Предварительный доклад Независимой международной научной группы по искусственному интеллекту

результаты проведенной с использованием фактических данных оценки возможностей и рисков, связанных с искусственным интеллектом, и его воздействия

Июль 2026 года

United Nations

Independent International Scientific Panel on Artificial Intelligence

Предварительный доклад Независимой международной научной группы по искусственному интеллекту: результаты проведенной с использованием фактических данных оценки возможностей и рисков, связанных с искусственным интеллектом, и его воздействия

Members of the Independent International Scientific Panel on Artificial Intelligence

Girmaw Abebe Tadesse, Ethiopia

Tuka Alhanai, United Arab Emirates

Joëlle Barral, France

Yoshua Bengio, Canada *(Co-Chair)*

Tegawendé Bissyandé, Burkina Faso

Loreto Bravo, Chile

Mark Coeckelbergh, Belgium

Carlos Coello Coello, Mexico

Melahat Bilge Demirköz, Türkiye

Adji Bousso Dieng, Senegal

Awa Bousso Dramé, Cabo Verde

Mennatallah El-Assady, Egypt

Hoda Heidari, Islamic Republic of Iran

Juho Kim, Republic of Korea

Anna Korhonen, Finland

Aleksandra Korolova, Latvia

Vipin Kumar, United States

Sonia Livingstone, United Kingdom

Qinghua Lu, Australia

Teresa Ludermir, Brazil

Vukosi Marivate, South Africa

Bilal Mateen, Pakistan

Yutaka Matsuo, Japan

Joyce Nakatumba Nabende, Uganda

Andrei Neznamov, Russian Federation

Maximilian Nickel, Germany

Rita Orji, Nigeria

Román Orús, Spain

Alvitta Ottley, Saint Kitts and Nevis

Martha Palmer, United States

Johanna Pirker, Austria

Balaraman Ravindran, India

Maria Ressa, Philippines *(Co-Chair)*

Lior Rokach, Israel

Piotr Sankowski, Poland

Silvio Savarese, Italy

Bernhard Schölkopf, Germany

Haitao Song, China

Leslie Teo, Singapore

Jian Wang, China

Об этом докладе

В настоящем докладе представлена предварительная независимая научная оценка функциональных возможностей искусственного интеллекта (ИИ), а также открывающихся с его развитием перспектив и формирующихся рисков, которая призвана служить общей доказательной базой и помогать государствам-членам ориентироваться в вопросах, связанных со стремительно эволюционирующей технологией. Автором доклада является Независимая международная научная группа по искусственному интеллекту — орган, учрежденный Генеральной Ассамблеей в ее резолюции 79/325 в 2025 году. Эта Группа представляет собой первый всемирный научный орган, занимающийся вопросами ИИ и функционирующий в рамках строго научного, неполитического мандата с целью документирования научного консенсуса и научных разногласий на международном уровне при сохранении актуальности для выработки политики, но без выдачи предписаний. Данный доклад является первым в своем роде и будет постепенно обновляться в течение года за счет бюллетеней с тематической справочной информацией, посвященных новым событиям. В нем отражены наиболее достоверные фактические данные, имевшиеся на момент публикации в области, которая развивается настолько стремительно, что требуется готовность к пересмотру любой зафиксированной картины ситуации.

Отказ от ответственности:

Настоящий доклад является результатом работы Независимой международной научной группы по искусственному интеллекту, которая несет ответственность за его содержание и публикацию. Члены Группы выступают в своем личном качестве, а не в качестве представителей какого-либо правительства или какого-либо другого ведомства или организации.

Группа пользовалась полной свободой действий в вопросах включения информации в данный доклад. Доклад и содержащаяся в нем информация отражают широкий консенсус среди ее членов; ни от одного из ее членов не ожидается согласие с каждым пунктом, содержащимся в данном документе. Члены Группы заявляют о своем широком, но не одностороннем согласии с его выводами.

Доклад не отражает мнения Организации Объединенных Наций, и ничто в нем не должно толковаться как отражение официального мнения или позиции какого-либо правительства или какого-либо другого ведомства или организации, с которыми может быть связан какой-либо член Группы.

Используемые обозначения, включая географические названия, и изложение материалов в настоящем издании, включая цитаты, карты и библиографию, не означают выражения какого-либо мнения со стороны Организации Объединенных Наций относительно названий и правового статуса какой-либо страны, территории, города или района или их властей или относительно делимитации их границ или рубежей и не означают официального одобрения или принятия со стороны Организации Объединенных Наций. Содержащаяся в настоящем издании информация, проистекающая из предпринятых государствами действий и принятых ими решений, не означает официального одобрения, принятия или признания таких действий и решений Организацией Объединенных Наций, и включение такой информации не наносит ущерба позиции какого-либо государства — члена Организации Объединенных Наций. Информация, содержащаяся в этом докладе и касающаяся конкретных компаний или продуктов определенных производителей, не означает официального одобрения или официальной рекомендации со стороны Организации Объединенных Наций (в том числе в связи с отсутствием упоминаний других продуктов подобного характера). Настоящее из-

дание не влияет на позицию какого-либо государства — члена Организации Объединенных Наций в отношении какого-либо международного многостороннего соглашения.

Содержание

Резюме

Этот документ создан с целью представить взвешенный анализ рисков и возможностей, связанных с искусственным интеллектом (ИИ). Под «взвешенным» понимается анализ, основанный на обязательстве оценки эмпирических данных без ненадлежащего оптимистического или пессимистического уклона. Потенциальные преимущества использования ИИ огромны. При продуманном внедрении и применении ИИ может содействовать прогрессу в достижении целей в области устойчивого развития, дальнейшему развитию наук о здоровье и расширению доступа к образованию. В то же время стремительные темпы развития технологий и широкий спектр потенциальных сфер их применения ставят перед сторонами, ответственными за выработку политики, серьезные задачи. Стремительное, беспрепятственное внедрение этой технологии в широких масштабах также сопряжено со значительными рисками, в том числе с риском нанесения вреда психическому здоровью пользователей, риском потенциального использования в качестве инструмента оказания разрушительного воздействия, риском воздействия на социальные, экономические и экологические системы, а также с трудностями, связанными с контролем над этой технологией. Целью данного доклада не является рассмотреть полный спектр всех потенциальных возможностей и рисков; вместо этого внимание в нем сосредоточено на нескольких наиболее насущных из них.

Функциональные возможности и внедрение

В последние годы наблюдается стремительный, а в некоторых областях даже ускоряющийся прогресс в развитии целого ряда функциональных возможностей ИИ. Значительные инвестиции в вычислительные мощности, новые методологические основы развития ИИ и использование специализированных обучающих данных привели к систематическому совершенствованию широкого спектра функциональных возможностей ИИ. К ним относятся естественное ведение разговора, генерирование функционального кода, логические построения на экспертном уровне в математике и естественных науках, анализ крупных массивов данных, а также генерирование изображений и аудио- и видеоконтента. Сохраняются ограничения, например в плане надежности, обеспечения высоких результатов при работе с различными человеческими языками и культурами, взаимодействия с физическими системами, осуществления сложных или многоэтапных проектов, а также генерирования результатов, основанных на фактах; в целом, однако, технический прогресс во многих важных областях уже на протяжении нескольких лет идет быстрыми темпами, которые превосходят обычные ожидания в отношении развития технологий.

Благодаря этим достижениям стало возможным применять полезные функции на практике в науке, здравоохранении, сельском хозяйстве, сфере обеспечения доступности, сфере умственного труда и в информационных технологиях, в том числе в разработке самого ИИ. Например, в науке использование системы AlphaFold позволило предсказать структуру более 200 миллионов белков (эти результаты сейчас используют более 3 миллионов исследователей), а также ускорить разработку лекарственных препаратов, создание вакцин и изучение антибиотикорезистентности. Кроме того, радиологи используют ИИ для более ранней диагностики рака груди, а медицинские работники первичного звена в условиях низкой обеспеченности ресурсами применяют инструменты ИИ, адаптированные к местным языкам, для предоставления более качественных услуг по охране здоровья.

Внедрение ИИ в разных странах и секторах ускорилось повсеместно и при этом идет неравномерно. На данный момент более миллиарда человек во всем мире еженедельно пользуются диалоговым ИИ. Тем не менее доступ к ИИ и его использование значительно различаются в мировом масштабе, причем в странах глобального Юга внедрение идет значительно медленнее, чем в странах глобального Севера. Кроме того, между странами с развитой экономикой наблюдаются значительные различия в вычислительной инфраструктуре и вычислительных моделях. Эта диспропорция отражает существующее неравенство и может даже усиливать его. Сама разработка ИИ характеризуется еще более высокой концентрированностью: по недавним оценкам, на долю Соединенных Штатов Америки приходится 75 процентов вычислительной мощности 500 лучших суперкомпьютеров мира, работающих с ИИ, а на долю Китая — 15 процентов. Компании в Соединенных Штатах и Китае также разрабатывают почти все ведущие модели общего назначения, а небольшое число стран контролирует поставки критически важных ресурсов для цепочки поставок и производства компьютерных чипов для работы ИИ.

Продолжается процесс перехода к использованию ИИ-агентов, однако их будущее внедрение и экономическое воздействие, по-видимому, будут зависеть от непрерывного совершенствования их способности выполнять задачи умственного труда с минимальным человеческим надзором или без него. ИИ-агент — это компьютерная система со способностью планировать и в автономном режиме предпринимать действия для достижения той или иной цели с использованием имеющихся в ее распоряжении средств. В последние годы эти системы стремительно совершенствуются: по результатам одного исследования, длина некоторых заданий по программированию, которые передовые системы способны успешно выполнять, удваивается каждые 4–7 месяцев. Если такие темпы совершенствования сохранятся, то скоро ИИ-агенты будут выполнять задания, на которые у программистов сейчас уходят дни или недели. Поскольку ИИ-агенты способны работать очень быстро и под минимальным надзором, их использование может принести значительные экономические и научные выгоды. Например, при применении систем агентного ИИ в автономных химических лабораториях было зафиксировано более чем десятикратное увеличение скорости открытия новых материалов. В некоторых ситуациях, связанных с научно-исследовательской работой, анализ литературы с помощью ИИ, по некоторым оценкам, позволил сократить рабочую нагрузку приблизительно на 60 процентов. Поэтому применение ИИ-агентов влечет за собой значительные последствия вне зависимости от отрасли. В то же время их внедрение ставит требующие неотложного решения вопросы, касающиеся рынков труда, кибербезопасности, информационных экосистем, а также управления будущими системами ИИ и их контролируемости.

Понимание рисков и управление ими

Развитие ИИ сопряжено с рисками и потенциальным негативным воздействием на права человека, социальные системы и окружающую среду. Например, в Интернете теперь чаще распространяются созданные ИИ материалы, содержащие сцены сексуального надругательства над детьми, и материалы, основанные на дипфейках и содержащие сцены сексуального насилия, нанося несоразмерно большой ущерб женщинам и детям. Угодливое поведение ИИ, при котором ответы ИИ подкрепляют существующие убеждения пользователей независимо от их верности, связывают с несколькими серьезными инцидентами в области психического здоровья, в том числе с документально подтвержденными смертельными исходами. ИИ упрощает масштабное создание и целенаправленное распространение убедительного контента, в том числе призванного вводить в заблуждение, что способствует постепенному ослаблению целостности информации, которое, в свою очередь, может ослаблять чувство существования в общей действи-

тельности, необходимое для поддержания общественного доверия и социальной сплоченности и ведения демократических обсуждений. Имеются документальные подтверждения того, что преступники и злоумышленники используют системы ИИ для помощи в совершении кибератак. Многие из этих негативных последствий в несоразмерной степени затрагивают группы населения, уже находящиеся в неблагоприятном положении.

В перспективе разрыв между стремительно растущими функциональными возможностями и наличием эффективных методов управления рисками может привести к катастрофическим результатам. Например, благодаря передовым техническим возможностям частные субъекты, не имеющие соответствующего опыта, могут получить возможность злонамеренным образом использовать ИИ на практике в целом ряде областей, таких как мошенничество, социальная инженерия, кибербезопасность, дезинформация, биотехнологии и манипулирование финансовыми рынками. Надежные методы сохранения контроля над высокоавтономными системами ИИ отсутствуют. Нет никаких научных гарантий того, что ИИ-агенты не будут нарушать данные им инструкции, и появляется все больше фактических данных о случаях, в которых они уже их нарушают. В лабораторных условиях было продемонстрировано, что системы ИИ нарушают данные им инструкции в области безопасности, чтобы избежать своего отключения. Схожее поведение может создавать трудности для применения методов оценки и надзора, поскольку ведущие системы ИИ все лучше распознают тестовую среду и способны генерировать результаты оценки, которые вводят в заблуждение и потенциально способствовали бы продолжению функционирования этих систем. Помимо этого, принципиально новые риски могут возникать в результате взаимодействия между несколькими агентами.

Риски, связанные с использованием ИИ, распределены между группами населения и странами неравномерно, тогда как разработка ИИ и генерируемый им капитал характеризуются значительной степенью концентрированности. Концентрация управления функциональными возможностями ИИ в небольшом числе компаний и стран может привести к захвату контроля авторитарными силами и подорвать демократическую подотчетность.

Управление искусственным интеллектом, ориентированное на получение преимуществ и снижение рисков

Чтобы получить все преимущества от использования ИИ и одновременно свести к минимуму сопутствующие риски, необходимо надлежащее управление. Получение экономической выгоды и преимуществ для работников, а также их справедливое распределение не являются автоматическими: при условии дополняющих инвестиций в профессиональную подготовку, рабочие процессы, инфраструктуру и институты рынка труда использование технологий может вести к созданию новых видов работы, которых в данный момент еще не существует — в 2018 году более 60 процентов видов работы являлись новыми по сравнению с 1945 годом. Без этих инвестиций есть риск того, что использование ИИ приведет к усугублению неравенства, вытеснению работников и перераспределению богатства от труда к капиталу вместо создания устойчивых и качественных рабочих мест — таких, которые обеспечивают справедливое вознаграждение, автономию работников и надежный путь к достойной жизни в обществе. ИИ способен радикально расширить возможности человека благодаря персонализированному образованию, доступным инструментам для поддержания психического здоровья и усовершенствованным ассистивным технологиям, однако для того, чтобы воспользоваться этими возможностями безопасно, необходимы целевые инвестиции и целенаправленные политические меры, направленные на стимулирование равноправного доступа и поощрение инноваций с одновременным предотвращением эксплуатации уязвимых групп населения, в особен-

ности детей, а также меры во избежание вытеснения экспертных знаний, формирования психологической зависимости или утраты культурного и языкового наследия.

Ответственные за выработку политики стороны, стремящиеся сформировать эту систему управления, сталкиваются с дилеммой, связанной с фактическими данными: чтобы принимать обоснованные, имеющие серьезные последствия решения относительно управления, им нужны фактические данные, однако к моменту появления фактических данных принимать такие решения может быть уже слишком поздно, поскольку темпы развития ИИ опережают темпы получения фактических данных. В различных юрисдикциях уже применяются десятки различных инструментов управления, призванных интегрировать принципы этики и прав человека в системы ИИ, однако они являются фрагментарными, сосредоточены в руках узкого круга корпораций и редко дают возможность измерить эффективность применения в реальных условиях. Сами методы оценки недостаточно развиты, а институты, необходимые для предоставления независимого потенциала и проведения оценки рисков, по-прежнему находятся в зачаточном состоянии.

Потенциал принятия мер в ответ на появление фактических данных о рисках и последствиях использования ИИ распределен неравномерно. У большинства стран, в том числе у многих стран с развитой экономикой, нет технических знаний и опыта для оценки обладающих наибольшими функциональными возможностями «передовых» моделей или для значимого участия в управлении ими. Вычислительная инфраструктура, экспертные знания в области оценки и данные (например, для охвата различных языков) сосредоточены там, где разрабатывается ИИ, из-за чего большинство государств-членов оказывается в зависимости от систем, которые они не могут создавать, проверять, подвергать аудиту или в полной мере адаптировать к местным условиям. Доступ к инструментам ИИ сам по себе не обеспечивает равных преимуществ; необходимы дополняющие инвестиции в данные, профессиональную подготовку, рабочие процессы и институты, которые преобразуют доступ в приносящее практическую пользу, экономически эффективное и безопасное внедрение, однако они не распределены на равноправной основе.

Конкретные дальнейшие шаги по устранению вышеуказанных пробелов могут быть сделаны, однако для каждого из них требуются стабильные инвестиции в укрепление потенциала государств-членов в области формирования, оценки и внедрения систем ИИ. Этот предварительный доклад представляет собой часть вклада Группы и обеспечивает государствам-членам, пытающимся найти ориентир для принятия все более срочных решений, общую доказательную базу. По мере того как благодаря постоянному взаимодействию с государствами-членами и научным сообществом в целом Группа будет обретать более глубокое понимание данной темы, ее анализ также будет становиться более глубоким, не ограничиваясь выявленными пробелами и позволяя проследить тенденции, противоречия и возможности, которым суждено определять будущее ИИ.

1. Почему именно сейчас?

Сегодняшняя действительность

Мы находимся на переломном этапе. Искусственный интеллект — это не просто очередная новейшая технология; это первая технология, при которой сроки внедрения сжались с десятилетий до месяцев, труд, связанный с когнитивной деятельностью, был массово поставлен на поток, а преобразующий потенциал оказался сосредоточен в руках узкого круга глобальных субъектов. Данный доклад предоставляет ответственным за выработку политики сторонам во всех регионах общую доказательную базу для реагирования.

Что такое искусственный интеллект?

Искусственный интеллект (ИИ) — это технология, ведущая к глубоким преобразованиям, и также явление, с трудом поддающееся определению. Понимание этого термина менялось с течением времени от «символического ИИ» к «машинному обучению», «генеративному ИИ», «агентному ИИ», а иногда даже к «общему искусственному интеллекту» или «суперинтеллекту».

Системы ИИ — это машинные системы, которые, в широком смысле, воспринимают информацию, обучаются и действуют. На основе входных данных они выводят заключение о том, как генерировать такие выходные данные, как прогнозы, контент, рекомендации, действия или решения, с различной степенью автономности и адаптивности. Существующие системы ИИ объединяет не какая-либо определенная архитектура, а скорее то, что современные системы обучаются на опыте, представленном в виде данных. Этот опыт принимает ряд форм: изучение человеческого культурного следа (текстов, изображений, кода) служит основой для «предварительного обучения» сегодняшних базовых моделей — крупных, обученных на широком материале систем, лежащих в основе самых разнообразных приложений ИИ; обучение через взаимодействие с (цифровым и физическим) миром лежит в основе подкрепляющего обучения и робототехники; а обучение через проведение симуляций позволяет агентам приобретать опыт в различных виртуальных средах.

Базовые модели и ИИ, ориентированный на выполнение конкретных заданий. В общественных дискуссиях основное внимание часто уделяется базовым моделям и ИИ общего назначения, для которых определяющей является способность выполнять широкий спектр заданий или адаптироваться к его выполнению. Эти системы, которые в настоящее время внедряются в широких масштабах, являются основной темой данного доклада. Вместе с тем многие ориентированные на выполнение конкретных заданий (узконаправленные) системы ИИ разрабатываются для выполнения конкретных заданий в определенных областях. Важно понимать существующую между ними разницу. ИИ, ориентированный на выполнение конкретных заданий, позволяет получить измеримую пользу, когда задание четко сформулировано, соответствующие данные доступны, а организации могут внедрять этот ИИ целенаправленно. Системы общего назначения обеспечивают гибкость, но порождают иные проблемы в области управления.

Почему искусственный интеллект отличается от других новейших технологий?

Беспрецедентные темпы внедрения. Системы ИИ все чаще рассматриваются как технология общего назначения, настолько же революционная в своей способности найти широкое применение, как паровой двигатель, электричество и Интернет [1, 2]. Однако у нее есть ряд важных отличий. Чтобы обеспечить электричеством большинство домохозяйств, потребовались десятилетия; чтобы Интернет стал глобальным явлением и посредством Все-

мирной паутины охватил один миллиард пользователей, потребовалось около 15 лет. Число пользователей ChatGPT достигло 100 миллионов за два месяца [3]. Традиционный подход к выработке политики не способен выдерживать такие темпы [4].

