными о коренных народах. Кроме того, государство должно поощрять научно-исследовательские круги и частный сектор страны к разработке мер управления данными и их совместного использования в соответствии с теми же принципами.

➤ Сбор и накопление качественных данных

Существующие данные должны быть собраны и сохранены в одном хранилище, например, из различных государственных ведомств, где они могли храниться разрозненно. Кроме того, пробелы в данных должны быть заполнены путём сбора новых данных. Данные должны быть точными и актуальными. Кроме того, данные должны быть полными и не должны, например, игнорировать меньшинства или неформальные сектора экономики. Данные о людях должны быть дезагрегированы, где это уместно, в том числе по доходам, полу, возрасту, расовой и этнической принадлежности, миграционному статусу, инвалидности и географическому положению (Inter-Agency and Expert Group on Sustainable Development Goal Indicators, 2017).

➤ Развитие потенциала в области открытых данных

Развитие потенциала должно проходить по двум направлениям: государственные служащие и потенциальные пользователи данных. Ключевым

⁸⁹ <https://theodi.org/article/how-to-write-a-good-open-data-policy/>

понятием является «грамотность в использовании данных», которая подпадает под компетенции, связанные с медийной и информационной грамотностью, предложенные ЮНЕСКО⁹⁰, и для которых существуют курсы и учебные программы. Для государственных служащих развитие потенциала включает в себя понимание преимуществ открытых данных и преодоление распространённых страхов и непонимания. В этой связи также был введён термин «ответственное управление данными» (data stewardship), означающий «систематическое, устойчивое и ответственное управление данными на благо общества» (GovLab, 2020). Для потенциальных пользователей развитие потенциала включает демонстрацию возможностей открытых данных, например, по их повторному использованию, а также по осознанному (информированному) выбору.

➤ Подготовка данных для использования в ИИ

Если данные используются не только человеком, но и предполагается их использование в системах искусственного интеллекта, то они должны соответствовать ещё нескольким критериям. Первым шагом в этом направлении является подготовка данных в машиночитаемом формате. Некоторые форматы данных более пригодны к читаемости системами ИИ, чем другие⁹¹. Если данные, кроме того, будут использоваться в качестве обучающих данных для целей контролируемого обучения⁹² (в отличие от неконтролируемого⁹³ или обучения с подкреплением⁹⁴), то они должны быть очищены и промаркированы, что зачастую требует много времени, а, следовательно, и затрат средств. Успех работы системы искусственного интеллекта зависит от качества обучающих данных, в том числе от их согласованности и релевантности. Необходимый объем обучающих данных трудно определить заранее, и его приходится отслеживать с помощью проверок эффективности процесса. Данные должны охватывать все сценарии, для работы которых создавалась система искусственного интеллекта.

90 <https://iite.unesco.org/mil/>

91 См. обзор: <http://opendatahandbook.org/guide/en/appendices/file-formats/>

92 Контролируемое обучение основано на использовании размеченных данных, например, фотографий людей, которые были размечены человеком. Затем строится модель, которая может быть применена к аналогичным данным, например, для автоматической идентификации тех же людей на новых фотографиях (UNESCO, 2021c).

93 Неконтролируемое обучение основано на данных, которые не были классифицированы или размечены. Цель в том, чтобы выявить в данных скрытые закономерности, которые используются для классификации новых данных. Примером может служить автоматическая идентификация букв и цифр в рукописном тексте путём поиска закономерностей в большом количестве образцов (UNESCO, 2021c).

94 Обучение с подкреплением предполагает непрерывное совершенствование модели за счёт обратной связи. На основе исходных данных ИИ формирует модель, которая оценивается как правильная или неправильная и, соответственно, следует «вознаграждение» или «наказание». Затем это подкрепление используется для обновления модели, которое итеративно повторяется с течением времени (UNESCO, 2021c).

3.3. ОТКРЫТИЕ ДАННЫХ

➤ Выбор наборов данных для открытия

Первым шагом в открытии данных является принятие решения о том, какие наборы данных следует открыть. Критерии в пользу открытия данных следующие: поступали ли ранее запросы на открытие этих данных, открывались ли такие данные другими государствами и принесло ли это пользу. Критериями против открытия являются вопросы, связанные с неприкосновенностью частной жизни, правами на интеллектуальную собственность или национальной безопасностью. Другими словами, открытие данных не должно нарушать национальные законы, например, о конфиденциальности данных.

➤ Юридическое открытие наборов данных

Перед раскрытием наборов данных государство должно точно указать, на каких условиях, если таковые имеются, можно использовать эти данные. Разработано множество лицензий на открытые данные, и государство может выбрать ту, которая наиболее соответствует его целям. Широко известны лицензии Creative Commons и Open, «которые предоставляют каждому человеку и организации в мире бесплатный, простой и стандартизированный способ выдать разрешение в области авторских прав на творческие и научные произведения, обеспечить надлежащее указание авторства и позволяют выдать разрешение другим лицам копировать, распространять и использовать эти произведения»⁹⁵. Помимо отказа от всех прав на данные, можно рассмотреть следующие варианты ограничений, для которых доступны варианты лицензии Creative Commons:

- Attribution (атрибуция): разрешено свободное использование данных при условии указания их источника;
- Non-Commercial (некоммерческое использование): запрещено использование данных в целях получения прибыли;
- No Derivatives (никаких производных): запрещено распространять изменённые данные;
- Share Alike (сохранение условий): разрешено распространять изменённые данные только на условиях лицензии исходных данных.

Другим вариантом является разработка государствами собственных лицензий, как это, например, сделало правительство Великобритании⁹⁶.

⁹⁵ <https://creativecommons.org/about/>

⁹⁶ <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>

➤ Техническое открытие наборов данных

Наиболее распространённым способом открытия данных является их публикация в электронном формате для скачивания на каком-либо сайте — либо на собственном сайте правительства, либо на сайте сторонней организации. Кроме того, данные могут быть опубликованы в архиве или хранилище данных, которое, согласно Рекомендации ЮНЕСКО по открытой науке, должно «поддерживаются и обслуживаются академическими учреждениями, научными обществами, государственными органами или авторитетными некоммерческими организациями, которые ради общего блага обеспечивают открытый доступ к материалам, их неограниченное распространение, функциональную совместимость и долгосрочное сохранение и архивирование в цифровом формате» (UNESCO, 2021b, p. 9). Данные должны быть представлены в формате, обеспечивающем возможность их поиска, доступность, функциональную совместимость и повторное использование, что соответствует принципам FAIR, представленным выше. Для обеспечения удобства поиска и доступности существуют различные системы управления данными с открытым исходным кодом, позволяющие создать портал открытых данных по принципу «одного окна»⁹⁷. Наборы данных должны сопровождаться метаданными, которые ещё более конкретизируют данные и для которых существуют четыре общих стандарта: Dublin Core⁹⁸, Data Catalog Vocabulary⁹⁹, DataCite¹⁰⁰ и Schema.org¹⁰¹. Что касается конкретных форматов данных, то ранее в этом докладе была представлена схема «пяти звёзд» для категоризации открытых данных. Чем больше звёзд, тем более открытыми являются данные¹⁰². Форматы данных должны поддерживать совместимость, например, с интерфейсами прикладного программирования, и возможность повторного использования, например, другими пользователями, а также машинами¹⁰³. Кроме того, необходимо помнить о пригодности для чтения средствами ИИ, как обсуждалось в шаге 4. Также, крайне важна реализация мер безопасности для предотвращения манипуляций с данными со стороны неавторизованных пользователей.