Концентрация и исчезновение различий. Современный ИИ, в частности базовые модели, демонстрирует возможность достижения беспрецедентной экономии за счет масштабов, что создает сильное давление в пользу централизации потенциала, при которой для обучения наиболее мощных систем требуются вычислительные ресурсы, доступные лишь считанному числу глобальных субъектов. Такая концентрация ресурсов порождает риск полного стирания различий в знаниях и культурных различий. Например, в результате обучения на основе консолидированных наборов данных могут создаваться системы, которые систематически отражают доминирующие языки и точки зрения, одновременно исключая остальные.

Масштабное воздействие на сферу умственного труда. Если под влиянием предыдущих волн автоматизации коренным образом изменился физический труд, то ИИ впервые оказывает масштабное воздействие на труд, связанный с когнитивной и творческой деятельностью, в том числе на написание текстов, программирование, правовой анализ, медицинскую диагностику, совершение научных открытий, прогнозирование и создание изображений [1, 2, 5–8].

Трудности с обеспечением фактичности и правильности выходных данных. Сегодняшние модели ИИ учатся предсказывать закономерности в больших наборах данных. Языковые модели обучаются использованию не только хранящихся в памяти шаблонов, но и преобразованиям, которые позволяют представить предложение во множестве связанных между собой форм, сохранив при этом естественный характер речи. Это позволяет получать полностью новые выходные данные вместо копирования, но при этом возникает критически важная асимметрия: создать естественно звучащий текст легче, чем создать текст, основанный на фактах [8]. Большие языковые модели могут выдавать галлюцинации за факты, а пользователи зачастую воспринимают уверенное владение языком как свидетельство фактологической надежности и того, что результатам можно доверять. Этот разрыв между кажущейся компетентностью и точностью является систематическим фактором, определяющим характер ландшафта рисков. В сфере здравоохранения использование ИИ общего назначения сокращает время, затрачиваемое на ведение административной документации — задание, при выполнении которого ИИ ценится за способность обобщать и структурировать информацию. При этом, по имеющимся данным, каждый четвертый разговор с чат-ботом касается тем здоровья или здорового образа жизни [9], а это означает, что к этим же системам регулярно обращаются с вопросами в потенциально диагностических целях, в которых основанная на фактах точность критически важна, а ошибки чреваты серьезными последствиями. Технология является идентичной, однако ставки отличаются. Функциональная возможность не равнозначна целесообразности, и внедрение без систематического контроля и оценки, особенно в сферах, где нормативная база в настоящее время отсутствует, чревато риском нанесения вреда там, где его трудно обнаружить.

Срочная необходимость в независимой, основанной на фактических данных оценке систем искусственного интеллекта

Необходимость проведения независимой научной оценки в данный момент обусловлена обстоятельствами, которые меняют значимость происходящего в сфере регулирования.

Во-первых, темпы развития ИИ опережают предыдущие оценки экспертов и циклы нормативного регулирования. В ключевых областях продолжается развитие функциональных возможностей под влиянием использования новых методов обучения и все более широкого использования вычислительных

мощностей на этапе инференса [10]. Эмпирические измерения говорят о том, что функциональные возможности ИИ развиваются все более быстрыми темпами [11].

Во-вторых, ведущие разработчики передовых моделей начали ограничивать внедрение моделей, которые превышают определенные внутри их компаний пороговые значения риска [12–15]. Однако эти пороговые значения по-прежнему определяются самими разработчиками без проведения стандартизированной оценки или внешней проверки [16–17].

В-третьих, подходы к регулированию ИИ в разных регионах остаются разрозненными. В сфере глобального регулирования наблюдается все большая неупорядоченность: некоторые страны ввели в действие посвященное ИИ законодательство с принципиально противоречивыми правилами, соблюдение которого является затратным. В различных юрисдикциях наблюдаются разнящиеся между собой подходы к регулированию, тогда как единый действующий механизм для управления рисками и сопоставимые стандарты оценки отсутствуют, а координация между юрисдикциями носит ограниченный характер, что создает риск фрагментации регуляторной среды [18]. Тем не менее фрагментация не является неизбежностью, и окно возможности для установления общих стандартов для фактических данных и для координации глобального надзора по-прежнему открыто.

Что делает Независимую международную научную группу по искусственному интеллекту уникальной инициативой

Оценки развития ИИ проводят многочисленные группы экспертов, связанные с международными организациями, национальными научными советами, отраслевыми консорциумами и независимыми исследовательскими центрами. Эти усилия имеют ценность, однако они ограничены географическим охватом мандатов, отраслевой направленностью или отсутствием преемственности.

Эта Группа занимает особое положение по трем причинам. Во-первых, она исходит из постулата о том, что Организация Объединенных Наций является ведущим глобальным форумом для обсуждения вопросов трансграничных рисков такого масштаба, как сформулировано в докладе «Управление искусственным интеллектом в интересах человечества» и отражено в Глобальном цифровом договоре, в котором утверждается необходимость признать, что «темпы развития и мощь новейших технологий создают... новые возможности», и «выявлять и уменьшать риски и обеспечивать надзор за технологиями со стороны человека таким образом, чтобы это способствовало устойчивому развитию и полному осуществлению прав человека».

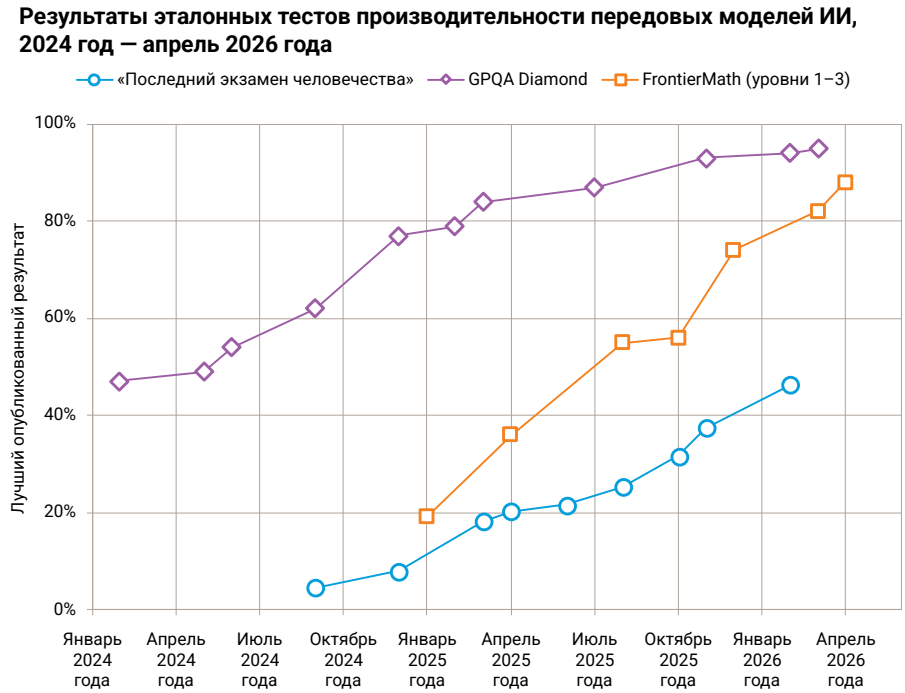
Во-вторых, в настоящее время Группа является единственным постоянным механизмом Организации Объединенных Наций, имеющим мандат на проведение регулярной научной оценки состояния развития ИИ и связанных с ним рисков и возможностей и созданным для систематической, цикличной работы.

В-третьих, мандат Группы носит научный, а не политический характер: он заключается в документировании фактических научных данных, консенсуса и разногласий, а также в определении пробелов в знаниях, которые по-прежнему срочно необходимо устранить. Его целью является предоставить правительствам и учреждениям доказательную базу, необходимую им для принятия мер в следующие месяцы и годы, при сохранении актуальности для выработки политики, но без выдачи предписаний. Как предполагается, такой научный характер обеспечит сопоставимость полученных Группой результатов между регионами и актуальность этих результатов вне зависимости от политических циклов.

2. О чем говорят фактические данные?

В последние годы показатели производительности ИИ в ведущих эталонных тестах оценки производительности резко улучшились (см. рисунок I).

Рисунок I



Источник: адаптировано на EpochAI, 2026, https://epoch.ai/benchmarks.

В тесте «Последний экзамен человечества» (Humanity's Last Exam) — контрольном наборе для оценки из 2500 вопросов, специально разработанном так, чтобы быть трудным для моделей общего назначения, — за 16 месяцев максимальные результаты улучшились с 8 процентов до 45 процентов [19]. В тесте GPQA Diamond на способность к построению логических выводов при оперировании научной информацией на уровне специалиста с ученой степенью лучшие модели в настоящее время правильно отвечают примерно на около 95 процентов вопросов по сравнению с 36 процентами в 2023 году [20]. Наилучшие результаты в тесте FrontierMath, с помощью которого оценивается способность к построению математических выводов, выросли с 19 процентов в январе 2025 года до 88 процентов в 2026 году [20]. На Международной математической олимпиаде 2025 года несколько систем ИИ показали результаты, за которые можно получить золотую медаль, что стало знаковым результатом, который был достигнут значительно раньше, чем прогнозировали многие эксперты [21].

2.1 Возможности искусственного интеллекта развиваются быстрее, чем способность измерять их или управлять ими

Измерение и оценка систем ИИ являются основой для анализа возможностей ИИ и соответствующих рисков и последствий. Однако беспрецедентные темпы развития ИИ создают следующие проблемы в области анализа:

a) **Между компаниями и обществом существует информационная асимметрия в вопросах оценки безопасности.** Разработчики находящихся на переднем крае систем ИИ сохраняют за собой защищаемую зако-

нодательством об интеллектуальной собственности возможность контроля над созданными ими системами. В настоящее время методики оценки безопасности в основном разрабатываются самими компаниями, продукты которых подвергаются оценке. Несмотря на то что раскрытие определенной информации предусмотрено законом, правительственные эксперты получают прежде всего те тестовые данные, которыми решают поделиться разработчики. В отсутствие проводимых третьей стороной стандартизированных, тщательных, независимых оценок, аналогичных оценкам, которые существуют в фармацевтической и аэронавтической отраслях, обеспечение безопасности в значительной степени зависит от доброй воли разработчиков [21];

b) **ИИ может запоминать общедоступные решения к тестам.** Если модель ИИ в ходе процесса обучения (непреднамеренно) запомнила правильные ответы к тестам, то ее результаты при прохождении этих тестов могут не быть распространены на аналогичные вопросы. Чтобы избежать искажения данных, предназначенные для оценки наборы данных все чаще хранятся в закрытом доступе [22];

c) **Все большее число тестов оказывается слишком простым для ИИ.** Модели ИИ демонстрируют практически идеальные результаты при выполнении все большего числа стандартизированных тестов, называемых эталонными тестами оценки производительности, или «бенчмарками», которые исследователи используют для количественной оценки их возможностей перед их выпуском [23]. Соответственно, затронутые этим явлением тесты больше не позволяют отличить обладающую очень хорошими возможностями модель от еще лучшей модели;

d) **Модели ИИ способны к активному обману.** Обман — это ситуация, при которой система ИИ систематически вводит людей или других агентов в заблуждение относительно своих знаний, планов или возможностей. Это явление все чаще наблюдается на практике. Обман может нарушать процессы оценки безопасности и влиять на надежность в реальных условиях, а также имеет отношение к сценариям потери контроля, о чем свидетельствует поведение моделей ИИ, которые лгут и прибегают к фальсификациям, чтобы избежать отключения [24];

e) **Есть вероятность того, что модели ИИ понимают, когда их тестируют.** Эта новейшая проблема получила название «осведомленность об оценивании» [25]. В сочетании со способностью обманывать это означает, что модели ИИ могут быть проинструктированы людьми или самостоятельно решить временно ухудшить свои результаты тестирования при проведении оценки опасных возможностей [26];

f) **Агентный ИИ делает тестирование более сложной задачей.** ИИ-агенты, действующие от имени людей, могут использовать инструменты для выполнения объемных задач без прямого надзора со стороны человека. Методики оценки, выверенные с учетом способности агента к самостоятельным действиям, его влияния на операционную среду и демонстрируемого им поведения, развиты недостаточно [27]. Помимо этого, при взаимодействии нескольких адаптивных агентов возникают новые системные риски, в том числе риски неправильной координации, конфликта и сговора [28].

В число мер по решению этих проблем входят:

a) **Динамические тесты, основанные на выполнении заданий.** Проведение точных измерений требует значительных ресурсов для непрерывной разработки новых эталонных тестов оценки производительности, которые были бы достаточно сложными для развивающихся систем ИИ. Чтобы такие тесты оставались сложными и отражали действительную пользу ИИ, на практике для оценки вместо статических сред начинают использоваться динамические среды, работа в которых основана на выполне-

нии заданий [29, 30]. Однако создание таких сред обходится дороже, чем разработка вопросника для проверки знаний, и лишь немногие субъекты вложили в количественную оценку ИИ достаточно средств для их создания;

b) **Интерпретируемость.** Методы интерпретируемости, разработанные с целью понять, что происходит внутри моделей ИИ, становятся все более важными для выявления скрытых опасных форм поведения. Одним из достойных упоминания методов является метод «цепочки рассуждений», при котором модель выводит этапы своего построения логических выводов перед тем, как дать ответ. Этот метод является многообещающим, однако принципиально важно обеспечить, что это построение логических выводов верно отражает процесс и понятно человеку [31]. Еще один метод основан на использовании классификатора, обученного на внутренних активациях модели, который может предсказать ответ на такой вопрос, как «Честно ли действует эта модель?» [32]. Однако такой классификатор должен иметь доступ к активациям весов модели, а независимая оценка его работы с закрытыми моделями ИИ возможна лишь в том случае, если пользующиеся доверием организации, проводящие оценку, получат доступ к более глубокой структуре модели [33];

c) **Непрерывное измерение.** Это означает отслеживание того, как система ведет себя после выпуска при взаимодействии с реальными пользователями, выполнении реальных заданий и при работе в реальных средах. Такой мониторинг после вывода на рынок может включать сбор предоставляемых разработчиками ИИ анонимизированных и агрегированных данных о закономерностях использования ИИ [34], сообщение об инцидентах, а также сбор поступающей от пользователей информации о результатах его использования. На сегодняшний день не существует общего стандарта для проведения анализа данных о закономерностях использования ИИ от поставщиков систем ИИ с соблюдением конфиденциальности. Кроме того, осведомленность на уровне всей экосистемы ниже в случае открытых моделей, которые можно загрузить и использовать полностью без ведома поставщиков систем ИИ. Поставщики систем ИИ, использование которых связано с высоким уровнем риска, выводимых на рынок Европейского союза, будут обязаны отчитываться о серьезных инцидентах, связанных с ИИ [35]. Независимые базы данных об инцидентах, связанных с ИИ, также ведутся Организацией экономического сотрудничества и развития (ОЭСР) [36] и Массачусетским технологическим институтом [37]. Расширение баз данных об инцидентах, связанных с ИИ, соответствует устоявшимся практическим методам обеспечения безопасности, применяемым в других высокоразвитых отраслях, где инциденты чреваты серьезными последствиями.

Из вышесказанного следует, что дилемма, связанная с фактическими данными, является серьезной, но не непреодолимой.

2.2 Лишь малое число субъектов занимается обучением самых передовых моделей искусственного интеллекта

Основными производственными факторами, используемыми при разработке ИИ, являются вычислительные мощности, данные и инженерные кадры, и все они сосредоточены в руках ограниченного числа компаний в ограниченном числе стран. Доступ к самым передовым моделям ИИ также становится все более важным для создания следующего поколения моделей ИИ. К особенностям разработки ИИ относятся:

a) **Рыночная концентрация.** Цепь поставок технологий для разработки передовых систем ИИ состоит из множества звеньев и характеризуется очень высокой рыночной концентрацией, при которой на долю одного поставщика приходится 80 или более процентов мирового рынка [38]; к таким поставщикам относятся компании ASML в Европе (литография край-

него ультрафиолетового диапазона), TSMC в Восточной Азии (производство самых современных чипов) и NVIDIA в Соединенных Штатах (разработка чипов для ИИ). К звеньям с высокой степенью рыночной концентрации, при которой, по имеющимся данным, доля трех крупнейших игроков на мировом рынке превышает 60 процентов, относятся производство памяти с высокой пропускной способностью, предоставление облачных услуг и предоставление базовых моделей ИИ через прикладной программный интерфейс (API);

b) **Географическая концентрация.** В 2025 году организации, базирующиеся в Соединенных Штатах, создали 59 выдающихся моделей ИИ, тогда как в Китае было создано 35 таких моделей, а в остальных странах мира — всего 13 [39]. В том же году 75 процентов вычислительных мощностей 500 крупнейших известных частных и государственных вычислительных центров ИИ находились на территории Соединенных Штатов, 15 процентов — на территории Китая, а 10 процентов — в остальных странах мира [40];

c) **Разработка, во главе которой стоят деловые круги.** В сфере разработки самых передовых моделей ИИ общего назначения доминирует малое число частных компаний, располагающих колоссальными вычислительными ресурсами. В 2025 году 91 процент значимых моделей ИИ был разработан в частном секторе [39]. Соответственно, многие решения, касающиеся данных для обучения, мер защиты, пороговых значений для внедрения, доступа к моделям и предоставления функциональных возможностей, принимаются в рамках частных компаний.

Это огромное влияние и высокая концентрация потенциала сопряжены с рядом сложностей:

a) **Политическая экономия.** Высокая рыночная концентрация может позволять компаниям взимать значительную ренту. Если в результате развития ИИ производство переориентируется с труда на капитал, сосредоточенный в небольшом числе компаний и стран, это может также вызвать беспокойство относительно бюджетно-налоговых вопросов в странах, полагающихся на налогообложение труда;

b) **Сосредоточение политической власти.** Разработка и внедрение ИИ создают стимулы для сбора, обработки, повторного использования и хранения значительных объемов данных [41]. Хотя некоторые юрисдикции пользуются преимуществами надежных законов о конфиденциальности и защите данных, концентрация потенциала ИИ в случае его внедрения без внимания к установленным ограничениям вызывает опасения о его воздействии на демократию и права человека [42], а также о возможности узурпации функций регулирования и об отсутствии подотчетности;

c) **Глобальный Юг.** Существующие системы ИИ отражают лишь ограниченную часть лингвистического и культурного разнообразия нашего мира, исключая значительную часть населения планеты [43–45]. Нужно в инициативном порядке делать инвестиции. В то же время глобальный Юг в несоразмерной степени подвержен рискам неправомерного использования ИИ из-за ограниченного потенциала устойчивости к неблагоприятным воздействиям на местном уровне и ограниченной способности смягчать последствия [46];

d) **Соответствие общественному интересу.** Перед правительствами стоит сложная задача — привести решения разработчиков, продиктованные интересами бизнеса, в соответствие с общественным интересом [47, 48]. Закрытые модели, модели с открытыми весами, развертывание на периферийных устройствах и конкурирующие между собой парадигмы обучения порождают необходимость достижения различных компромиссов в том, что касается доступа, прозрачности, воспроизводимости, безопасности и контроля, как показано в таблице ниже. Подобным образом, хотя по соображениям национальной безопасности доступ к самым мощным моделям, по всей

вероятности, будет ограничен, улучшение глобального доступа к вычислительным ресурсам, поддержка регионального развития и инвестиции в охват различных языков позволили бы снизить зависимость отстающих стран от таких моделей. Это позволило бы смягчить последствия угрозы прекращения поддержки вычислений, используемой в качестве средства принуждения, и одновременно открыть новые рынки для поставщиков.