➤ Формирование культуры открытых данных

Опыт показывает, что для построения культуры открытых данных, помимо юридического и технического открытия данных, необходимо выполнить ещё как минимум две задачи. Зачастую сотрудники, особенно в государственных ведомствах, не привыкли делиться данными, а, скорее, были приучены

97 См., например: <https://ckan.org/>, <https://magda.io/> or <https://www.re3data.org/>

98 <http://dublincore.org/>

99 <https://www.w3.org/TR/vocab-dcat-2/>

100 <https://datacite.org/>

101 <https://schema.org/>

102 <https://5stardata.info/en/> и <https://www.w3.org/DesignIssues/LinkedData.html>

103 См. обзор: <http://opendatahandbook.org/guide/en/appendices/file-formats/>

придерживаться замкнутого мышления. Более того, данные должны стать, по возможности, исключительной основой для принятия решений, другими словами, решения должны основываться на данных. Обе эти темы затрагиваются в шаге 3 «Развитие потенциала в области открытых данных», однако помимо потенциала требуются изменения в культуре всех вовлечённых сотрудников. Необходимо развивать проактивное раскрытие данных, которое может гарантировать, что данные будут доступны ещё до того, как появится запрос на них. Этого можно достичь путём включения соответствующих положений в законы о порядке предоставления доступа к информации¹⁰⁴.



3.4. ОБЕСПЕЧЕНИЕ ПОВТОРНОГО ИСПОЛЬЗОВАНИЯ И УСТОЙЧИВОСТИ

➤ Поддержка участия граждан

После открытия данных о них должны узнать потенциальные пользователи. Для этого необходимо разработать стратегию пропаганды, которая может включать в себя анонсирование открытия данных в сообществах открытых данных и соответствующих каналах социальных сетей, например, с использованием хэштега #opengov. Ещё одним важным мероприятием является проведение предварительных консультаций и взаимодействие с потенциальными пользователями. Помимо информирования потенциальных пользователей об открытых данных, необходимо также поощрять их к использованию и повторному применению данных, а также к дальнейшему участию. Свою эффективность доказали такие механизмы вовлечения, как хакатоны (англ. hackathon — hacker «хакер» + marathon «марафон», соревнования разработчиков) и другие конкурсы по использованию открытых данных для решения конкретных задач и проблем¹⁰⁵.

104 См., например: <https://en.unesco.org/themes/access-information-laws>

105 <http://opendatatoolkit.worldbank.org/en/demand.html#stage-3>



©Miha Creative/Shutterstock.com

➤ Поддержка международного участия

Международные партнёрства, например, по линиям Юг-Юг и Север-Юг, позволят ещё больше повысить отдачу от открытых данных. Особенно важны партнёрства, направленные на поддержку и наращивание потенциала для эффективного повторного использования данных, как с применением искусственного интеллекта (см. ниже), так и без. В частности, заслуживают высокой оценки Альянс «Открытость во благо» (Open for Good Alliance), Глобальное партнёрство по искусственному интеллекту (GPAI), а также межправительственные и региональные инициативы, которые были представлены выше.

➤ Поддержка полезного применения ИИ

Открытые данные предоставляют множество возможностей для систем искусственного интеллекта. Для того чтобы полностью раскрыть потенциал данных, необходимо убеждать разработчиков использовать эти данные и создавать соответствующие системы ИИ. В то же время необходимо предотвращать злоупотребление открытыми данными, например, при безответственном и вредоносном применении ИИ. Одним из примеров этого являются лицензии на ответственное применение искусственного интеллекта (Responsible AI Licences, RAIL)¹⁰⁶. Этот шаг связан с предыдущим. Также рекомендуется ведение публичного учёта того, какие данные использовались системами ИИ, и каким образом.

¹⁰⁶ <https://www.licenses.ai/about>

➤ Поддержание высокого качества данных

Многие данные быстро устаревают. Поэтому наборы данных необходимо регулярно обновлять.

Шаг «Поддержание высокого качества данных» придаёт циклический характер рекомендациям, содержащимся в настоящих руководящих принципах, поскольку он связан с шагом «Сбор и накопление качественных данных».



©UnderhilStudio/Shutterstock.com

ЗАКЛЮЧЕНИЕ

Настоящие руководящие принципы служат призывом к действию в соответствии с Рекомендацией ЮНЕСКО об этических аспектах искусственного интеллекта (UNESCO, 2021d). Если государства-члены ЮНЕСКО будут следовать этим принципам, открывать свои данные на устойчивой основе и создавать соответствующий потенциал, а также культуру открытых данных, то изложенная в предыдущей части концепция со всеми её преимуществами может стать реальностью.

Открытые данные являются необходимым условием для мониторинга и достижения целей устойчивого развития. Ввиду масштабности поставленных задач государства должны не только принять решение об открытии данных, но и создать благоприятные условия для полезного участия технологий искусственного интеллекта, позволяющих создавать на основе открытых данных новые знания для принятия решений на основе фактических данных.

Вкратце: сделайте свои данные доступными для поиска, функционально совместимыми и пригодными к повторному использованию, а также готовыми к применению в ИИ, или, иначе говоря, приведите в соответствие с принципами FAIR, чтобы их мог обрабатывать и анализировать любой желающий на благо общества.



©3rdtimeluckystudio/Shutterstock.com

4. БИБЛИОГРАФИЯ

Access Partnership (2018). Artificial Intelligence for Africa: An opportunity for growth, development and democratisation. South Africa: University of Pretoria.

https://www.up.ac.za/media/shared/7/ZP_Files/ai-for-africa.zp165664.pdf

Alamo, T., Reina, D. G., Mammarella, M., & Abella, A. (2020). Covid-19: Open-data resources for monitoring, modeling, and forecasting the epidemic. *Electronics*, 9(5), 827.

<https://www.mdpi.com/2079-9292/9/5/827/htm>

Bauer, F., & Kaltenböck, M. (2011). Linked open data: The Essentials. *Edition mono/monochrom*, Vienna, 710.

<https://www.reeeep.org/LOD-the-Essentials.pdf>

Berners-Lee, T. (2012). Raw data, now! Wired.

<https://www.wired.co.uk/article/raw-data>

Big Data for Migration Alliance (2021). Designing Data Collaboratives to Better Understand Human Mobility and Migration in West Africa.

<https://odpl.thegovlab.com/uploads/ccm-directus/originals/aa45abf2-743d-4d58-8471-e25bff956ba5.pdf>

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317-331.

<https://arxiv.org/pdf/1712.03141.pdf>

Bodor, Á., Varjú, V., & Grünhut, Z. (2020). The effect of trust on the various dimensions of climate change attitudes. *Sustainability*, 12(23), 10200.

<https://www.mdpi.com/2071-1050/12/23/10200>

von Borzyskowski, I., Mazumder, A., Mateen, B., & Wooldridge, M. (2021). Data science and AI in the age of COVID-19. The Alan Turing Institute.

<https://www.turing.ac.uk/research/publications/data-science-and-ai-age-covid-19-report>

Bostrom, N. (2011). Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy*, (10), 44-79.

<https://www.nickbostrom.com/information-hazards.pdf>

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

Cascella, M., Rajnik, M., Aleem, A., Dulebohn, S. C., & Di Napoli, R. (2022). Features, evaluation, and treatment of coronavirus (COVID-19). *Statpearls* [internet].

<https://www.ncbi.nlm.nih.gov/books/NBK554776/>

Davies, T. (2019). Issues in Open Data – Algorithms and AI. In T. Davies, S. Walker, M. Rubinstein, & F. Perini (Eds.), *The State of Open Data: Histories and Horizons*. Cape Town and Ottawa: African Minds and International Development Research Centre.

<https://idl-bnc-idrc.dspacedirect.org/bitstream/handle/10625/57585/The%20State%20of%20Open%20Data.pdf>

Dodds, L. & Wells, P. (2019). Issues in Open Data – Data Infrastructure. In T. Davies, S. Walker, M. Rubinstein, & F. Perini (Eds.), *The State of Open Data: Histories and Horizons*. Cape Town and Ottawa: African Minds and International Development Research Centre.

<https://idl-bnc-idrc.dspacedirect.org/bitstream/handle/10625/57585/The%20State%20of%20Open%20Data.pdf>

European Parliament and Council (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.

<https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX-02016R0679-20160504&from=EN>

European Union (2012). Charter of Fundamental Rights of the European Union. Official Journal of the European Union, C 326/391.