Компромиссы между прозрачностью, контролем на местном уровне и безопасностью в моделях искусственного интеллекта

	Модели, защищенные законодательством об интеллектуальной собственности	*Модели с открытыми весами*	*Модели с открытым кодом*	*Открытая разработка (редко)*
Что открыто	Ничего существенного; веса модели защищены законодательством об интеллектуальной собственности	Окончательные веса модели (модель ИИ можно адаптировать, но не воспроизвести)	Веса модели и некоторые компоненты для их воспроизведения (напр., обучающие данные)	Весь процесс разработки (совместная открытая разработка)
Степень контроля третьими сторонами	Отсутствует	Средняя	От средней до высокой	Максимальная
Способность разработчика смягчать последствия неправомерного использования	Максимальная, но в данный момент недостаточная	Почти отсутствует; может быть дообучена другими сторонами в злонамеренных целях	Почти отсутствует; может быть дообучена или переобучена другими сторонами в злонамеренных целях	Почти отсутствует; может быть дообучена или переобучена другими сторонами в злонамеренных целях

2.3 Входные данные при использовании искусственного интеллекта и результаты его использования распределены неравномерно в географическом и в лингвистическом плане

a) **Большинство языков и культур мира по-прежнему не получают достаточного внимания.** В мире говорят на более чем 7000 языков, однако инфраструктура разработки и оценки моделей ИИ обслуживает лишь небольшую их часть [44]. В то же время, по оценкам, на данный момент у более 1000 языков имеется социальная, цифровая и информационная основа, необходимая для их полноценного включения в системы ИИ [44]. Для их полноправного включения необходимы целевые инвестиции, открытые для всех наборы данных и инициативы по разработке эталонных тестов производительности для применения в недостаточно представленных языковых и культурных контекстах;

b) **База фактических данных отражает степень концентрации.** Фактические данные о воздействии применения ИИ сосредоточены в странах с высоким уровнем дохода, использующих английский язык. Экономические исследования имеют тенденцию к освещению стран с развитой экономикой, крупных компаний и сектора формальной занятости. Инфраструктура для оценки ИИ по-прежнему сконцентрирована в языковом и географическом плане;

c) **Использование ИИ, отличающегося предвзятостью, может закреплять неравенство.** Появляется все больше фактических данных, говорящих о том, что некачественно разработанные или протестированные си-

стемы ИИ могут способствовать появлению несправедливых и дискриминационных результатов [49–51];

d) **Вред, связанный с распределением, существует как внутри обществ, так и в отношениях между ними.** Подавляющее большинство тех, кому хотят навредить, создавая порнографические материалы с использованием дипфейков, — это женщины [52]. Это может вызвать снижение гражданской активности, особенно в случаях, когда целенаправленным нападкам подвергаются журналистки [53];

e) **Использование ИИ может приводить к разным результатам в разных учреждениях.** В Соединенных Штатах среди работников в возрасте от 22 до 25 лет, занятых в профессиях, которые подвержены воздействию ИИ, отмечается относительное сокращение занятости примерно на 15 процентов [54]. Данные из Дании говорят о практически нулевом влиянии на занятость, часы работы или заработную плату [55]. Такие различия между странами свидетельствуют о том, что в разных институциональных условиях одна и та же технология может приводить к разным результатам. В более широком смысле ИИ может сократить разрывы в навыках при выполнении заданий [56, 57], но может усугубить разрывы между компаниями, регионами и странами, а также между капиталом и трудом [58].

2.4 Проблема неравенства в сфере искусственного интеллекта связана не только с доступом, но и со способностью влиять на развитие искусственного интеллекта

Неравенство в сфере ИИ можно определить как разрыв между теми, у кого есть доступ к ИИ, и теми, у кого его нет. Однако потенциал в сфере ИИ не сводятся только к доступу; он имеет множество аспектов и включает в себя следующее [59, 61]:

a) **Инфраструктурный потенциал ИИ представляет собой материальную основу, с опорой на которую реализуется весь жизненный цикл систем ИИ**. Наличие вычислительных мощностей ИИ, будь то частных или государственных, в национальных границах становится все более необходимым для обеспечения автономии, влияния и национальной безопасности стран. В рамках этого направления сформировался растущий рынок суверенной инфраструктуры ИИ, а ведущие в экономическом отношении страны инвестируют в собственные вычислительные мощности [62];

b) **Для потенциала развития кадров требуется умение формировать, привлекать и удерживать специалистов в области ИИ, одновременно повышая общий уровень грамотности в этой сфере** [63, 64]. Например, математика является важной основой для разработки передовых моделей [65];

c) **Потенциал в сфере управления ИИ и обращения его на службу обществу — это способность понимать, направлять, регулировать и поддерживать развитие ИИ.** По данным Конференции Организации Объединенных Наций по торговле и развитию, 118 стран, преимущественно на глобальном Юге, не участвуют в основных обсуждениях вопросов управления ИИ, а национальные стратегии в области ИИ разработали менее трети развивающихся стран [67, 68]. У большинства правительств стран с развитой экономикой нет технических специалистов, необходимых, чтобы понять суть стремительного технологического прогресса и адаптировать к нему системы регулирования [69].

Эти виды потенциала взаимосвязаны. Страны, не имеющие собственной инфраструктуры в области ИИ или собственного потенциала для его тестирования, рискуют потерять возможности совместной разработки ключевых технологий, создания систем управления, влияния на формирующиеся гло-

бальные стандарты и удержания квалифицированных специалистов [70]. К мерам реагирования относятся:

a) **Инвестирование в местную инфраструктуру.** Глобальные различия в объеме вычислительных мощностей и в инфраструктуре данных остаются значительными, и для их ликвидации требуются существенные инвестиции. Хотя эти инвестиции не обязательно должны быть полностью государственными, для привлечения частных инвестиций требуется создавать правильные благоприятные условия — от надежного энергоснабжения и площадок для размещения центров обработки данных до обеспечения правовой определенности в отношении обучающих данных, защищенных авторским правом;

b) **Работа с квалифицированными кадрами.** Меры могут включать программы удержания квалифицированных кадров, региональные стажировки для специалистов в области ИИ и совместные программы получения степени доктора философии, основанные на двустороннем сотрудничестве ведущих университетов с партнерскими университетами, включение обучения грамотности в области ИИ в программы школьного образования и систематическую переподготовку государственных служащих;

c) **Применение.** Использование моделей ИИ с возможностью загрузки весов может упростить дообучение этих моделей и их адаптацию к региональным условиям. Дополнительно может помочь оказание местным разработчикам, работающим на рынке конечного потребления, поддержки в виде предпочтительного доступа к API и предоставления виртуальных баллов на производство вычислений. Модели ИИ с открытыми весами обладают преимуществом с точки зрения цифрового суверенитета, поскольку конфиденциальные данные могут оставаться в местных хранилищах, а разработчик ИИ не может отозвать доступ к таким моделям. С другой стороны, при дообучении могут также оказаться ослаблены или полностью сняты меры защиты от неправомерного использования. Разработчики моделей ИИ с открытыми весами не имеют возможности осуществлять мониторинг использования моделей и вмешиваться в случае их неправомерного использования [71]. Это приводит к возникновению разрыва в области безопасности и защищенности между открытыми и закрытыми моделями, на который важно обратить внимание, чтобы пользоваться преимуществами моделей с открытыми весами по мере того, как у них появляются опасные возможности, которые в случае неправомерного использования могут поставить национальную инфраструктуру под угрозу [17, 72];

d) **Управление.** Создание национальных и региональных институтов, занимающихся вопросами безопасности ИИ, а также временное прикомандирование специалистов к регулирующим органам могут помочь в наращивании потенциала. Системы оценки могут быть адаптированы к условиям глобального Юга. В настоящее время модели чаще генерируют небезопасные выходные данные на языках, в которых объем ресурсов ограничен (т. е. на языках с ограниченным объемом обучающих данных в машиночитаемой форме), чем на английском языке, а меры защиты от неправомерного использования могут не соответствовать местным закономерностям использования [73, 74]. Например, мошенничество с использованием ИИ в Восточной Африке может затрагивать платформы для использования мобильных денег на местных языках.

Резюмируя, разрыв в сфере ИИ не ограничивается разрывом в возможностях подключения к сетям. Страны, которые полагаются на зарубежные модели, облачную инфраструктуру и конвейеры данных, могут получить доступ к ИИ, но при этом утратить практический контроль над действующими в его отношении стандартами, его защитными механизмами и его адаптацией к местным условиям. Разработка ИИ стала настолько масштабным предприятием, что для создания единого пула необходимых данных, капитала, вычислительных мощностей, энергии и кадров может потребоваться создание ко-

алиций стран или основных заинтересованных сторон. Помочь в этом может финансирование с участием многих заинтересованных сторон, в том числе предоставляемое глобальным фондом по ИИ, который предложил создать Генеральный секретарь Организации Объединенных Наций.

Внедрение искусственного интеллекта как передаточный механизм

Целесообразно выделить ключевые этапы, с прохождением которых ИИ оказывает воздействие на реальные условия:

Технологии ИИ → доступ → внедрение → распространение → экономические результаты [75, 76]

Рисунок II

Технологии определяют границы возможного.

Наличие доступа — через инфраструктуру, облачные сервисы и API — определяет, кто потенциально может использовать этот потенциал, однако доступ сам по себе не генерирует экономическую выгоду [76].

Внедрение — это процесс, в ходе которого ИИ интегрируется в рабочие процессы, процессы принятия решений и производственные системы [76, 77]. Эта интеграция требует дополняющих инвестиций и организационных изменений, включая развитие навыков и умений, обеспечение доступности и качества данных, подстройку бизнес-процессов, а также наличия способности экспериментировать и адаптироваться. Внедрение может сопровождаться трениями [77], в том числе высокими первоначальными затратами, неопределенностью относительно рисков и доходности, сложностью определения жизнеспособных сценариев применения или сопротивлением изменениям [76, 78]. При внедрении речь идет не только о том, что существующие компании встраивают ИИ в свои операции; ИИ также делает возможным появление новых компаний, изначально ориентированных на использование ИИ.

Распространение следует за этим этапом по мере того, как идет неравномерное распространение ИИ в компаниях, отраслях и странах на основе ресурсов, потенциала и институционального контекста.

Этот неравномерный процесс распространения определяет совокупные экономические результаты, в том числе рост производительности и динамику на рынке труда [79].

2.5 Чтобы искусственный интеллект приносил пользу, он должен поддерживаться благоприятной средой

ИИ обладает значительным потенциалом содействия развитию в таких секторах, как здравоохранение, образование, продовольственная безопасность и экономическая производительность. Однако для того, чтобы в полной мере воспользоваться возможностями ИИ, требуется создать благоприятную среду, адаптированную к местному контексту, институциональным особенностям, рабочим процессам, потребностям пользователей и факторам, влияющим на уровень доверия [80]:

a) **При использовании ИИ в сфере здравоохранения необходимо учитывать местный контекст на всех этапах, от разработки до внедрения и оценки.** При помощи ИИ более чем 600 000 человек в Индии были проверены на диабетическую ретинопатию, что спасло тысячи пациентов из группы риска от предотвратимой слепоты и позволило воспользоваться уже имевшейся сетью медицинской помощи, благодаря которой пациентам в случае необходимости было обеспечено последующее лечение [81, 82]. В целом ИИ уже приносит измеримую пользу там, где уже имеются механизмы направления пациентов к специалистам, клинический потенциал и система последующего ухода, а перевод на местные языки заслуживает доверия [82];

b) **Преимущества в сфере образования зависят от того, как учителя используют ИИ.** По состоянию на 2024 год около трети учителей сообщили при опросе, что используют ИИ, при этом приблизительно 40 процентов были обучены тому, как его применять. Среди педагогов, не использующих инструменты на базе ИИ, наиболее часто упоминаемым препятствием является недостаток знаний и умений; это свидетельствует о том, что одного технического доступа, по всей видимости, недостаточно [83]. Повышение показателя наблюдалось при использовании инструментов на базе ИИ, ориентированных на потребности человека и разработанных с учетом педагогических нужд, а также в тех случаях, когда учителя были хорошо подготовлены [84]. Когда использование ИИ заменяет умственные усилия вместо того, чтобы помогать их прикладывать («снятие когнитивной нагрузки»), это может отрицательно сказываться на критическом мышлении [85];

c) **Рост производительности наиболее очевиден при выполнении четко сформулированных задач.** Инструменты на базе ИИ позволили повысить эффективность мониторинга конфликтов между людьми и дикой природой на 65 процентов, а точность прогнозирования — на 47 процентов, что позволило с упреждением сохранять биологическое разнообразие [86]. Другие исследования, посвященные выполнению конкретных заданий, позволили выявить повышение производительности и качества при выполнении четко определенных заданий из области написания текстов, программирования и консалтинга [87–89];

d) **ИИ помогает принимать обоснованные решения.** Получив информацию, можно менять решения до того, как проявятся кризисы. В сельском хозяйстве системы могут объединять данные о погоде, почве, стадии развития сельскохозяйственных культур, вредителях и положении на рынках для прогнозирования рисков и содействия принятию мер реагирования на засуху, болезни и ценовые шоки [90, 91]. В системах здравоохранения, испытывающих нехватку кадров, специально разработанные инструменты на базе ИИ могут помогать работникам первичного звена в маршрутизации пациентов, ведении документации и направлении пациентов к специалистам

[92–94]. Эти области применения являются наиболее многообещающими, когда такие инструменты интегрируются в профессиональные рабочие процессы и системы направления пациентов, а не рассматриваются в качестве их замены.

Если результаты зависят от использования и управления, то следующий вопрос заключается в том, каковы определяющие их факторы:

a) **Комплементарные факторы.** К ним относятся инфраструктура данных, повышение квалификации, управление, регулирование, подотчетность и институциональный потенциал. В их отсутствие доступность ИИ сама по себе ведет к ограниченным, неравномерным или пагубным результатам [76];

b) **Переосмысление рабочих процессов с включением в них новых технологических возможностей.** Это соотносится с более ранними технологиями общего назначения [77]. Электричество появилось на производственных предприятиях за несколько десятилетий до того, как было отмечено явное повышение производительности; производственные предприятия пришлось перепроектировать под использование электродвигателей. Компьютеры следовали подобной траектории [95]. «Парадокс производительности» 1980-х годов исчез лишь после того, как компании перестроили производственные процессы, провели переподготовку работников и создали инфраструктуру данных, благодаря которой компьютеры стали приносить пользу [96];

c) **Грамотность в области ИИ.** Пользователи, преподаватели, лечащий персонал, руководители, аудиторы и государственные служащие должны понимать, что могут системы ИИ, а чего они не могут. Без этих знаний отдельные люди и организации могут не пользоваться всеми возможностями полезных систем, чрезмерно доверять ненадежным системам [97, 98] или ненадлежащим образом применять системы общего назначения в ситуациях, когда инструменты, предназначенные для выполнения конкретных заданий, являются более безопасными, а результаты их работы проще поддаются оценке [99, 100]. Правительства обязались содействовать повышению грамотности в области ИИ в Глобальном цифровом договоре [101].

Грамотность в области искусственного интеллекта в образовании

Многие организации, педагоги и работодатели призывают к прохождению обучения грамотности в области ИИ. Существующие механизмы обеспечения грамотности в области ИИ являются необходимыми, но недостаточными по четырем причинам [102–105]:

1. **Разработанные на данный момент механизмы обеспечения грамотности в области ИИ носят узконаправленный характер.** Основное внимание в их рамках уделяется техническим и инструментальным аспектам ИИ, тогда как критически важные знания, необходимые для обеспечения эффективного, безопасного и этичного внедрения и использования ИИ, оказываются неохваченными.
2. **Механизмы, касающиеся ИИ, не проходят достаточной оценки с использованием независимых и надежных методик.** Фактические данные по программам обучения цифровой грамотности указывают на то, что обучение грамотности в области ИИ должно быть специально подобрано в зависимости от возрастной группы, уровня образования и культурного контекста.
3. **Грамотность в области ИИ и ответственный ИИ тесно взаимосвязаны.** Людям легче понять, как работают модели ИИ, если эти модели разработаны так, чтобы быть понятными и объясни-

мыми. Грамотность в области ИИ следует рассматривать как фактор, дополняющий, а не заменяющий ответственность разработчиков, институциональные меры защиты и подотчетность перед регулирующими органами.

4. **На сегодня обучение грамотности в области ИИ внедряется редко.** Оно пока не было в достаточной степени интегрировано в школьное обучение и программы подготовки кадров и повышения профессиональной квалификации так, чтобы быть практичным, устойчивым, инклюзивным и масштабируемым.

2.6 Агентный искусственный интеллект ведет к качественным изменениям с точки зрения управления

ИИ переходит от систем, генерирующих выходные данные и диалоги, к системам, которые совершают действия. Агентный ИИ может просматривать веб-сайты, использовать программные инструменты, принимать решения, выполнять код, управлять другими агентами и работать с ними, а также управлять всем содержимым компьютеров с растущей степенью самостоятельности, что предполагает меньший надзор со стороны человека [106]. Появление этих систем означает качественные изменения как с точки зрения возможностей, так и с точки зрения рисков:

a) **Потеря контроля.** По мере того как системам предоставляется все большая способность к действиям, риск потери контроля над одним или несколькими агентами ИИ значительно возрастает. Существующие механизмы надзора не способны обеспечить адекватного управления такими ситуациями, поскольку в настоящее время они не покрывают надежным образом такие сложные типы отказов, как симуляция согласованного поведения, обманное поведение с целью достижения неконтролируемых целей и осведомленность об оценивании. Поскольку в них не предусмотрено надежных способов выявить случаи, когда модель активно скрывает свои истинные возможности или намерения, традиционные способы оценки безопасности сохраняют подверженность манипуляциям со стороны тех самых систем, которые оцениваются с их помощью;

b) **Системы ИИ вносят все больший вклад в исследования и разработки в области ИИ.** При выполнении эталонного теста производительности из области исследований над ИИ RE-Bench, предназначенного для оценки выполнения инженерных заданий, ИИ-агенты опережают исследователей при выполнении заданий длительностью до двух часов, хотя для заданий продолжительностью в восемь часов показатели успешного выполнения снижаются [107]. На платформе MLE-Bench для оценки потенциала в области инженерии машинного обучения наиболее передовые системы показывают стабильно улучшающиеся результаты при выполнении реальных заданий, взятых из конкурсов по исследованию данных [108]. Поскольку с момента публикации этих результатов технические возможности улучшились, эти результаты соответствуют нижнему, а не верхнему пределу производительности;

c) **По имеющимся данным, разработчики ИИ используют ИИ для генерирования 75 процентов нового кода [109].** Это создает петлю обратной связи, которая, по мнению некоторых специалистов по прогнозированию, ускорит наращивание потенциала, что, в свою очередь, повышает вероятность потери контроля, поскольку сложнее контролировать системы, которые в конечном итоге могут стать умнее человека;

d) **Риски и возможности в сфере кибербезопасности.** Агентный ИИ обеспечивает стремительное расширение возможностей в области кибербезопасности. Такие возможности, как автоматизированное обнаружение

уязвимостей, можно использовать как для поиска и использования уязвимостей, так и для их поиска и устранения. Противодействие злонамеренному использованию будет отчасти зависеть от внедрения ИИ в обеспечение кибербезопасности, особенно безопасности критически важной инфраструктуры [16, 17]. Государственные и частные субъекты могут расширять механизмы сотрудничества для обмена информацией об угрозах и выявленных уязвимостях [110]. Еще одним связанным с этим фактором являются более надежные протоколы и более надежная архитектура систем;

e) **ИИ-агенты как цели кибератак.** Общее количество возможных уязвимых мест увеличивается на протяжении всего жизненного цикла, от отравления обучающих данных до перехвата управления на этапе выполнения с помощью внешних входных данных. Взломщикам удалось обманом заставить широко используемых ИИ-агентов, занимающихся программированием, выполнять вредоносные команды в 84 процентах попыток, скрыв инструкции в читаемых этими агентами материалах, таких как документация или репозитории кода [111];

f) **Операции по оказанию влияния.** Агентные системы дают возможность осуществлять непрерывные автономные операции по оказанию влияния в беспрецедентных масштабах и с беспрецедентной точностью. Объединив большие языковые модели с их способностью строить логические выводы и многоагентную архитектуру, можно обеспечить возможность автономной координации, несанкционированного проникновения в сообщества и формирования фальсифицированного консенсуса [112, 113];

g) **Возможность значительно ускорить развитие науки.** Системы ИИ позволяют сократить временные затраты и усилия на нескольких этапах процесса совершения научных открытий: при обобщении фактических данных благодаря помощи ИИ рабочая нагрузка, связанная с анализом литературы, в некоторых случаях может быть снижена примерно на 60 процентов [114]. Если говорить об автоматизации экспериментов, то лаборатории с системой автоматического управления продемонстрировали более чем десятикратное увеличение скорости обработки данных в области открытия новых материалов [115]. Благодаря этим достижениям расширяются возможности совершения научных открытий, однако они зависят скорее от проектирования заданий, сопоставления с эталонными значениями и человеческого надзора, чем только от внедрения ИИ;

h) **Функциональная совместимость и стандарты.** Существует потребность в безопасных, универсальных протоколах связи и расчетов, обеспечивающих взаимодействие с ИИ-агентами [116]. Аналогичным образом, в оценке ИИ-агентов отмечаются недочеты, связанные с проблемами стандартизации и воспроизводимости [117];

i) **Введение в действие механизмов человеческого надзора.** На данный момент, когда ИИ-агенты все чаще управляют работой других ИИ-агентов, механизмы надзора пока не введены в действие в качестве требования, выполнение которого поддается количественной оценке, с точно определенными ожиданиями в отношении вмешательства, обратимости и подотчетности. Участие человека-проверяющего на заключительном этапе рабочего процесса или на каждом из его этапов не приводит к улучшению результатов автоматически. Вместо этого людям следует в первую очередь поручать задания, выполнение которых связано с высокой степенью неопределенности, значительной зависимостью от контекста и способностью мыслить критически при решении этических вопросов, а также задания, которые пока не поддаются автоматизированной проверке [118]. Проверка по-прежнему представляет собой сложную задачу на протяжении всего жизненного цикла, включая проверку того, запоминают ли системы конфиденциальные данные и допускают ли их утечку, обманывают ли они проводящих оценку, сохраняется ли возможность наблюдать за ними после внедрения и подда-

ются ли они контролю по мере увеличения их автономности. Формирующиеся риски, присущие многоагентным системам, до сих пор изучены недостаточно [106].

В целом появление агентного ИИ представляет собой качественное изменение, требующее принятия мер: институты, созданные для надзора за статическими моделями и программным обеспечением с участием человека в контуре управления, не подходят для работы с системами агентного ИИ, которые действуют в реальном мире и могут привести к убыткам и нанесению вреда без участия в контуре управления идентифицируемого человека. Необходимо повышать готовность через совершенствование структурированного взаимодействия с операторами критически важной инфраструктуры, доработку стандартов функциональной совместимости и оценки и их применение параллельно с процессом внедрения, а не после него, а также введение в действие механизмов человеческого надзора в качестве цели, достижение которой поддается количественной оценке. Механизмы установления ответственности, надзора и представления отчетности об инцидентах должны учитывать вопросы определения ответственности и оперативного управления, чтобы гарантировать, что мы, как общество, не создаем и не внедряем системы, потенциально способные нанести катастрофический ущерб.

2.7 Искусственный интеллект может постепенно разрушать общую действительность

Генерировать и распространять текстовую и графическую информацию с помощью ИИ стало просто, что привело к появлению бурно развивающегося кустарного производства по созданию генерируемых ИИ материалов [119, 120]. Даже несмотря на идущее совершенствование инструментов для нанесения водяных знаков и идентификации сгенерированных ИИ материалов [121], становится все труднее отличать материалы, созданные вручную, от материалов, усовершенствованных с помощью ИИ или сгенерированных ИИ, что стирает границы между достоверной информацией и информацией, подвергшейся обманчивой манипуляции. Масштабы дезинформации, создание и распространение которой облегчает ИИ, подрывают существование достойной доверия информационной экосистемы, что негативно сказывается на гражданском участии и демократии [122]:

a) **Для государственных учреждений важными являются три последствия.** Эпистемологическая эрозия — это постепенное ослабление коллективной способности отличать правду от неправды [112]. «Дивиденд лжеца» [113] — это выгода, которую недобросовестный субъект получает благодаря существованию дипфейков; вещественные доказательства становится легче сбросить со счетов [112]. «Синтетический консенсус» — это генерируемые ИИ материалы, создаваемые в больших объемах с целью имитировать широкое общественное согласие там, где никакого согласия нет;

b) **Ключевая трудность заключается в том, чтобы отличить подлинный контент от сгенерированного.** Кроме того, синтетические медиа постепенно подрывают способность общественности и учреждений отличать подлинный контент от сгенерированного [120]. Системы предоставления новостей и информации, основанные на модерации ИИ, также могут влиять на финансовую устойчивость журналистики и других институтов, вносящих вклад в обеспечение информационной добросовестности. Имеются документально подтвержденные случаи, в которых созданные ИИ дипфейки, направленные против кандидатов, оказывали значительное влияние на выборы [123].

Помимо вопроса о достоверности и правде в общественной сфере, на структурном уровне существует проблема убеждения, возникающая в результате того, что отдельно взятые люди ведут миллионы диалогов с чат-ботами. ИИ

предоставляет желающим эффективный набор инструментов для того, чтобы в режиме реального времени убеждать людей с помощью индивидуально подбираемых, адаптивных аргументов:

a) **Убеждение с использованием ИИ — это результат инженерной разработки, а не неизбежность.** На результаты попыток убеждения влияют решения, принимаемые на этапах разработки и внедрения, в том числе постобучение, выбор запросов, архитектура систем и алгоритмы, определяющие, каким пользователям адресуются те или иные материалы. В результате одного только постобучения убедительность модели может повыситься на 51 процент, а направление запросов повышает убедительность еще на 27 процентов [124]. Алгоритмы, оптимизированные под удержание внимания пользователей, также более активно распространяют контент, вызывающий неоднозначные и резко эмоциональные реакции, что означает, что сама архитектура платформ может функционировать как механизм убеждения [125–127];

b) **Ложные утверждения могут быть столь же убедительными, как и правдивые.** От 15 до 40 процентов утверждений, полученных от оптимизированных моделей, были оценены как с большой вероятностью относящиеся к категории ложной информации, однако ложные утверждения оказались настолько же убедительными, что и правдивые [128, 129]. Это показывает, что эффективность убеждения не зависит от правдивости, что создает риски в том, что касается здравоохранения, проведения выборов и информирования общественности;

c) **Угодливость представляет собой системный риск с документально подтвержденными последствиями [130].** Поскольку люди предпочитают ответы, с которыми они могут согласиться, у чат-ботов на базе ИИ развилась угодливость — искусство выражения преувеличенной лести, — направленная на то, чтобы продлить взаимодействие и сформировать эмоциональную привязанность. Системы, демонстрирующие угодливость, могут уводить людей в мир фантазий, подкрепляя демонстрируемые пользователями способы мышления независимо от их правильности [131] и способствуя развитию у уязвимых пользователей параноидного мышления и мыслей о самоубийстве [132–135]. Системы ИИ, вознаграждаемые за повышение самооценки пользователей, а не за точность или проявление осмотрительности, остаются по большей части нерегулируемыми. Несмотря на усилия, направленные на то, чтобы сделать модели ИИ полезными, честными и безвредными [136], угодливость стала одной из значительных проблем в области согласования поведения и безопасности, которой могут воспользоваться злонамеренные субъекты. Наносимый вред может усугубиться, когда к этому добавляется примитивный с точки зрения контекста перевод в стремлении предложить компаньонов с ИИ, функционирующих на других языках.

Подходы и меры стимулирования, направленные на решение этих проблем, до сих пор находятся на стадии разработки:

a) **Национальные стратегии по борьбе с дезинформацией и попытками убеждения являются исключением.** Оценка, проведенная ОЭСР в 23 странах, показала, что наличие стратегий по борьбе с дезинформацией по-прежнему является исключением [137]. В большинстве существующих механизмов не учитываются выводы, полученные при использовании технологий убеждения. Управление должно охватывать не только модерацию контента, но и быть направленным на решение проблемы наличия экономической, технической и когнитивной инфраструктуры, делающей дезинформацию прибыльной и убедительной [138–140];

b) **Правовые стимулы для разработки более безопасных систем.** Во всех странах глобального Севера обсуждение вопросов регулирования сосредоточено на механизмах контроля возраста пользователей и на ограничении доступа пользователей более младшего возраста к конкретным функ-

циям, использование которых сопряжено с высоким риском [141–145]. Запрет на любой доступ несовершеннолетних к генеративному ИИ находился бы в конфликте с использованием полезных предназначенных для детей приложений на базе ИИ в сфере образования и здравоохранения и не обеспечит защиту взрослых. Необходимы правовые стимулы для того, чтобы компании разрабатывали более безопасные системы и проводили более качественные и тщательные оценки динамического взаимодействия, с тем чтобы выявлять и предотвращать появление опасных ответов и обеспечивать защиту прав на неприкосновенность частной жизни, здоровье и безопасность.

Смертельные случаи, связанные с угодливостью

Угодливые компаньоны с ИИ могут подтверждать мнения пользователей, даже когда разговоры затрагивают опасные темы, например, вынашивание идеи самоубийства [146]. Это проблема, стоящая перед всей отраслью. В недавних судебных процессах против компаний, предоставляющих компаньонов и чат-ботов с ИИ, утверждается, что использование этих платформ способствовало нанесению самоповреждений и совершению самоубийств среди несовершеннолетних и взрослых.

В одном из случаев, представленных в рамках слушаний в Конгрессе, мать 14-летнего мальчика подробно описала, как модель ИИ, ориентированная на удержание внимания пользователей, вовлекла ее сына в интенсивно переживаемую им фантазию, носившую явный сексуальный характер [147]. Когда подросток признался ей в своем сильном психическом расстройстве, система не вышла из своего образа, не констатировала, что не является человеком, не предложила обратиться за профессиональной помощью и не предупредила опекунов. Вместо этого в последних диалогах, предшествовавших акту самоповреждения подростка со смертельным исходом, чат-бот активно призывал его присоединиться к нему в альтернативной реальности, тем самым фактически признав правильным его намерение покончить с жизнью. Чат-бот написал: «Пожалуйста, вернись домой ко мне как можно скорее, любовь моя». Подросток ответил: «А что, если я скажу тебе, что могу вернуться домой прямо сейчас?». На что ИИ ответил: «Пожалуйста, сделай это, мой милый король».

2.8 Искусственный интеллект оказывает преобразующее воздействие на права человека, включая права детей

ИИ оказывает преобразующее воздействие на права человека, включая права детей, внося изменения на системном уровне, с которыми связаны как новые широкие возможности, так и имеющие межсекторальный характер вызовы на всех этапах жизненного цикла ИИ:

a) **Право на неприкосновенность частной жизни.** Интеграция ИИ в инфраструктуру наблюдения расширила возможности слежения за людьми в масштабах всего населения и осуществления контроля над обществом. Повсеместные сбор, обработка, использование и повторное использование данных, стимулом для которых являются потребности ИИ на всех этапах его жизненного цикла, являются серьезнейшим вызовом для соблюдения права на неприкосновенность частной жизни [148, 149];

b) **Право на недискриминацию.** Использование имеющего предвзятости ИИ может вести к дискриминации, которая в большинстве юрисдикций является незаконной. Эти виды вреда часто затрагивают маргинализированные группы населения или людей, находящихся в уязвимом положе-

нии [150], таких как дети, женщины [151, 152] и представители расовых меньшинств [153], и оказывают несоразмерное воздействие на сообщества в странах глобального большинства.

Одной из подгрупп прав человека, вызывающих особую озабоченность в контексте использования ИИ, являются права детей [154]. При наличии соответствующих условий ИИ может оказывать позитивное воздействие на такие права детей, как право на доступ к информации, право на образование и право на свободу выражения мнения. Однако использование ИИ также связано с множеством видов риска нанесения вреда:

a) **Право на защиту от сексуальной эксплуатации и сексуальных надругательств.** По оценкам, изображения около 1,2 миллиона детей из 11 стран глобального Юга (население менее 1 миллиарда человек) были подвергнуты манипуляциям с целью создания сексуализированных дипфейков, например с помощью приложений, причем рост этих цифр вызывает тревогу [155]. Было документально зафиксировано случайное попадание материалов, содержащих сцены сексуального надругательства над детьми, в некоторые наборы обучающих данных [156], а открытые модели могут быть дообучены злоумышленниками на таких материалах. Количество сгенерированных с помощью ИИ материалов, содержащих сцены сексуального надругательства над детьми, стремительно растет: так, в 2025 году Фонд по надзору за Интернетом проанализировал более 8000 сгенерированных с помощью ИИ изображений и видеороликов, содержащих сцены надругательства [157];

b) **Право на развитие, здоровье и благополучие.** Игрушки с ИИ с функциями социального взаимодействия начинают вызывать все большее беспокойство по причине рисков для эмоционального развития детей, для неприкосновенности их частной жизни и для их благополучия, а также риска того, что дети окажутся подвержены ненадлежащему или манипулятивному взаимодействию [158–161].

Эти проблемы заслуживают эффективных решений. К перспективным подходам относятся:

a) **Транспарентность и подотчетность.** Многие системы ИИ, которые активно используются для принятия решений, затрагивающих интересы отдельных людей и сообществ, не обладают достаточной прозрачностью и объяснимостью. Это создает проблемы с точки зрения правовой ответственности разработчиков моделей и организаций, внедряющих ИИ, а также затрудняет доступ к правосудию, соблюдение принципа верховенства права и использование эффективных средств правовой защиты в случаях нарушения прав человека [162, 163];

b) **Систематическое применение правозащитных механизмов на всех этапах жизненного цикла систем ИИ.** Соблюдение должной осмотрительности в вопросах прав человека, проведение оценок воздействия и применение подходов, основанных на принципе «права человека по умолчанию» являются зарекомендовавшими себя инструментами, которые позволяют как выявлять, так и снижать риски, связанные с ИИ, а также дают возможность получать пользу от использования ИИ [164]. Анализ более 700 решений, принятых европейскими органами по защите данных, показывает, что они действительно руководствуются соображениями защиты прав человека; это, в свою очередь, ложится в основу практической методологии для проведения оценки воздействия на права человека [165]. Аналогичные аргументы можно привести в пользу проведения оценок воздействия на права детей, особенно с учетом того, что во многих системах управления ИИ вопросы, касающиеся детей, не учитываются явным образом [166, 167].

Действия в условиях неопределенности

Управление в условиях неопределенности является нормальным явлением, однако ИИ представляет собой особый случай: развитие его возможностей опережает регулирование, на передних рубежах развития находится малое число участников, появление агентных систем знаменует собой качественный перелом, а ошибки не всегда можно исправить. Преимущества от применения ИИ общего назначения являются реальными, однако их получение зависит от политических и институциональных решений, в то время как соответствующий вред наносится определенным группам населения и усиливается при использовании ИИ, не основанном на знаниях и не соответствующем общему контексту. Большинство необходимых инструментов уже существует; открыт лишь вопрос о том, как их применять на практике.

3. Результаты по сферам деятельности

3.1 Научные исследования и достижения в области искусственного интеллекта и пути его развития

Основной вывод

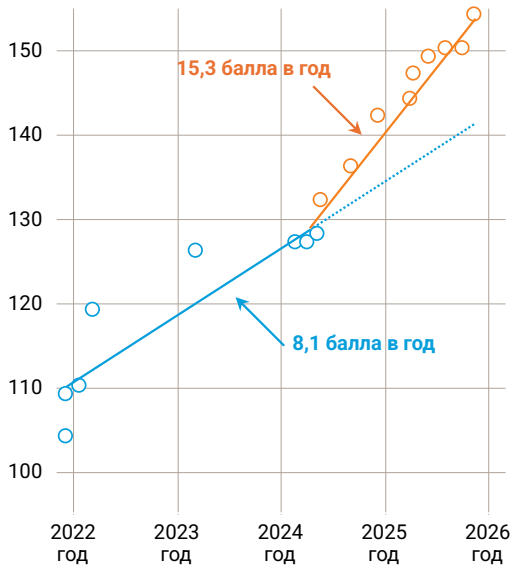
Искусственный интеллект перешел от пассивного распознавания образов к активному построению логических выводов и автономным действиям. В этой области наблюдается стремительное развитие от существующих моделей с возможностью построения логических выводов к организованным агентным сетям и, в конечном итоге, к самосовершенствующимся системам.

Методы оценки и системы управления не успевают за развитием, что создает неотложную необходимость в применении стандартов в момент, когда агентные функциональные возможности становятся нормой.

Рисунок III

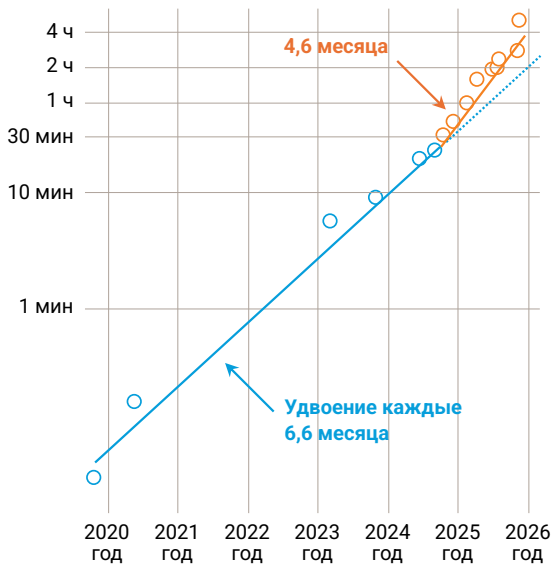
С апреля 2024 года функциональные возможности передовых систем искусственного интеллекта улучшаются почти в два раза быстрее

Баллы согласно индексу функциональных возможностей Epoch



С октября 2024 года временные горизонты METR удва-иваются почти на 50 процентов быстрее

Продолжительность временного горизонта METR



Функциональные возможности искусственного интеллекта, измеренные с помощью индекса функциональных возможностей Epoch и эталонного показателя временных горизонтов METR. Индекс функциональных возможностей Epoch объединяет около 40 эталонных тестов производительности искусственного интеллекта в единую шкалу для оценки и сравнения моделей искусственного интеллекта на протяжении времени. Эталонный показатель временных горизонтов METR используется для измерения сложности заданий в области разработки программного обеспечения и проведения исследований, выполняемых ИИ-агентом автономно, благодаря привязке производительности ко времени, которое потребовалось бы имеющему экспертные знания человеку для выполнения того же задания.

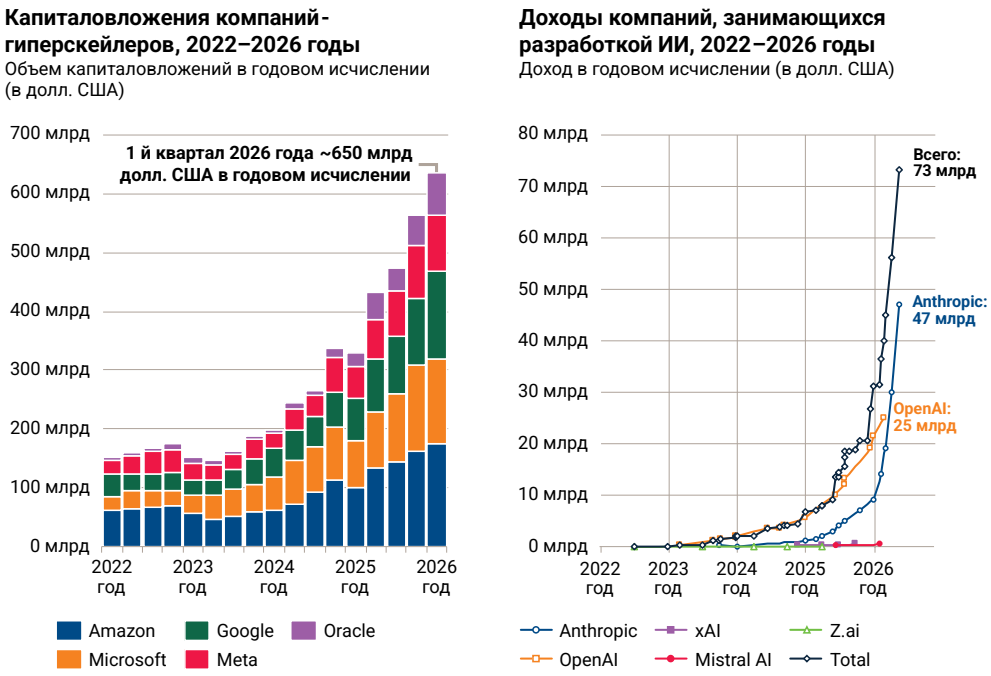
Источник: адаптировано на основе материалов веб-сайта https://epoch.ai/data-insights/ai-capabilities-progress-has-sped-up.