<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:12012P/TXT&from=EN>

Executive Office of the President of the United States of America – National Science and Technology Council Committee on Technology (2016). *Preparing for a Future of Artificial Intelligence*.

https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Giovannini, G., Haick, H., & Garoli, D. (2021). Detecting COVID-19 from breath: A game changer for a big challenge. *ACS sensors*, 6(4), 1408-1417.

<https://pubs.acs.org/doi/10.1021/acssensors.1c00312>

GLZ (2020). *A Study on Open Voice Data in Indian Languages*.

https://toolkit-digitalisierung.de/app/uploads/2021/02/Study-on-Open-Voice-Data-in-Indian-Languages_GLZ-BizAugmentor.pdf

GovLab (2020). *Wanted: Data Stewards*.

<https://thegovlab.org/static/files/publications/wanted-data-stewards.pdf>

Harrus, I., & Wyndham, J. (2021). Artificial intelligence and COVID-19: applications and impact assessment. *AAAS AI Report*.

https://www.aaas.org/sites/default/files/2021-05/AlandCOVID19_2021_FINAL.pdf

Heaven, W.D. (2021). Hundreds of AI tools have been built to catch covid. None of them helped. MIT Technology Review.

<https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/>

Hussain, A. A., Bouachir, O., Al-Turjman, F., & Aloqaily, M. (2020). AI techniques for COVID-19. IEEE access, 8, 128776-128795.

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9136710>

Independent Expert Advisory Group on a Data Revolution for Sustainable Development (2014). *A world that counts: Mobilizing a Data Revolution for Sustainable Development*.

<http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf>

Inter-Agency and Expert Group on Sustainable Development Goal Indicators (2017). Revised list of global Sustainable Development Goal indicators. Annex III. E/CN.3/2017/2.

<https://unstats.un.org/sdgs/indicators/official%20revised%20list%20of%20global%20sdg%20indicators.pdf>

International Monetary Fund and Financial Stability Board (2021). G20 Data Gaps Initiative (DGI-2). The Sixth Progress Report — Countdown to December 2021.

<https://www.fsb.org/wp-content/uploads/Po81021-1.pdf>

International Telecommunication Union and CISCO Systems (2016). Harnessing the Internet of Things for Global Development.

<http://www.itu.int/en/action/broadband/Documents/Harnessing-IoT-Global-Development.pdf>

Khemasuwan, D., & Colt, H. G. (2021). Applications and challenges of AI-based algorithms in the COVID-19 pandemic. BMJ Innovations, 7(2).

<https://innovations.bmj.com/content/bmjinnov/7/2/387.full.pdf>

Klein, T., & Verhulst, S. (2017). Access to new data sources for statistics: Business models and incentives for the corporate sector.

<https://www.thegovlab.org/static/files/publications/paris-21.pdf>

Landi, A., Thompson, M., Giannuzzi, V., Bonifazi, F., Labastida, I., da Silva Santos, L. O. B., & Roos, M. (2020). The "A" of FAIR—as open as possible, as closed, as necessary. *Data Intelligence*, 2(1-2), 47-55.

<https://direct.mit.edu/dint/article/2/1-2/47/9998/The-A-of-FAIR-As-Open-as-Possible-as-Closed-as>

Lee, C., Yang, T., Inchoco, G. D., Jones, G. M., & Satyanarayan, A. (2021). Viral visualizations: How coronavirus skeptics use orthodox data practices to promote unorthodox

science online. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-18).

<https://arxiv.org/pdf/2101.07993.pdf>

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4).

<https://arxiv.org/pdf/0712.3329.pdf>

Lohn, A.J. (2021). Poison in the Well - Securing the Shared Resources of Machine Learning. CSET Policy Brief.