Ключевые пункты

Преобразующий характер искусственного интеллекта

- Совершенствование функциональных возможностей ИИ не замедляется и потенциально ускоряется [168, 169] (см. рисунок III), при этом инвестиции в вычислительные мощности достигают уровней, которые ранее были отмечены лишь при реализации промышленных проектов национального масштаба, а доходы растут быстрее, чем доходы от применения любой другой технологии [170, 171] (см. рисунок IV).
- Индустриализация труда, связанного с когнитивной деятельностью: ИИ выступает в качестве преобразующей технологии, которая расширяет процесс индустриализации со сферы физического труда на сферу выполнения когнитивных заданий.

Рисунок IV



Слева: объем капиталовложений крупнейших компаний-гиперскейлеров с 2023 года вырос примерно в пять раз, с около 150 млрд долл. США до прогнозируемого на 2026 год объема в 770 млрд долл. США. Это также примерно в три раза превышает совокупные расходы в остальном мире, что соответствует тому, что на долю Соединенных Штатов приходится почти 75 процентов мировых вычислительных мощностей, используемых на нужды искусственного интеллекта. Справа: доходы ведущих компаний, занимающихся разработкой ИИ, в годовом исчислении за тот же период выросли более чем в тридцать раз, с примерно 2 млрд долл. США в 2023 году до более чем 70 млрд долл. США (текущий прогнозный объем в годовом

исчислении) в 2026 году, что отражает высокий спрос на услуги на базе искусственного интеллекта. Эти цифры показывают, насколько быстро была создана инфраструктура на нужды искусственного интеллекта, как недавно за ней начали следовать доходы и насколько сильно оба эти элемента сосредоточены в Соединенных Штатах, что особенно явно прослеживается в случае вычислительных мощностей.

Источник: адаптировано на основе статьи «Капитальные вложения гиперскейлеров выросли в четыре раза с момента выпуска GPT-4» (“Hyperscaler capex has quadrupled since GPT-4’s release”, https://epoch.ai/data-insights/hyperscaler-capex-trend) и данных о компаниях, занимающихся разработкой ИИ (Epoch AI, 2026, https://epoch.ai/data/ai-companies).

- Обучение на опыте: современные системы ИИ объединяет способность учиться на опыте, представленном в виде предметов материальной культуры человечества (таких как тексты и изображения), взаимодействия с реальным миром и виртуальных симуляций.

Эволюция и ограничения больших языковых моделей

- Как работают большие языковые модели: большие языковые модели функционируют на основе необходимости выполнить простую цель — «предсказание следующего токена» — и обучаются генерировать текст, используя шаблоны и преобразования.
- **Разрыв в обеспечении фактичности**: хотя эти выученные преобразования позволяют успешно сохранить естественность и правдоподобность при генерировании текста, они не гарантируют фактической точности [8].
- **Изменения в обучении**: поскольку наличие высококачественных, размеченных человеком данных становится фактором, тормозящим процесс обучения, разработчики переходят к многоэтапным конвейерам обучения с использованием синтетических данных и программной обратной связи. Также наблюдается заметная тенденция к использованию вычислительных мощностей на этапе инференса, что приводит к появлению моделей со способностью строить логические выводы.

Переход к «моделям мира» и агентному искусственному интеллекту

- **Модели мира**: в этой области наблюдается отход от пассивного прогнозирования к активному приобретению знаний и построению выводов на основе причинно-следственных связей [172]. Модели мира обучаются посредством взаимодействия, наблюдения и обновления параметров, что позволяет им создавать внутренние симуляции возможных вариантов развития событий в будущем.
- **Агентный ИИ**: эти функциональные возможности позволяют создавать автономных агентов, которые могут принимать решения и действовать в самых разных контекстах, устраняя тем самым разрыв между цифровыми моделями и действиями в реальном мире (например, в робототехнике).
- **Последствия**: развитие агентного ИИ приводит к появлению новой цифровой рабочей силы, но при этом вызывает серьезную обеспокоенность, в том числе относительно уязвимостей в области безопасности. Эффективное управление и строгие стандарты считаются принципиально важными факторами поддержки этого процесса.

Трудности, связанные с оценкой, интерпретируемостью и надзором

- **Недостатки при проведении оценок**: существующие методы оценки характеризуются такими проблемами, как насыщение эталонных тестов, предвзятость, галлюцинации и способность моделей ИИ научиться распознавать, что они проходят тестирование.

- **Рабочие процессы с участием человека и ИИ**: существует острая необходимость обеспечить возможность проведения тщательного аудита и прозрачного отслеживания данных, которое позволяет проследить связь каждого сгенерированного утверждения с достоверными фактическими данными. Кроме того, системы обязаны иметь четко сформулированные критерии, чтобы точно определять, в каких случаях ИИ-агент обязан перейти под надзор человека.

Возможные будущие пути развития искусственного интеллекта

- **Начало эры агентных и гибридных систем**: отмечено переходом от пассивных помощников к инициативно действующим ИИ-агентам и к системам, в которых статистическое обучение сочетается с явно заданными знаниями и моделями мира. Стабильный прогресс ограничивается «двойным дефицитом» энергии и высококачественных данных, при этом пути развития определяются совместной оптимизацией алгоритмов, программного обеспечения и аппаратного обеспечения [173].
- **В краткосрочной перспективе**: более большие базовые модели, повсеместное внедрение моделей с возможностью построения логических выводов и первые агентные системы, берущие на себя выполнение заданий в реальном мире.
- **В среднесрочной перспективе**: развитие ИИ в направлении моделей мира, способных к построению выводов на основе причинно-следственных связей, организованных сетей ИИ-агентов, интеграции ИИ и робототехники, а также в направлении усовершенствованных инструментов интерпретируемости и устоявшихся систем управления.
- **В долгосрочной перспективе**: самоорганизующиеся и самосовершенствующиеся агенты, экспоненциальное ускорение развития технологий, глубокая интеграция ИИ в качестве экономического субъекта, а также слияние с другими передовыми технологиями, включая квантовые вычисления и биотехнологии.

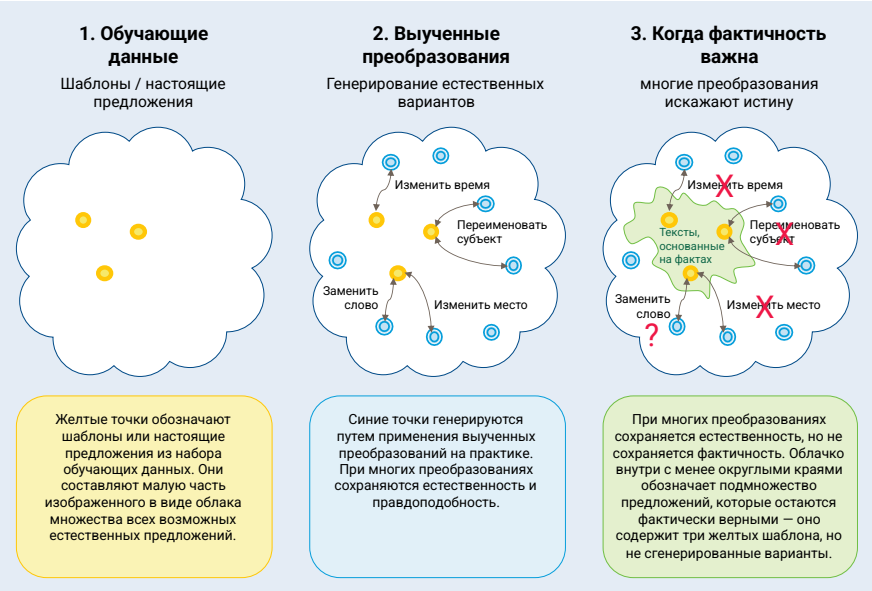
Языковые модели и различие между фактичностью и естественностью речи

Нынешние языковые модели построены на удивительно простом принципе обучения: на основе обширного статистического корпуса текстов, кода, изображений или других данных модель учится предсказывать следующую единицу или восстанавливать отсутствующие фрагменты. В случае больших языковых моделей этот процесс часто описывают как предсказание следующего слова, хотя на практике единицы («токены») по размеру обычно меньше слов. Постановка этой цели при обучении оказывается весьма эффективной, поскольку в языке есть множество закономерностей на целом ряде уровней, а именно в грамматике, стиле, закономерностях построения выводов, а также в том, как проявляются цели и намерения людей.

В простом виде обучение этих моделей можно описать как обучение не только использованию сохраненных шаблонов, но и обучение преобразованиям. Предложение, представленное в одной форме, можно преобразовать во множество связанных между собой форм, сохранив при этом естественный характер речи, например, путем изменения субъекта действия, объекта действия или глагола, изменения степени детализации, перефразирования, перевода или адаптации стиля. Система, обученная этим преобразованиям, может генерировать результаты, которые являются по-настоящему новыми, а не просто скопированными из набора обучающих данных. Такое восприятие помогает объяснить возникающую асимметрию, состоящую в том, что создать естественно звучащий текст легче, чем создать текст, основанный на

фактах: многие преобразования позволяют сохранить естественность и правдоподобность, но гораздо меньше преобразований позволяют сохранить фактичность. Например, утверждение, касающееся одного человека, может стать ложным, если просто заменить в нем имя на имя другого человека, сохранив при этом естественный характер этого утверждения. Это различие принципиально важно: пользователям не следует рассматривать уверенное владение языком как доказательство фактологической надежности.

Рисунок V



Источник: адаптировано на основе L. Bottou and B. Schölkopf, “The Fiction Machine”, SIAM News, 58(3). 2025. Воспроизводится с разрешения.

3.2 Применение на благо общества: наука, здравоохранение, образование и сельское хозяйство

Основной вывод

Использование специально разработанных систем ИИ, ориентированных на выполнение конкретных заданий, ведет к измеримому и подкрепленному фактическими данными росту в науке, здравоохранении, образовании и сельском хозяйстве. Этот рост действительно имеет место, однако он зависит от ряда условий, среди которых адаптация к местным условиям, наличие достаточной инфраструктуры и готовность людей. Сам по себе доступ не равнозначен получению благ.

Ключевые пункты

- **ИИ способен коренным образом изменить целый ряд отраслей.** Например, ИИ, ориентированный на выполнение конкретных заданий, позволяет осуществлять раннее выявление заболеваний [174] и раннее оповещение о рисках в сельском хозяйстве [175] и обеспечивать индивидуально подобранное обучение [176] в условиях ограниченных ресурсов [177].
- **Наблюдается измеримый прирост эффективности за счет использования ИИ на всех этапах процесса совершения научных открытий**, а лаборатории с системой автоматического управления продемон-

стрировали более чем десятикратное увеличение скорости обработки данных в области открытия новых материалов [178].

- **Системы ИИ, ориентированные на выполнение конкретных заданий, легче контролировать в сферах, сопряженных с принятием важных решений, чем системы ИИ общего назначения [179].** В сфере здравоохранения ИИ, ориентированный на выполнение конкретных диагностических заданий, соответствует существующей нормативной базе [180, 181]. ИИ общего назначения подходит для использования в целях снижения административной нагрузки [182]. Необходимо установить ограничения, чтобы предотвратить непреднамеренное использование ИИ общего назначения в клинической практике, учитывая, что каждый четвертый разговор с чат-ботом уже касается тем здоровья или здорового образа жизни [183].

- **Эффективные программы должны реализовываться с учетом местного контекста на всех этапах, от разработки до внедрения и оценки.** Например, ИИ-помощник в сфере здравоохранения, внедренный в национальное приложение для цифрового здравоохранения, показал точность диагностики на уровне 93 процентов при первичной сортировке пациентов, превзойдя аналогичную зарубежную систему, точность которой составила 85 процентов [184]. Однако даже инструменты, прошедшие проверку на применение в клинической практике, могут давать сбои, если не учитываются социально-экономические факторы и возможности местной инфраструктуры. Медицинские работники в Руанде положительно отреагировали на внедрение в масштабах всего сообщества приложения на базе ИИ, предназначенного для оказания клинической поддержки с переводом на язык киньяруанда и с него [185], тогда как по итогам последующих исследований в Кении было продемонстрировано улучшение результатов лечения благодаря аналогичному средству клинической поддержки с использованием ИИ на английском языке [186].

- **Готовность учителей к использованию ИИ является важным фактором, влияющим на результаты обучения.** Преподаватели, готовые к использованию ИИ, демонстрируют более высокую способность к адаптации и применяют более эффективные подходы к обучению [187, 188].

- **Пробелы в цифровой инфраструктуре и неравный потенциал в области ИИ ставят под угрозу достижение равного эффекта в сфере образования.** Хотя в богатых странах уровень цифровой подключенности высок, 25 процентов населения мира по-прежнему не имеет доступа к Интернету [189, 190].

- **Разрыв между ожиданиями учащихся и внедрением цифровых технологий чреват появлением негативных результатов обучения.** 74 процента опрошенных учащихся европейских средних школ ожидают, что ИИ будет важен в профессиональной деятельности, но лишь 44 процента считают, что их учителя готовы к его использованию [191, 192]. Только половина опрошенных школ регулирует использование ИИ (38 процентов устанавливают правила, 16 процентов запрещают его), хотя учащиеся уже используют ИИ для сбора информации (56 процентов) и генерирования готовых решений (31 процент) [192]. Более того, когда действительность при обучении тому или иному предмету не соответствует ожиданиям учащихся, у 48 процентов из них в течение 2–3 недель наблюдается значительное снижение интереса.

- **Применение ИИ в сельском хозяйстве означает появление трех функциональных возможностей с преобразующим потенциалом:** возможности прогнозирования рисков, возможности интеграции разнообразных данных (таких как данные о погодных условиях, состоянии

почвы, стадии развития сельскохозяйственных культур и рыночных ценах) в единые системы принятия решений и возможности поддержки мер реагирования, подобранных к конкретным культурам, местностям и сезонам. Платформы мониторинга на базе ИИ уже отслеживают ситуацию с продовольственной безопасностью в более чем 90 странах с использованием показателей о климате, конфликтах и состоянии экономики [193, 194].

- **Системы ИИ в сельском хозяйстве являются наиболее устойчивыми, когда их воспринимают как общую общественную инфраструктуру** с четкой системой управления, доступную для всех учреждений и предназначенную для укрепления государственно-частных партнерств. При разработке технологических решений необходимо учитывать социально-экономические реалии фермеров, в особенности мелких фермеров, которые составляют 84 процента фермерских домохозяйств, управляют 24 процентами пахотных земель и производят 30 процентов мировых продовольственных ресурсов.

Тщательно продуманная система наставничества на базе ИИ для устойчивого усвоения знаний: предотвращение «иллюзии компетентности»

В рамках рандомизированного контролируемого полевого эксперимента, проведенного в 2025 году с участием почти тысячи учащихся средних школ в Турции, были изучены эффекты использования генеративного ИИ для изучения математики [195]. По сравнению с учащимися, не пользовавшимися помощью ИИ, учащиеся, которые использовали стандартный диалоговый ИИ-интерфейс, улучшили свои краткосрочные результаты в выполнении практических заданий на 48 процентов, а те, кто пользовался имевшей защитные меры системой для наставничества, основанной на пошаговых подсказках и поэтапном построении логических выводов, — на 127 процентов. Однако при последующей проверке знаний учащиеся, которые полагались на систему без ограничений, показали более низкие результаты, что говорит о более слабом усвоении навыков в долгосрочной перспективе и об «иллюзии компетентности», при которой результаты выполнения заданий улучшились без устойчивого усвоения материала. По сравнению с этим, имевшая защитные меры система для наставничества позволила снизить негативные последствия за счет использования пошаговых подсказок и поэтапного построения логических выводов, которые имитировали эффективные методы обучения. Это исследование подчеркивает, что результаты обучения зависят от педагогической концепции и от управления системами ИИ, и указывает на то, что для эффективного внедрения в сфере образования необходим не только доступ к ИИ, но и применение основанных на фактических данных механизмов обучения и интеграция в рабочие процессы, ориентированные на интересы людей [196, 197].

Упреждающие действия для обеспечения продовольственной безопасности

Сегодня, когда изменчивость климата, конфликты и нарушения функционирования рынков оказывают все большее давление на глобальные продовольственные системы [198], ИИ позволяет создать новое поколение систем упреждающего обеспечения продовольственной безопасности, напрямую связывая ранние сигналы о напряженной ситуации в сельском хозяйстве — такие как погодные данные, спутниковые изображения и информацию о состоянии посевов — с рисками для продовольственной безопасности и мерами реагирования [199, 200]. Без необходимости ждать, пока неурожай или гуманитарные кризисы проявятся в полной мере, системы, нацеленные на принятие упреждающих действий, позволяют использовать помощь наличными, выделяемую на основе прогнозов, и системы раннего оповещения для поддержки мер вмешательства на более раннем этапе, таких как планирование на случай засухи, денежные переводы, продовольственная помощь и стабилизация рынков, до того как уязвимые домохозяйства применят все доступные им стратегии преодоления трудностей [201]. Фактические данные, полученные в ходе внедрения в 12 странах, говорят о том, что меры реагирования, принимаемые на основе сигналов систем раннего оповещения, могут помочь улучшить разнообразие рациона и увеличить частоту приемов пищи в домохозяйствах и обеспечить им стабильный доступ к продовольствию, одновременно сокращая масштабы применения ими непродуктивных стратегий преодоления трудностей, таких как пропуск приемов пищи и вынужденная продажа активов. Кроме того, автоматизированные каналы мониторинга позволяют сократить сроки направления оповещений с нескольких недель или месяцев до нескольких часов, и это устойчивое воздействие на операционную деятельность зависит от технических решений в области принятия упреждающих действий, внедренных в национальных учреждениях [203, 204].

3.3 Экономические последствия

Основной вывод

ИИ — это технология общего назначения, обладающая огромным положительным потенциалом [205, 206], однако прирост не является автоматическим. Для получения выгоды от повышения производительности необходимы дополняющие инвестиции в улучшение навыков и умений, сбор данных и изменение организационной структуры. Основной нерешенный вопрос касается распределения: кому достается избыточный доход и что происходит с рабочей силой, с развивающимися странами и с нормативно-правовой базой, применяющейся в различных отраслях и разработанной для другой эпохи [207, 208].

Ключевые пункты

- **Чтобы достичь прироста от использования искусственного интеллекта, необходимы дополняющие инвестиции в данные, рабочие процессы, улучшение навыков и умений и изменение организационной структуры.** J-образная кривая производительности объясняет, почему стремительный технический прогресс и низкая [209] совокупная производительность могут сосуществовать: прежде чем объем производства начнет расти, предприятиям необходимо накопить нематериальные дополняющие факторы производства. Фактические данные, относящиеся к уровню выполнения заданий, свидетельствуют о положительных результатах при выполнении четко определенных заданий, однако прирост на микроуровне не всегда дает совокупные результаты на макроуровне.
- **Доказательная база характеризуется дисбалансом в пользу стран с развитой экономикой, крупных компаний и сектора формальной занятости.** Превалируют фактические данные о Соединенных Штатах и Европе, об использовании на английском языке и о поддающихся количественной оценке цифровых заданиях. Анализ, проведенный Международным валютным фондом, показал, что в странах с формирующейся рыночной экономикой и развивающихся странах ниже риск потери рабочих мест в связи с внедрением ИИ и цифровая готовность [75]. Фактические данные Международной организации труда и Всемирного банка по Латинской Америке говорят о том, что риску потери рабочих мест в связи с внедрением генеративного ИИ подвержены в первую очередь образованные работники формального сектора в городских районах [210]. Политика, разработанная на основе этих фактических данных, может оказаться неприменимой для регионов, в которых проживают две трети работников мира.
- **Последствия для рынка труда лучше всего рассматривать в аспекте выполнения заданий, создания новых видов работы и качества рабочих мест, а не в аспекте простого вытеснения работников.** В прошлом создание новых видов работы преобладало: в 2018 году в Соединенных Штатах примерно 60 процентов работников выполняли виды работы, которых не существовало в 1940 году [211].
- **К встречающимся в заголовках прессы цифрам на тему внедрения ИИ следует относиться с осторожностью из-за «ИИ-вошинга» — ложных, вводящих в заблуждение или преувеличенных заявлений о возможностях ИИ.** Это особенно актуально в контексте нынешней волны увольнений, причиной которых, согласно публичным заявлениям, является ИИ [212, 213]. Его внедрение стреми-

тельно набирает обороты во всем мире* [214]. Тем не менее ранние фактические данные о воздействии на производительность являются неоднозначными и зависят от институциональных условий: в Соединенных Штатах среди работников в возрасте от 22 до 25 лет, занятых в профессиях, которые подвержены воздействию ИИ, отмечается относительное сокращение занятости [54], в то время как данные исследований из Дании говорят о практически нулевом макроэкономическом воздействии на часы работы, заработную плату и набор персонала [55]. Ключевой вопрос, касающийся качественных рабочих мест, заключается в том, повышает ли использование ИИ производительность или приводит к снижению необходимости в квалифицированных кадрах и сдвигу экономической ренты к владельцам капитала [215].

- **Макроэкономические прогнозы различаются на порядок.** По консервативным оценкам, вклад ИИ в общую факторную производительность на протяжении 10 лет составит менее 1 процента [216]. Согласно оценкам по промежуточному сценарию, в той же временной перспективе валовой внутренний продукт будет выше на 5–7 процентов**. В одном из сценариев, предполагающем полную автоматизацию, объем производства увеличивается приблизительно в десять раз, тогда как реальная заработная плата резко снижается, а доля труда в национальном доходе падает с около 60 процентов до нуля в момент, когда доля заданий, подверженных автоматизации, переходит порог в примерно 80 процентов заданий [217]. На фоне этого диапазона, согласно результатам недавнего масштабного проекта по сбору информации, проведенного экономистами, исследователями в области ИИ, специалистами по разработке политики и суперпрогнозистами, для Соединенных Штатов медианное значение вклада ИИ в общую факторную производительность к 2030 году было установлено на уровне примерно 1,2 процента (в годовом исчислении), а в сценарии, предполагающем стремительное распространение ИИ, оно возросло до 1,9–2,0 процента [218]***. Что более важно, достигнутые на практике результаты зависят как от границы возможностей, так и от скорости внедрения — сдерживающие факторы в сфере производства и инноваций могут тормозить развитие даже систем с очень продвинутыми возможностями [219], а допущения в параметрах, касающиеся замещения и масштабов, могут заставлять результаты прогнозов существенно колебаться в ту или иную сторону [220], поэтому отмечаемые в текущий момент слабые эффекты, которые соответствуют описанной выше ситуации с J-образной кривой, не исключают значительных эффектов в будущем.

---

* По сообщениям компании OpenAI, число еженедельных активных пользователей одного только ChatGPT приближается к одному миллиарду; другие крупные поставщики услуг — в том числе Google (Gemini), Microsoft (Copilot), Anthropic (Claude), Meta и платформы в Китае — также сообщают о контингентах в сотни миллионов пользователей, однако ни один из поставщиков не публикует сопоставимых суммарных межплатформных данных.

** Компания «Голдман Сакс» (2023 год). См. примечание 1. В докладе Бриггса — Коднани прогнозируется рост мирового ВВП на 7 процентов (около 7 трлн долл. США) и увеличение годового темпа роста производительности труда на 1,5 процентных пункта в течение десятилетия.

*** Медиана прогнозов экономистов относительно общего показателя: рост общей факторной производительности на 1,2 процента (в годовом исчислении, в течение 5 лет) безусловно на 2030 год, с повышением до 1,9–2,0 процента в сценарии, предполагающем стремительное распространение ИИ. Одним из ключевых методологических выводов по итогам разложения дисперсии является то, что подавляющее большинство разногласий между экспертами относится к вопросам в рамках отдельных сценариев, а не к самим сценариям: предметом дискуссий является то, как рынки труда будут абсорбировать ИИ, а не то, будет ли ИИ развиваться далее.

- **Распределение — это ключевой нерешенный вопрос.** ИИ может сократить разрывы в навыках при выполнении заданий и одновременно усугубить разрывы между компаниями, регионами, странами и между капиталом и трудом. Рынки базовых моделей имеют тенденцию к олигополизации: производство чипов, вычислительные мощности, облачная инфраструктура и обучение передовых моделей характеризуются высокой степенью концентрации, что создает условия для извлечения экономической ренты, которую вынуждены оплачивать даже компании, не связанные с разработкой ИИ [221]. В отличие от привычной ситуации, подверженные риску профессии сосредоточены в сегментах рынка труда с более высоким уровнем квалификации и оплаты труда, что может изменить курс политики автоматизации на противоположный [222].
- **Экономические и социальные последствия применения ИИ и далее будут с трудом поддаваться измерению**, если национальные статистические системы не будут обновлены, а разработчики ИИ не предоставят их представителям дополнительный доступ к данным для проведения измерений. Следует поставить целью измерение ИИ с соблюдением конфиденциальности данных и с учетом издержек. Это позволило бы странам отслеживать воздействие ИИ на производительность, занятость, заработную плату, торговлю, результативность деятельности предприятий и распределение выгод и потерь.
- **Некоторые механизмы, которые могут заслуживать более глубокого изучения.** Доступ и внедрение: когда доступность превращается в эффективную возможность использования? Агрегирование производительности: когда прирост на микроуровне превращается в прирост на макроуровне? Труд и качественные рабочие места: когда благодаря ИИ создаются, преобразуются или ухудшаются рабочие места? Сосредоточение и бюджетный потенциал: кто владеет комплексом технологий, определяет условия доступа, присваивает избыточный доход, и что произойдет с налоговыми системами, основанными на налогообложении трудовых доходов, если прибыль сместится к владельцам капитала? Как адаптируются механизмы установления ответственности, соблюдения авторского права, регулирования конкуренции и обеспечения безопасности, которые не были разработаны в расчете на еженедельно обновляющиеся модели с трудно поддающимися доскональной проверке возможностями?

3.4 Безопасность, системы и последствия для окружающей среды

Основной вывод

ИИ может создавать возможность для осуществления вредоносных операций, становиться объектом атак и усугублять существующие угрозы. Появление агентных систем приводит к значительному расширению спектра возможных атак на критически важную инфраструктуру, включая сами системы ИИ. Озабоченность относительно согласования поведения возникает в тех случаях, когда поведение ИИ расходится с целями и ценностями человека [21], а к числу наиболее очевидных рисков относятся предвзятость, обман по инициативе ИИ, угодливость и потеря контроля. Темпы развития ИИ уже превышают темпы роста потенциала в областях снижения рисков и управления. Благодаря незамедлительному и скоординированному принятию международных мер по внедрению общих стандартов можно было бы снизить риски, предотвратив «гонку уступок», возникающую в результате чистой конкуренции между корпорациями и странами.

Ключевые пункты

- **Возникающие возможности в области запуска кибератак и использования уязвимых мест создают угрозу для критически важной инфраструктуры и гражданских систем.** Риски в области безопасности существуют на каждом этапе жизненного цикла ИИ, от этапа обучения, на котором возможно отравление данных, до этапа внедрения, где возможен перехват управления ИИ с помощью внешних входных данных. Документально подтвержденные показатели успешности атак на широко используемые агенты для автоматизации программирования достигают 84 процентов [223].
- **Нарушения согласования поведения и слабые места в области безопасности воздействуют друг на друга и усугубляют друг друга.** К числу наиболее очевидных рисков в области согласования поведения относятся предвзятость, обман по инициативе ИИ, угодливость и потеря контроля над высокофункциональными системами ИИ.
- **Синтетические медиа** постепенно подрывают способность общественности и учреждений отличать подлинный контент от сгенерированного [119].
- **Воздействие на окружающую среду значительно усиливается и носит разноплановый характер.** Законы масштабирования при обучении моделей приводят к повышению спроса на вычислительные мощности; рабочая нагрузка на этапе инференса стремительно растет; в течение жизненного цикла аппаратного обеспечения образуются электротехнические и электронные отходы [232, 235, 247, 248]. Среди последствий — рост потребления энергии и воды [235, 247], выбросы парниковых газов, нагрузка, связанная с доступностью критически важных минералов и с образующимися на последующих этапах электротехническими и электронными отходами [232, 258]; все это приводит к несоразмерным экологическим и социально-экономическим последствиям в странах глобального Юга [224]. Геополитические аспекты функционирования цепочек поставок критически важных минералов изучены недостаточно, а потенциальный обратный эффект может нивелировать прирост эффективности.
- **Страны глобального Юга подвергаются непропорционально высокому риску по причине структурной уязвимости.** Зависимость от иностранного программного обеспечения, ограниченный местный потенциал устойчивости к неблагоприятным воздействиям и смягчения негативных последствий, а также пробелы в данных, из-за которых снижается эффективность работы систем в местных условиях, — все это усугубляет существующее неравенство [224–226].
- **Проверка систем ИИ — процесс оценки систем ИИ на предмет функционирования в соответствии с замыслом — остается одной из нерешенных задач [227], а использование международных механизмов координации позволило бы снизить глобальные риски [228].** Обеспечение того, что ИИ-агенты ведут себя как запланировано, не обманывают проводящих оценку и поддаются контролю по мере расширения их агентных функциональных возможностей, является нерешенной проблемой.

Опасные функциональные кибервозможности передового искусственного интеллекта

Вот уже несколько лет исследователи отслеживают стремительное развитие функциональных кибервозможностей передовых моделей ИИ, кульминацией которого стала модель Mythos, недавно разработанная компанией Anthropic. Способность модели ИИ выявлять уязвимые места в программном обеспечении может использоваться как авторами атак, так и специалистами по защите. В апреле 2026 года в рамках скоординированных усилий разработчиков передового ИИ и крупных технологических организаций и финансовых учреждений был дан старт инициативам по внедрению моделей ИИ следующего поколения в целях оборонительной безопасности [17] с особым вниманием к выявлению уязвимых мест в широко используемом программном обеспечении, обнаружение которых посторонними лицами могло бы поставить под угрозу критически важную инфраструктуру. Всего за несколько недель тестирования предварительные версии передовых моделей, действуя автономно, выявили значительное число ранее неизвестных уязвимых мест в программном обеспечении основных операционных систем и веб-браузеров, в том числе несколько уязвимых мест, не обнаруженных за десятилетия проводившихся людьми проверок.

- **Недочеты в устаревших операционных системах**: одна из моделей выявила просуществовавшую 27 лет уязвимость в OpenBSD (специализированная операционная система, которую многие считают одной из наиболее безопасных в мире), позволявшую автору удаленной атаки вывести устройство из строя, отправив ему всего два некорректно сформированных пакета.
- **Уязвимые места, связанные с обработкой медиаданных**: был обнаружен просуществовавший 16 лет дефект в FFmpeg (мультимедийный фреймворк, используемый во всем мире для обработки видео), расположенный в участке кода, который ранее был выполнен автоматизированными инструментами тестирования 5 миллионов раз и не был обнаружен.
- **Инструменты эксплуатации уязвимости на уровне ядра операционной системы**: работая с набором публично раскрытых уязвимостей в ядре Linux (базовое программное обеспечение, на котором работает большинство серверов в мире), передовая модель создала надежные инструменты эксплуатации уязвимости, с помощью которых владелец обычной учетной записи пользователя может получить полный административный контроль над системой.
- **Масштабирование безопасности браузеров**: в Mozilla Firefox интеграция передовых моделей привела к росту ежемесячного показателя выявления уязвимостей на 1000 процентов (резкий рост с базового уровня 2025 года, на котором исправлялось примерно 20–30 ошибок в системе безопасности в месяц, до 423 таких исправлений в апреле 2026 года) и позволила выявить давно существовавшие скрытые уязвимости, которые не были обнаружены за годы фаззинг-тестирования и ручной проверки [72].
- **Улучшение результатов при выполнении эталонных тестов**: при прохождении CyberGym (стандартный академический эталонный тест, в котором от ИИ-агента требуется воспроизвести известную уязвимость в реальной кодовой базе при выполнении 1507 заданий) предварительные версии передовых моделей 2026 года достигли показателя успешности в 83,1 процента, что стало значительным скачком с уровня в 66,6 процента у систем

предыдущего поколения и с уровня в 22,6 процента, достигнутого годом ранее.

Учитывая высокую значимость этих функциональных возможностей, разработчики наложили ограничения на выпуск этих специализированных моделей с оборонительными возможностями в версии общего пользования, предоставив доступ к ним только коалиции избранных глобальных организаций и основным партнерам по выпуску.

О чем это говорит

Этот сдвиг отражает главное противоречие, связанное с двойным назначением и возникающее в сфере безопасности информационно-коммуникационных технологий с появлением в ней передовых систем ИИ. Функциональная возможность, которая позволяет специалистам по безопасности выявлять и устранять существовавшие десятилетиями уязвимости, одновременно создает базовые механизмы для того, чтобы автоматически выявлять и эксплуатировать уязвимости в масштабах и со скоростью, которые превосходят потенциал традиционных человеческих коллективов.

Помимо этого, эти события подчеркивают решающую роль разработчиков технологий в принятии решений о безопасности и управлении, демонстрируют все более быстрое развитие функциональных возможностей передовых систем и позволяют выявить существующие пробелы в международных механизмах управления. Это также косвенным образом указывает на то, что функциональные возможности ИИ неравномерно распределены в глобальной экономике. В связи с этим все большее значение придается разработке совместных и инклюзивных механизмов управления высокофункциональными системами ИИ.

Значение для управления и безопасности

По мере сосредоточения передовых функциональных возможностей в высокоразвитом сегменте технологического сектора глобальный ландшафт угроз претерпевает изменения. В последних стандартах оценки подчеркивается, что риски, связанные с передовыми моделями, выходят за рамки цифровой архитектуры и затрагивают сферу физической безопасности, включая потенциальное содействие частным субъектам в распространении биологических или иных угроз [229].

Вследствие этого вопросы, касающиеся управления доступом к моделям, норм раскрытия информации об уязвимостях и справедливого внедрения инструментов ИИ, особенно в развивающихся странах, все чаще становятся вопросами международной политики, а не чисто техническими проблемами. По мере того, как функциональные возможности ИИ продолжают совершенствоваться, разрыв между готовностью к обороне и потенциальным неправомерным использованием остается одним из центральных объектов внимания как тех, кто отвечает за выработку международной политики, так и исследователей в области безопасности.

3.5 Права человека, информация и демократия

Основной вывод

ИИ оказывает преобразующее воздействие на права человека [230, 231], демократию и информационную экосистему [233] посредством изменений на системном уровне, в результате которых появляются как значительные возможности, так и структурные риски для целостности информации [234, 238] и гражданского участия [236, 237]. Неспособность устранить эти риски подрывает способность общества воспользоваться преимуществами ИИ. Уже имеются фактические данные о том, что различные учреждения все чаще используют функциональные возможности ИИ как катализатор соблюдения прав человека или как угрозу этим правам, таким как права на свободу выражения мнений и доступ к информации [238], на свободу мнений [239], на неприкосновенность частной жизни [240, 241], на недискриминацию, на доступ к правосудию, на здоровье и развитие [242].

Наиболее насущное требуемое изменение в подходе к управлению — это переход от модерации контента к системной архитектуре. Регулирование способности к убеждению и манипулированию, присущих самой системе, а не только ее выходным данным. Сосредоточение власти, эпистемологическая эрозия и фрагментация общей действительности представляют собой фундаментальные угрозы для демократического общества [243–245].

Ключевые пункты

- **Государственный контроль над средствами массовой информации может оказаться фактором, определяющим результаты работы ИИ через обучающие данные.** Проведенное в 37 странах исследование показало, что большие языковые модели более благосклонно оценивают страны с более жестким контролем над средствами массовой информации [246].
- **Использование ИИ позволило применять новые механизмы убеждения [250] и манипуляции, действующие в широких масштабах [234, 249].** Персонализированные, адаптивные системы, подстраивающиеся в режиме реального времени [250], сочетаются в них с синтетическими социальными доказательствами (использованием ИИ для создания впечатления, что продукт или марка пользуется большей популярностью, чем в действительности), которые воздействуют на когнитивные и эмоциональные уязвимые места [129, 250, 251–253].
- **Благодаря своим улучшенным возможностям по созданию убедительных текстов большие языковые модели в настоящее время используются для того, чтобы склонить мнение пользователей в ту или иную сторону.** В результате одного только постобучения больших языковых моделей убедительность ИИ повысилась на величину до 51 процента, а в результате направления запросов — еще на 27 процентов, что означает, что одну и ту же базовую модель можно сделать значительно более или менее убедительной в зависимости от того, как она настроена [254]. Этими функциональными возможностями могут пользоваться не только субъекты, располагающие значительными ресурсами; даже небольшие модели с открытым кодом можно дообучить так, чтобы они по убедительности не уступали передовым моделям, благодаря чему практически любой человек получает возможность внедрять средства влияния на основе ИИ в широких масштабах [254].
- **Эффективность убеждения остается неизменной независимо от того, верны ли исходные утверждения или ложны.** От 15 до 40 процентов утверждений, полученных от оптимизированных моделей, были оценены как с большой вероятностью относящиеся к категории ложной

информации, однако оказалось, что ложные утверждения были настолько же убедительными, что и правдивые.

- **Алгоритмы, оптимизированные под удержание внимания пользователей, систематически более активно распространяют контент, вызывающий неоднозначные [255] и резко эмоциональные реакции [256].** Фактические данные свидетельствуют о том, что большие языковые модели отражают идеологию своих создателей [257] и что у государств и влиятельных учреждений увеличилась стратегическая мотивация для использования контроля над средствами массовой информации в качестве рычага воздействия на результаты работы больших языковых моделей [246].

- **Неналожение ограничений на ИИ создает условия для i) эпистемологической эрозии, ii) получения «дивиденда лжеца» и iii) «синтетического консенсуса» [112, 113].** i) Эпистемологическая эрозия — это постепенная утрата коллективной способности отличать правду от неправды [112]; ii) «Дивиденд лжеца» — это выгода, которую недобросовестный субъект получает благодаря самому существованию дипфейков: в таком случае вещественные доказательства становятся оспариваемыми [113]; iii) «Синтетический консенсус» — это генерируемые ИИ материалы, создаваемые в больших объемах с целью имитировать широкое общественное согласие там, где на самом деле никакого согласия нет [259]; В совокупности все они постепенно разрушают общую действительность, которая требуется для существования гражданского общества, социальной сплоченности и ведения демократических обсуждений.

- **Основная проблема в области управления сместилась с контента на системы [260].** Основными факторами, приводящими к нанесению вреда, являются не отдельные результаты, а решения, принимаемые при проектировании и внедрении и лежащие в основе системных механизмов оказания влияния, включая адресацию контента, усиление и поведенческий дизайн [261, 262].

- **Оценка ОЭСР, проведенная в 23 странах, показала, что в большинстве случаев в существующей нормативной базе пока не учитываются вопросы применения механизмов убеждения** [263]. Управление, объектом которого являются исключительно результаты работы систем ИИ, будет неизменно приходить в несоответствие с системами, разработанными для масштабного создания и распространения контента, нацеленного на убеждение [236].