<https://cset.georgetown.edu/publication/poison-in-the-well/>

Lozano, M. A., Piñol, E., Rebollo, M., Polotskaya, K., Garcia-March, M. A., Conejero, J. A., ... & Oliver, N. (2021). Open Data Science to Fight COVID-19: Winning the 500k XPRIZE Pandemic Response Challenge. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 384-399). Springer, Cham.

https://2021.ecmlpkdd.org/wp-content/uploads/2021/07/sub_638.pdf

Milan, S., & Treré, E. (2020). The Rise of the Data Poor: The COVID-19 Pandemic Seen From the Margins. *Social Media+ Society*, 6(3), 2056305120948233.

https://journals.sagepub.com/doi/full/10.1177/2056305120948233?fbclid=IwAR-ojUNmHgCSSZmOXaByn3X4EyKoWRDCxM6qXnCA1FuN-yrgF_s8rnoCgVqM&

Montes, M. & Slater, D. (2019). Issues in Open Data - Data Literacy. In T. Davies, S. Walker, M. Rubinstein, & F. Perini (Eds.), *The State of Open Data: Histories and Horizons*. Cape Town and Ottawa: African Minds and International Development Research Centre.

<https://idl-bnc-idrc.dspacedirect.org/bitstream/handle/10625/57585/The%20State%20of%20Open%20Data.pdf>

Odriozola-González, P., Planchuelo-Gómez, Á., Iruetia, M. J., & de Luis-García, R. (2020). Psychological effects of the COVID-19 outbreak and lockdown among students and workers of a Spanish university. *Psychiatry research*, 290, 113108.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7236679/>

OECD (2020). Why open science is critical to combatting COVID-19. OECD Policy Responses to Coronavirus (COVID-19).

https://read.oecd-ilibrary.org/view/?ref=129_129916-31pgjnl6cb&title=Why-open-science-is-critical-to-combatting-COVID-19

OECD & GovLab (2021). Open Data in action. Initiatives during the initial stage of the COVID-19 pandemic.

<https://www.oecd.org/gov/digital-government/open-data-in-action-initiatives-during-the-initial-stage-of-the-covid-19-pandemic.pdf>

Ong, E., Wong, M. U., Huffman, A., & He, Y. (2020). COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. *Frontiers in immunology*, 11, 1581.

<https://www.frontiersin.org/articles/10.3389/fimmu.2020.01581/full>

Open Data Charter (2015). International Open Data Charter.

https://opendatacharter.net/wp-content/uploads/2015/10/opendatacharter-charter_F.pdf

Pahar, M., Klopfer, M., Warren, R., & Niesler, T. (2021). COVID-19 cough classification using machine learning and global smartphone recordings. *Computers in Biology and Medicine*, 135, 104572.

<https://arxiv.org/pdf/2012.01926.pdf>

Panchal, N., Kamal, R., Orgera, K., Cox, C., Garfield, R., Hamel, L., & Chidambaram, P. (2020). The implications of COVID-19 for mental health and substance use. Kaiser family foundation, 21.

<https://www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/>

Petrov, O., Gurin, J., & Manley, L. (2016). Open data for sustainable development.

<https://openknowledge.worldbank.org/bitstream/handle/10986/24017/Open-odataforosustainabledevelopment.pdf?sequence=4&isAllowed=y>

Ratcliffe, J. & Tuzeneu, M.-C. (2019). Maximizing the Power of Climate Data by Building Trust. Blog post.

<https://www.climatelinks.org/blog/maximizing-power-climate-data-building-trust>

Raveendran, A. V., Jayadevan, R., & Sashidharan, S. (2021). Long COVID: an overview. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*.

<https://www.sciencedirect.com/science/article/pii/S1871402121001193>

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3), 199-217.

<https://www.nature.com/articles/s42256-021-00307-0>

Russell, S. J. & Norvig, P. (2015). *Artificial Intelligence-A Modern Approach*, Third Edition. Upper Saddle River: Pearson.

Sachs, J. D., Kroll, C., Lafortune, G., Fuller, G. & Woelm, F. (2021). *Sustainable Development Report 2021*. Cambridge: Cambridge University Press.

<https://s3.amazonaws.com/sustainabledevelopment.report/2021/2021-sustainable-development-report.pdf>

Scassa, T. (2019). Issues in Open Data - Privacy. In T. Davies, S. Walker, M. Rubinstein, & F. Perini (Eds.), *The State of Open Data: Histories and Horizons*. Cape Town and Ottawa: African Minds and International Development Research Centre.