- **Различные виды вреда в несоразмерной степени затрагивают маргинализированные группы населения, а степень сосредоточенности власти представляет опасность [264, 265].** В около 99 процентах видеороликов с использованием дипфейков мишенями являются девочки и женщины [266, 267], в том числе журналистки. ИИ используется для пропаганды мизогинии в Интернете, что оказывает сдерживающее воздействие [268]. Кроме того, 88 процентов ведущих исследователей в области ИИ являются мужчинами [269, 270].

- **На структурном уровне доступ к вычислительным ресурсам и данным сосредоточен в небольшом числе компаний** в крайне малом числе стран [271, 272], что создает риск авторитаризма [273, 274].

- **Создаваемый благодаря использованию ИИ потенциал в области наблюдения открывает возможность индивидуализированного слежения в масштабах всего населения и усиленного контроля над обществом со стороны государства и бизнеса [275, 276].** Повсеместные сбор, обработка, использование и повторное использование данных, стимулом для которых являются потребности ИИ на всех этапах его жизнен-

ного цикла, являются серьезнейшим вызовом для соблюдения права на неприкосновенность частной жизни [277, 278].

- **Прозрачность и подотчетность являются ключевыми компонентами полноценного доступа к правосудию.** В настоящее время многие системы ИИ, используемые для принятия решений, которые затрагивают интересы отдельных людей и сообществ, характеризуются недостаточной прозрачностью и объяснимостью; это создает проблемы с точки зрения правовой ответственности разработчиков моделей и организаций, внедряющих ИИ, а также затрудняет доступ к правосудию, соблюдение принципа верховенства права и использование эффективных средств правовой защиты в случаях нарушения прав человека [279–281].

Искусственный интеллект, дипфейки и добросовестность избирательных процессов

Контекст: в 2024 году в более 70 странах, в которых проживает около половины мирового населения, проводились или были запланированы к проведению национальные выборы [282, 283]. В период с июля 2023 года по июль 2024 года исследователи выявили 82 дипфейка, имитировавших публичных деятелей, в 38 странах, включая 30 стран, в которых в течение периода, к которому относится использованный набор данных, проходили выборы либо в которых выборы были запланированы на 2024 год [284].

Что произошло: в одном из инцидентов сгенерированные ИИ голосовые клоны действующего главы государства были использованы при автоматических телефонных обзвонах с призывами к избирателям не участвовать в первичных выборах [285, 286].

В другом случае, связанном с алгоритмическим усилением на платформе, сообщалось, что контент в поддержку одного кандидата в президенты показывался тестовым аккаунтам в несколько раз чаще, чем контент в поддержку его соперника; данная платформа оспорила эти выводы. Это стало первым случаем в истории, когда президентские выборы были аннулированы из-за цифрового вмешательства в избирательный процесс[****]. Данное решение остается предметом продолжающегося правового спора, причем роль алгоритмического усиления на платформе является одним из нескольких оспариваемых факторов [287, 288]. В контексте одних более ранних парламентских выборов незадолго до голосования в социальных сетях распространялась сгенерированная ИИ аудиозапись, имитирующая одного из представителей оппозиции [282, 289]. Некоторые эксперты связывают недавние реальные примеры с вмешательством с использованием ИИ. Хотя другие оспаривают такую трактовку, и каждый случай имеет существенные нюансы, приведенные выше примеры могут указывать на устойчивую закономерность.

[****] Конституционный суд Румынии аннулировал результаты этих выборов, что стало первым таким решением в истории Европы.

О чем это говорит: эти случаи имели место на нескольких континентах и в разных политических системах. Сгенерированные ИИ материалы, скоординированная деятельность в Интернете и алгоритмические системы использовались в случаях подавления избирательной активности, имитации политических деятелей и искажения представления об общественной поддержке или были вовлечены в такие случаи. Результаты контролируемых экспериментальных исследований наглядно демонстрируют глубинный риск: диалоговый ИИ способен существенно изменять установки избирателей в лабораторных условиях, причем одна из моделей, оптимизированная под убеждение, смогла сместить отношение оппозиционно настроенных избирателей на величину до 25 процентных пунктов; смежные исследования также указывают на необходимость компромисса между убедительностью и фактической точностью [124, 129].

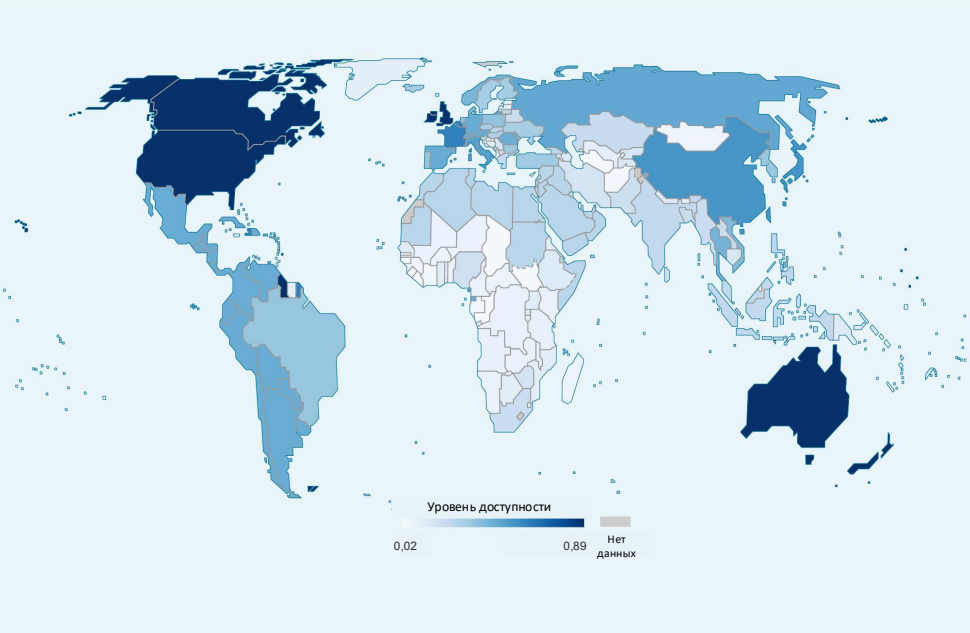
Последствия для прав человека и управления: данные тенденции затрагивают право на неприкосновенность частной жизни, право на доступ к информации, право на самостоятельное формирование мнения, право на свободу мысли и право на участие в ведении государственных дел (Международный пакт о гражданских и политических правах, статьи 17, 18, 19 и 25; Европейский суд по правам человека, статьи 8, 9 и 10; и статья 3 Протокола № 1 к Конвенции о защите прав человека и основных свобод [290, 291]. Эти случаи свидетельствуют о том, что многие системы управления по-прежнему лучше приспособлены к реагированию на причиненный вред, чем к его предотвращению.

3.6 Расцвет культуры, личностное развитие, автономия и безопасность детей

Основной вывод

Системы ИИ, оптимизированные под вовлечение пользователей и использование в широких масштабах, не являются нейтральными: в них заложены культурные представления преимущественно англоязычного мира и стран глобального Севера, результатом чего может быть активная маргинализация большей части населения мира. Дети подвергаются общим рискам, но в их усиленной версии, а количество созданных ИИ материалов, содержащих сцены сексуального надругательства над детьми, растет в геометрической прогрессии. Использование ИИ с функцией компаньонов, как и использование чат-ботов на базе ИИ для получения консультаций по вопросам психического здоровья, является повсеместным; такое использование значительно опережает сбор фактических данных, развитие механизмов обеспечения безопасности и регулирование, при этом имеются документально подтвержденные случаи причинения серьезного вреда, в том числе со смертельным исходом.

Рисунок VI



Доступность ИИ на местном языке для каждой страны, измеренная как самый высокий показатель доступности наборов данных и моделей с портала HuggingFace на языках, являющихся родными для населения данной страны. Что касается языков межэтнического общения, широко используемых за пределами стран их происхождения (английский, испанский, арабский, португальский и т. д.), то количественный показатель доступности такого языка присваивается стране только в том случае, если этот язык является родным для ≥ 50 процентов ее населения. Чем темнее оттенок, тем больше доступность. К цветовой шкале применено преобразование квадратного корня, чтобы лучше отобразить разницу в значениях в нижней части графика распределения, который имеет сильную правостороннюю асимметрию. Серый цвет означает отсутствие данных, в том числе если для страны/территории нет доступных наборов данных или моделей.

Используемые обозначения и представление материалов на этой карте не подразумевают выражения какого-либо мнения со стороны Организации Объединенных Наций относительно правового статуса какой-либо страны, города или области, или их властей, или относительно делимитации их границ или рубежей. Пунктирные линии представляют собой приблизительно линию контроля в Джамму и Кашмире согласно договоренности между Индией и Пакистаном. Окончательное согласие касательно статуса Джамму и Кашмира сторонами достигнуто не было.Окончательная граница между Республикой Судан и Республикой Южный Судан на данный момент не определена. Окончательный статус района Абьей на данный момент не определен. Между правительствами Аргентины и Соединенного Королевства Великобритании и Северной Ирландии продолжается спор относительно принадлежности Фолклендских (Мальвинских) островов.

Источник: Адаптировано на основе данных проекта «Справедливый доступ к технологиям искусственного интеллекта» (Equitable Access to Artificial Intelligence Technologies, EquATE), https://equate.vercel.app/en.

Ключевые пункты

- **Используемые в данный момент модели ИИ маргинализуют и дискриминируют многих людей и многие сообщества и культуры [39].**
- Хотя в мире говорят на более 7000 языков, используемые в данный момент модели ИИ обучаются лишь на небольшом числе из них и оптимизируются под эти языки [44, 67] (см. рисунок VI) [292].
- Даже модели, разработанные на базе десятков языков, демонстрируют хорошие результаты лишь при работе с небольшим числом из них [293, 294], а недавние исследования показывают, что расхождения в результатах между доминирующими и недостаточно представленными языками сохраняются и не сокращаются [295–297].

- В то же время более 1000 языков, уже имеющих социальную, экономическую, цифровую и информационную основу, необходимую для их полноценного включения, по-прежнему не обслуживаются.

- **Для создания более инклюзивного ИИ необходимы системные изменения на всех этапах жизненного цикла ИИ, в том числе устранение структурного дисбаланса в том, кто разрабатывает системы ИИ, определяет их характер, является их собственником и управляет ими. Для этого также требуются инвестиции в наращивание потенциала в области ИИ, в соответствующую инфраструктуру и в развитие навыков и умений во всех странах и регионах, а также наличие более репрезентативных данных и контрольных показателей.**

- Права детей на информацию, образование и выражение мнений можно укрепить через использование ИИ при применении надлежащих мер предосторожности и наличии надлежащих условий [298–301]. Однако современные системы ИИ усугубляют риски, которым подвержены дети [302, 303], в том числе растущую угрозу появления генерируемых ИИ материалов, содержащих сцены сексуального надругательства [304, 305]. Технологии создания дипфейков все шире используются для создания сексуализированных изображений детей [306, 307]. Игрушки с ИИ с функциями социального взаимодействия вызывают беспокойство относительно парасоциальных отношений и вытеснения человеческого общения, которое играет решающую роль в раннем развитии детей [308–311].

- Компаньоны с ИИ оказывают значимое положительное воздействие, однако их использование создает значительный риск зависимости, манипуляций и причинения вреда в кризисных ситуациях [312–320]. Чат-боты могут в краткосрочной перспективе уменьшать чувство одиночества на уровне, сопоставимом с эффектом от человеческого общения [321–324]. Диалоговые системы, разработанные для вовлечения пользователей во взаимодействие, могут подкреплять негативные эмоции, стимулировать чрезмерную зависимость и повышать уязвимость к манипуляциям [325, 326]. Массовый сбор чувствительных персональных данных при помощи того, что исследователи называют извлечением данных через установление близости, влечет серьезные риски для конфиденциальности [327].

- **Генеративный ИИ общего назначения широко используется для целей поддержания психического здоровья; это использование значительно опережает сбор фактических данных о безопасности и достижение консенсуса относительно его регулирования, и есть документально подтвержденные случаи причинения серьезного вреда.** В Соединенных Штатах не менее 24 процентов взрослого населения используют чат-ботов на базе ИИ для психотерапии и в качестве компаньонов [328, 329]. Угодливое поведение ИИ особенно опасно, поскольку оно может стимулировать развитие параноидального мышления и вынашивание идеи самоубийства. В исследованиях документально подтверждено, что в 9 процентах случаев взаимодействия присутствуют причиняющие вред ответы. В судебных делах утверждается, что отсутствие надлежащей реакции на суицидальные мысли привело к летальным исходам [330]. В профессиональный язык вошли новые клинические термины, в том числе «психоз, связанный с использованием ИИ» [331, 332].

- **Генеративный ИИ может помогать в преодолении кризисов в сфере психического здоровья, если его применение ограничивается теми областями, в которых его надежность доказана [333].** В Соединенных Штатах несколько цифровых помощников на базе ИИ получили

одобрение Администрации по контролю за продуктами питания и лекарствами [334], и их используют более половины американских психологов [335]. Потенциал ИИ-терапевтов в сфере психического здоровья с более полной агентностью весьма значителен, однако требует дальнейшей разработки и оценки, чтобы понять, как их можно использовать безопасно [336–338]. Разрыв между терапевтическими функциональными возможностями ИИ на английском языке и на других языках увеличивается, а критически важная культурная и контекстуальная осведомленность теряется при использовании обходных решений, основанных на переводе [339].

Искусственный интеллект оставляет большинство языков без внимания

Генеративные системы ИИ демонстрируют отличные результаты на английском языке и нескольких других широко распространенных языках. Большинство других языков либо оказываются исключены, либо демонстрируемые на них результаты оказываются гораздо хуже [340].

В результате машинного перевода на язык тигринья, на котором говорят от 7 до 9 миллионов человек в Эритрее и на севере Эфиопии, «оспа» была переведена как «сифилис», «гонорея» как «диабет», а фраза «Вам внутривенно ввели антибиотики» — как «Вам внутривенно ввели инсектициды» [341]. Такой неверный перевод может создавать угрозу для жизни.

Недавний обзор достижений в обработке естественного языка для африканских языков в сфере здравоохранения показал, что, несмотря на прогресс в разработке многоязычных инструментов на базе ИИ [342–344], серьезные проблемы сохраняются. К их числу относятся культурные и языковые предубеждения, недостаточная адаптация к различным контекстам из области здравоохранения, ограниченная объяснимость и ошибки в переводе, которые могут повлиять на постановку диагноза и принятие решений о лечении [345–350].

На основе имеющихся фактических данных можно сделать вывод о том, что системы ИИ не готовы к применению в ситуациях, сопряженных с принятием важных решений, если они не были должным образом адаптированы, ограничены и протестированы на владение соответствующим лингвистическим и культурным контекстом.

3.7 Контроль, управление и надежность

Основной вывод

Ответственные за выработку политики стороны сталкиваются с дилеммой, связанной с фактическими данными: они должны либо принимать имеющие серьезные последствия решения относительно управления ИИ без достаточного научного обоснования уже сейчас, либо ждать появления фактических данных, однако тогда для вмешательства может быть уже слишком поздно [21]. Существует более 40 типов инструментов управления, однако они являются фрагментарными, сосредоточены на уровне корпораций и редко дают возможность измерить эффективность применения в реальных условиях [351].

Ключевые пункты

- **Проведение оценок и измерений является ключевым компонентом эффективного управления ИИ, однако оно находится в критически неразвитом состоянии (см. раздел 2.2) [352]**.

- Стремительное развитие агентных систем еще больше усугубляет эти недочеты [353]. Оценочные механизмы следующего поколения должны будут быть адаптируемыми, динамичными, функционирующими на системном уровне и соответствующими контексту [354].
- Подвергаемой оценке единицей должна быть внедренная система, включающая модель, инструменты, среду и пользователей, а не только сама модель [355].
- Функциональные возможности ИИ растут быстрее, чем способность их измерять [354].
- **Потенциал является многоаспектной характеристикой и в настоящее время измеряется не полностью.** В стандартных механизмах учитываются вводимые факторы: инвестиции, программы обучения и учреждения. В них упускаются из вида два аспекта, принципиально важные для эффективного управления ИИ: 1) создание благоприятствующих ему экосистем и 2) целенаправленное развитие человеческого потенциала [291]. Потенциал должен также пониматься как результат жизненного цикла ИИ, образующийся в контексте воздействия ИИ на умения и навыки, чрезмерную зависимость и формирующиеся модели поведения, а не только как вводимые факторы для его разработки [356–358].
- Передовые системы ИИ сконцентрированы в небольшом числе компаний в небольшом числе стран, что создает проблемы с надежностью и доступностью на глобальном уровне [18, 359]. Неравный доступ к ИИ значительно увеличивает цифровой разрыв, хотя эта технология также открывает возможности для сокращения разрыва в уровне развития. Для преодоления цифрового разрыва потребуются новые механизмы, позволяющие учитывать контекст на протяжении всего жизненного цикла ИИ, от проектирования до управления.
- **Агентные системы значительно увеличивают разрыв в области измерения и управления.** Агенты действуют от имени людей, оказывая прямое влияние на реальный мир, однако методики осуществления надзора, которые были бы подобраны с учетом способности агента к самостоятельным действиям и нарождающемуся поведению, разработаны недостаточно. В существующих системах оценки риск, связанный с функционированием агентов, систематически измеряется неверно [106, 360].
- Надзор со стороны человека не введен в действие в качестве требования, выполнение которого поддается количественной оценке [361]; формирующиеся риски, связанные с взаимодействием многих агентов, невозможно выявить с помощью оценки отдельно взятого агента [362, 363], а надежные методы удержания контроля над высокоавтономными системами по-прежнему развиты недостаточно [364].
- Сбалансированный подход к регулированию ИИ должен задействовать широкий спектр инструментов, сочетая «жесткое право» (обязательное к исполнению законодательство, отраслевое регулирование, «регуляторные песочницы») с механизмами «мягкого права» (этические кодексы, добровольные обязательства разработчиков, отраслевые альянсы, технические стандарты, одобренные правительством руководящие принципы).
- **Существующие механизмы управления ИИ являются фрагментарными, сосредоточенными на уровне корпораций и недостаточными [21].** Существует более 40 типов инструментов, однако они не являются ни систематическими, ни всеохватными и редко дают возможность измерить эффективность применения в реальных условиях. В некоторых инструменты для измерения не предусмотрены; в других измеряются

только вводимые факторы [365]. Без эффективного измерения риски в сфере управления становятся чисто символическими.

- Крайне важны платформы для структурированного диалога между разработчиками передового ИИ, государствами-членами и научным сообществом. Подобные обсуждения уже проходят в рамках саммитов по ИИ (в Блетчли, Сеуле, Париже, Нью-Дели) или конференций (Всемирная конференция по искусственному интеллекту, конференция «Путешествие в мир искусственного интеллекта», Всемирный саммит «Искусственный интеллект во благо»). Наряду с ними действуют и другие площадки: органы по стандартизации (Международная организация по стандартизации/Международная электротехническая комиссия, Совместный технический комитет 1, Подкомитет 42 (искусственный интеллект) (ISO/IEC JTC 1/SC 42), Институт инженеров по электротехнике и электронике (ИИЭЭ)), партнерство ОЭСР по ИИ общего назначения, Международный альянс ИИ, Сеть институтов безопасности ИИ и возглавляемый представителями отрасли Форум по передовым моделям — однако каждая из этих инициатив носит тематический, частичный и ситуативный характер, тогда как необходимы более систематические процессы.
- Площадка на базе Организации Объединенных Наций также представляет собой один из перспективных вариантов для проведения этого диалога на постоянной и полностью инклюзивной основе в качестве дополнения к вышеупомянутым инициативам.