<https://idl-bnc-idrc.dspacedirect.org/bitstream/handle/10625/57585/The%20State%20of%20Open%20Data.pdf>

Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., ... & Shen, D. (2020). Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. *IEEE reviews in biomedical engineering*, 14, 4-15.

<https://arxiv.org/ftp/arxiv/papers/2004/2004.02731.pdf>

Shuja, J., Alanazi, E., Alasmay, W., & Alashaikh, A. (2021). COVID-19 open source data sets: a comprehensive survey. *Applied Intelligence*, 51(3), 1296-1325.

<https://link.springer.com/article/10.1007/s10489-020-01862-6>

UN General Assembly (2020). Road map for digital cooperation: implementation of the recommendations of the High-level Panel on Digital Cooperation. Report of the Secretary-General. A/74/821.

<https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/102/51/PDF/N2010251.pdf?OpenElement>

UNESCO (2020). Disinfodemic. Deciphering COVID-19 disinformation. Policy brief 1.

https://en.unesco.org/sites/default/files/disinfodemic_deciphering_covid19_disinformation.pdf

UNESCO (2021a). One year into COVID: prioritizing education recovery to avoid a generational catastrophe. ED/ADG/2021/01

<https://unesdoc.unesco.org/ark:/48223/pf0000376984>

UNESCO (2021b). Recommendation on Open Science. SC-PCB-SPP/2021/OS/UROS.

<https://unesdoc.unesco.org/ark:/48223/pf0000379949>

UNESCO (2021c). AI and Education Guidance for Policy-makers. Paris, UNESCO.

<https://unesdoc.unesco.org/ark:/48223/pf0000376709>

UNESCO (2021d). Recommendation on the Ethics of Artificial Intelligence. Paris, UNESCO.

<https://unesdoc.unesco.org/ark:/48223/pf0000381137>

UN System Chief Executives Board for Coordination (2020). Summary of deliberations. Addendum. System-wide Road Map for Innovating United Nations Data and Statistics. CEB/2020/1/Add.1

https://unsceb.org/sites/default/files/2021-01/CEB_2020_1_Add1.pdf

Van Noorden, R. (2021). Scientists call for fully open sharing of coronavirus genome data. *Nature*, 590(7845), 195-197.

<https://www.nature.com/articles/d41586-021-00305-7>

Waltman, L., Pinfield, S., Rzayeva, N., Henriques, S. O., Fang, Z., Brumberg, J., ... & Swaminathan, S. (2021). Scholarly communication in times of crisis: The response of the scholarly communication system to the COVID-19 pandemic.

https://rori.figshare.com/articles/report/_/17125394

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3), 1-34.

<https://arxiv.org/pdf/1904.05046.pdf>

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/>

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., ... & van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369.

<https://www.bmj.com/content/369/bmj.m1328>

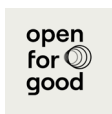
Yampolskiy, R. V. (Ed.). (2018). *Artificial intelligence safety and security*. CRC Press.

Yan, L., Zhang, H. T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., ... & Yuan, Y. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature machine intelligence*, 2(5), 283-288.

<https://www.nature.com/articles/s42256-020-0180-7.pdf>

Ziesche, S. (2017). Innovative big data approaches for capturing and analyzing data to monitor and achieve the SDGs. *Report of the United Nations Economic and Social Commission for Asia and the Pacific: Subregional Office for East and North-East Asia (ESCAP-ENEAS)*.

<https://www.unescap.org/publications/innovative-big-data-approaches-capturing-and-analyzing-data-monitor-and-achieve-sdgs>



Кафедра ЮНЕСКО по искусственному интеллекту (2019),
Университетский колледж Лондона, Соединённое
Королевство Великобритании и Северной Ирландии
Кафедра ЮНЕСКО по аналитике и науке о данных (2016)
Университет Эссекса, Соединённое Королевство
Великобритании и Северной Ирландии
Кафедра ЮНЕСКО по искусственному интеллекту и науке
о данных для общества (2022), Гаагский университет
прикладных наук (THUAS)