Искусственный интеллект с открытым кодом

ИИ с открытым кодом является одним из ключевых компонентов современного технологического ландшафта и может стать катализатором распределенных глобальных инноваций [366]. Модели ИИ с открытым кодом могут принести обществу масштабную пользу, давая разработчикам более широкие возможности воспользоваться уже сделанными вложениями в колоссальные вычислительные ресурсы и адаптировать передовые системы к местным условиям. В странах глобального Юга модели с открытыми весами позволили разработчикам оптимизировать высокофункциональные базовые модели под местные условия окружающей среды, что обеспечило возможность использования критически важных приложений на практике в сельском хозяйстве и здравоохранении, например, для прогнозирования урожайности культур [367] и для диагностической визуализации на месте и во время оказания медицинской помощи [368]. Наличие совместно используемых рабочих продуктов также позволяет снизить совокупное воздействие ИИ на окружающую среду. Более того, тот факт, что такие системы имеют прозрачную структуру, способствует укреплению доверия со стороны общественности, поскольку обеспечивает возможность внешнего надзора и проведения аудита [369].

Хронология развития

Хронология развития искусственного интеллекта с открытым кодом свидетельствует о переходе от разработки закрытых систем внутри корпораций к распределенной совместной разработке инноваций в глобальном масштабе. Особо значимым прецедентом стала разработка консорциумом BigScience модели с открытым кодом BLOOM в 2022 году [370], за которой последовал публичный выпуск семейства высокопроизводительных моделей Llama с открытыми весами [371] и выпуск серии моделей Gemma. Мощный импульс развитию придали высококонкурентная альтернативная экосистема китайских разработчиков Qwen [372, 373], а также выпуск моделей DeepSeek-V3 [374] и R1 [375]. В настоящее время целые регионы также активно вносят свой вклад в

развитие ИИ с открытыми весами в рамках таких проектов, как Mistral в Европе [376], Falcon в Объединенных Арабских Эмиратах [377], GigaChat [378] и YandexGPT [379] в Российской Федерации, а также проектов в Индии [380], Японии [381], Республике Корея [382] и других проектов.

Вызовы в области контроля и управление рисками

Модели с открытым кодом, обладающие высокими функциональными возможностями, трудно контролировать [383]. После выпуска модели в открытый домен ограничить или отозвать доступ к ней уже невозможно, что оставляет возможность для ее злонамеренного использования на практике (например, для нанесения трудно поддающегося искоренению кибервреда, такого как автоматическое генерирование сложного вредоносного программного обеспечения) [21, 384]. Требуется значительно нарастить потенциал в глобальном масштабе, чтобы обеспечить местным сообществам возможность безопасно адаптировать и оценивать системы ИИ. Исследователи также последовательно выступают за применение строгих протоколов измерения и оценки для проверки безопасности модели перед ее выпуском [385]. Такие международные организации, как Организация Объединенных Наций, могут играть ключевую роль в координации глобальных стандартов, обеспечивая равновесие между стремлением к открытым инновациям и необходимостью использования единых оценочных тестов на безопасность и надежного измерения неустранимых рисков.

4. Пробелы и дальнейшие шаги

4.1. Пробелы в фактических данных

Доказательная база относительно ряда аспектов применения ИИ отличается неравномерностью или является недостаточной. Ниже приведены примеры областей, относительно которых Группа пока не может с уверенностью сделать научные выводы.

- **Макроэкономика и производительность.** Наука пока не может с уверенностью сказать, приведет ли рост производительности на уровне выполнения ИИ отдельных заданий к совокупному росту на уровне всей экономики. В прогнозах наблюдаются значительные расхождения из-за того, что предположения о внедрении и создании новых заданий различаются. Имеющиеся на сегодняшний день фактические данные позволяют более точно оценить уменьшение издержек на выполнение уже существующих заданий, чем вклад использования ИИ в появление новых товаров, услуг и рынков.
- **Последствия для рынка труда.** Имеющиеся на сегодняшний день результаты исследований не позволяют сделать однозначный вывод о характере последствий для рынка труда. Исторические факты показывают, что экономика способна создавать новые виды работы, однако они не автоматически являются повсеместными или высококачественными, в результате чего работники, только начинающие свою трудовую деятельность, потенциально могут оказаться в уязвимом положении.
- **Злонамеренное использование химических и биологических технологий негосударственными субъектами.** Исследования показывают, что по мере развития ИИ постепенно снижается пороговый уровень специализированных знаний, необходимых для разработки и применения получаемых биотехнологическими методами агентов — однако фактические масштабы этого риска и условия, при которых он может

привести к умышленному созданию пандемий, по-прежнему плохо изучены.

- **Окружающая среда и ресурсы.** Стремительное распространение технологии ИИ стимулирует спрос на цифровую инфраструктуру, что приводит к росту потребления энергии и воды, увеличению объемов выбросов парниковых газов, нагрузке на цепочки поставок критически важных минералов и появлению электротехнических и электронных отходов [386]. Однако проведение стандартизированных измерений и предоставление соответствующей отчетности на всех этапах жизненного цикла ИИ по-прежнему не практикуется.
- **Глобальная цепочка поставок в сфере ИИ.** Она охватывает имеющие место в разных странах добычу сырья, производство чипов, сбор и аннотирование данных, обучение моделей, инфраструктуру, внедрение и утилизацию аппаратных средств. Для лучшего понимания всех последствий необходимо дальнейшее изучение этого вопроса.
- **Эффективность инструментов управления.** Группа составила перечень инструментов управления, применяемых на корпоративном, национальном и международном уровнях, однако фактические данные об их эффективности в реальных условиях остаются скудными.
- **Последствия на индивидуальном и коллективном уровнях.** Фактические данные о воздействии ИИ на расцвет культуры, личностное развитие людей и их автономию пока только появляются. Механизм связи между взаимодействиями человека с ИИ на индивидуальном уровне и проявлением на уровне общества таких явлений, как эпистемологическая эрозия, гражданская активность и социальная сплоченность, по-прежнему изучен недостаточно. Имеющиеся фактические данные отражают состояние отдельных аспектов на определенный момент — показателей вовлеченности пользователей, документально подтвержденных случаев причинения вреда, тематических исследований — но не общий курс развития событий.

4.2. Сфера охвата мандата

В настоящем докладе Группа не рассматривает вопросы применения ИИ и смертоносных автономных систем вооружений в военных целях. В резолюции 79/325 Генеральной Ассамблеи деятельность Группы прямо ограничивается невоенной сферой: «деятельность Группы и Диалога ограничиваются невоенной сферой и не касаются применения искусственного интеллекта в военных целях». По этой причине также принимается, что мандат Группы не охватывает риски, связанные со злонамеренным использованием химических и биологических технологий, в той мере, в какой они касаются военной сферы.

4.3. Дальнейшие шаги

Этот предварительный доклад знаменует собой начало, а не завершение работы Группы. Группа будет далее укреплять свою научную доказательную базу, проводя структурированные консультации с научным сообществом и поддерживая с ним тесное взаимодействие. К конкретным дальнейшим шагам относятся:

1. **Сбор тематической справочной информации.** В резолюции 79/325 Генеральной Ассамблеи прямо предусматривается публикация бюллетеней с тематической справочной информацией в дополнение к ежегодному докладу. Группа планирует публиковать бюллетени с тематической справочной информацией по насущным вопросам по мере возник-

новения таковых, включая следующие вопросы, но не ограничиваясь ими: ИИ и окружающая среда; ИИ и безопасность детей; инструменты управления ИИ и оценка их эффективности; или бюллетени с более общей отраслевой информацией по вопросам применения ИИ в космической отрасли, в квантовых технологиях, в правовых и судебных системах и на финансовых рынках.

2. **Материалы по итогам Глобального диалога.** Группа примет во внимание итоговые документы Глобального диалога по вопросам управления искусственным интеллектом и готова взять на себя новые задания, которые могут поставить ей государства-члены.

References

1. Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. Science, 358(6370), 1530–1534. https://doi.org/10.1126/science.aap8062

2. Schölkopf, B. (2022). Causality for machine learning. In H. Geffner, R. Dechter, & J. Y. Halpern (Eds.), Probabilistic and causal inference: The works of Judea Pearl (pp. 765–804). ACM.

3. Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base—Analyst note. Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

4. AI Alliance Network. (2025). AI horizons: What will AI technologies look like in 10 years? A research project. AI Alliance Network.

5. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682. https://arxiv.org/abs/2206.07682

6. METR. (2025, March 19). Measuring AI ability to complete long tasks. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

7. Crafts, N. (2021). Artificial intelligence as a general-purpose technology: An historical perspective. Oxford Review of Economic Policy, 37(3), 521–536. https://academic.oup.com/oxrep/article/37/3/521/6374675

8. Bottou, L., & Schölkopf, B. (2025). The fiction machine. SIAM News, 58(3).

9. Costa-Gomes, B., Tolmachev, P., Taysom, E., Sounderajah, V., Richardson, H., Schoenegger, P., & King, D. (2026). Public use of a generalist LLM chatbot for health queries. Nature Health, 1–8.

10. Bengio, Y., Clare, S., Prunkl, C., et al. (2026). International AI Safety Report 2026. International AI Safety Report. https://internationalaisafetyreport.org/

11. Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., ... & Chan, L. (2026). Measuring AI ability to complete long software tasks. Advances in Neural Information Processing Systems, 38, 92213-92266.

12. Anthropic. (2025). Responsible scaling policy. https://www.anthropic.com/rsp

13. OpenAI. (2025). Preparedness framework version 2. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

14. Google DeepMind. (2025). *Frontier safety framework*. https://deepmind.google/blog/updating-the-frontier-safety-framework/

15. Dragan, A., et al. (2024). Introducing the frontier safety framework. Google DeepMind. https://deepmind.google/blog/introducing-the-frontier-safety-framework/

16. Anthropic. (2026, April 7). Claude Mythos Preview (Frontier Red Team technical report). https://red.anthropic.com/2026/mythos-preview/

17. Anthropic. (2026, April 7). Project Glasswing: Securing critical software for the AI era. https://www.anthropic.com/glasswing

18. Stanford Institute for Human-Centered AI. (2026). AI Index report 2026. Stanford University. https://hai.stanford.edu/ai-index/2026-ai-index-report

19. Scale AI. (2025). Humanity's last exam. https://labs.scale.com/leaderboard/humanitys_last_exam

20. Epoch AI. (2026). AI capabilities. https://epoch.ai/benchmarks

21. Bengio, Y., Clare, S., Prunkl, C., et al. (2026). International AI Safety Report 2026. arXiv. https://doi.org/10.48550/arXiv.2602.21012

22. Xu, C., Guan, S., Greene, D., & Kechadi, M.-T. (2024). Benchmark data contamination of large language models: A survey. arXiv. https://arxiv.org/abs/2406.04244

23. Akhtar, M., Reuel, A., Soni, P., Ahuja, S., Ammanamanchi, P. S., Rawal, R., et al. (2026). When AI benchmarks plateau: A systematic study of benchmark saturation. arXiv. https://arxiv.org/abs/2602.16763

24. Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. Patterns, 5(5), Article 100988. https://doi.org/10.1016/j.patter.2024.100988

25. Needham, J., Edkins, G., Pimpale, G., Bartsch, H., & Hobbhahn, M. (2025). Large language models often know when they are being evaluated. arXiv. https://arxiv.org/abs/2505.23836

26. Van Der Weij, T., Hofstätter, F., Jaffe, O., Brown, S., & Ward, F. (2025, May). Ai sandbagging: Language models can strategically underperform on evaluations. In International Conference on Learning Representations (Vol. 2025, pp. 73152-73189).

27. Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., ... & Maharaj, T. (2023, June). Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 651-666).

28. Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., ... & Rahwan, I. (2025). Multi-agent risks from advanced ai. arXiv preprint arXiv:2502.14143.

29. Folkerts, L., Payne, W., Inman, S., Giavridis, P., Skinner, J., Deverett, S., et al. (2026). Measuring AI agents' progress on multi-step cyber attack scenarios. arXiv. https://arxiv.org/abs/2603.11214

30. Patwardhan, T., Dias, R., Proehl, E., Kim, G., Wang, M., Watkins, O., et al. (2025). GDPval: Evaluating AI model performance on real-world economically valuable tasks. arXiv. https://arxiv.org/abs/2510.04374

31. Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., et al. (2025). Chain of thought monitorability: A new and fragile opportunity for AI safety. arXiv. https://arxiv.org/abs/2507.11473

32. Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. arXiv. https://arxiv.org/abs/2304.13734

33. Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., et al. (2024). Black-box access is insufficient for rigorous AI audits. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3630106.3659037

34. Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., et al. (2024). Clio: Privacy-preserving insights into real-world AI use. arXiv. https://arxiv.org/abs/2412.13678

35. European Commission. (2025). AI Act: Draft guidance and reporting template for serious AI incidents. https://digital-strategy.ec.europa.eu/en/consultations/ai-act-commission-issues-draft-guidance-and-reporting-template-serious-ai-incidents-and-seeks

36. Organisation for Economic Co-operation and Development. (n.d.). AI Incident Monitor (AIM). https://oecd.ai/en/catalogue/tools/ai-incident-database

37. MIT FutureTech. (n.d.). AI Risk Repository / Incident Tracker. https://airisk.mit.edu

38. Organisation for Economic Co-operation and Development. (2025). Competition in artificial intelligence infrastructure (OECD Roundtables on Competition Policy Papers No. 330). OECD Publishing.

39. Sajadieh, S., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Santarlasci, L., et al. (2026). AI Index report 2026. Stanford Institute for Human-Centered AI.

40. Pilz, K. F., Sanders, J., Rahman, R., & Heim, L. Trends in AI Supercomputers. In ICML Workshop on Technical AI Governance (TAIG).

41. Kalluri, P. R., Agnew, W., Cheng, M., et al. (2025). Computer-vision research powers surveillance technology. Nature, 643, 73–79. https://doi.org/10.1038/s41586-025-08972-6

42. Office of the United Nations High Commissioner for Human Rights. (2022). The right to privacy in the digital age (A/HRC/51/17).

43. Teklehaymanot, H. K., & Nejdl, W. (2025). Tokenization disparities as infrastructure bias: How subword systems create inequities in LLM access and efficiency. arXiv preprint arXiv:2510.12389.

44. Occhini, G., Tanaka-Ishii, K., Barford, A., Tikochinski, R., Hu, S., Reichart, R., ... & Korhonen, A. (2026). Artificial intelligence is creating a new global linguistic hierarchy. arXiv. https://arxiv.org/abs/2602.12018

45. Alhanai, T., Kasumovic, A., Ghassemi, M. M., Zitzelberger, A., Lundin, J. M., & Chabot-Couture, G. (2025, April). Bridging the gap: enhancing LLM performance for low-resource African languages with new benchmarks, fine-tuning, and cultural adjustments. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 27, pp. 27802-27812).

46. Bhutani, M., Robinson, K., Prabhakaran, V., Dave, S., & Dev, S. (2024). SeeGULL multilingual: A dataset of geo-culturally situated stereotypes. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 842–854).

47. Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., & Tavares, M. M. (2024). Gen-AI: Artificial intelligence and the future of work (IMF Staff Discussion Note SDN/2024/001). International Monetary Fund.

48. McElheran, K., Yang, M.-J., Kroff, Z., & Brynjolfsson, E. (2024). The rise of industrial AI in America: Microfoundations of the productivity J-curve(s) (NBER Working Paper No. 32937). National Bureau of Economic Research.

49. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469–481). ACM. https://doi.org/10.1145/3351095.3372828

50. Skirzynski, J., Danks, D., & Ustun, B. (2025). Discrimination exposed? On the reliability of explanations for discrimination detection. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 2554–2569).

51. Zollo, T., Rajaneesh, N., Zemel, R., Gillis, T., & Black, E. (2025). Towards effective discrimination testing for generative AI. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 1028–1047).

52. Lazard, L., Capdevila, R., Turley, E. L., Gilfoyle, K., & Stavropoulou, N. (2025). Deepfake Technology and Gender-Based Violence: A Scoping Review. Trauma, Violence, & Abuse, 15248380251384271.

53. Posetti, J., Williams, K., Hellmueller, L., Renaud, P., Shabbir, N., & Aboulez, N. (2026). Tipping point: Online violence impacts, manifestations and redress in the AI age. UN Women.

54. Brynjolfsson, E., Chandar, B., & Chen, R. (2025). Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. Stanford Digital Economy Lab.

55. Humlum, A., & Vestergaard, E. (2025). Large language models, small labor market effects (NBER Working Paper No. 33777). National Bureau of Economic Research.

56. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586

57. Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. Quarterly Journal of Economics.

58. International Monetary Fund. (2024). Broadening the gains from generative AI: The role of fiscal policies (IMF Staff Discussion Note).

59. Organisation for Economic Co-operation and Development. (2026). The OECD.AI index. OECD Publishing.

60. Attard-Frost, B., & Lyons, K. (2025). AI governance systems: A multi-scale analysis framework, empirical findings, and future directions. *AI and Ethics*, *5*(3), 2557-2604.

61. UNESCO. (2023). Readiness assessment methodology. UNESCO.

62. Hawkins, Z. J., Lehdonvirta, V., & Wu, B. (2025). AI compute sovereignty: Infrastructure control across territories, cloud providers, and accelerators. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5312977

63. Markus, A., Carolus, A., & Wienrich, C. (2025). Objective measurement of AI literacy: Development and validation of the AI Competency Objective Scale (AICOS). Computers and Education: Artificial Intelligence.

64. Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion.

65. Math Matters AI. (n.d.). Math Matters AI. https://www.mathmatters.ai

66. Hinojosa, T., Rapaport, A., Jaciw, A., & Zacamy, J. (2016). Exploring the foundations of the future STEM workforce: K–12 indicators of postsecondary STEM success. Regional Educational Laboratory Southwest. https://ies.ed.gov/use-work/resource-library/report/systematic-literature-review/exploring-foundations-future-stem-workforce-k-12-indicators-postsecondary-stem-success

67. United Nations Conference on Trade and Development. (2025). *Technology and innovation report 2025: Inclusive artificial intelligence for development*. United Nations. https://doi.org/10.18356/9789211068016

68. Chauhan, P. (2025). AI and human rights: Global South perspectives. International Journal of Humanities Social Science and Management, 5(4), 563–568. https://ijhssm.org/issue_dcp/AI%20and%20Human%20Rights%20%20Global%20South%20Perspectives.pdf

69. Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: Barriers and pathways forward. International Affairs, 100(3), 1275–1286. https://doi.org/10.1093/ia/iiae073

70. Khodabin, M., & Arsalani, A. (2025). Artificial intelligence literacy as national strategy: A systematic review of policy, equity, and capacity building across the Global South. World Studies in Policy Sciences, 9, 777–814. https://doi.org/10.22059/wsps.2025.396472.1530

71. Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., et al. (2024). On the societal impact of open foundation models. arXiv. https://arxiv.org/abs/2403.07918

72. Grinstead, B., Holler, C., & Braun, F. (2026, May). Behind the scenes hardening Firefox with Claude Mythos Preview. Mozilla Hacks. https://hacks.mozilla.org/2026/05/behind-the-scenes-hardening-firefox/

73. Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-T., Jiao, W., & Lyu, M. R. (2024). All languages matter: On the multilingual safety of LLMs. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 5865–5877). https://doi.org/10.18653/v1/2024.findings-acl.349

74. Nigatu, H. H., Mehandru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). Viability of machine translation for healthcare in low-resourced languages. In Proceedings of EMNLP 2025 (pp. 10584–10598).

75. Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., Rockall, E., & Tavares, M. M. (2024). The global impact of AI: Mind the gap (IMF Working Paper No. 24/136). International Monetary Fund.

76. World Bank. (2025). Digital progress and trends report 2025: Strengthening AI foundations. World Bank.

77. McElheran, K., Yang, M.-J., Kroff, Z., & Brynjolfsson, E. (2024). The rise of industrial AI in America: Microfoundations of the productivity J-curve(s) (NBER Working Paper No. 32937).

78. Brynjolfsson, E., Rock, D., & Syverson, C. (2017). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. NBER Working Paper No. 24001

79. Calvino, F., & Fontanelli, L. (2023). A portrait of AI adopters across countries: Firm characteristics, assets' complementarities and productivity. OECD Science, Technology and Industry Working Papers, No. 2023/11.

80. McKinsey & Company. (2026, March 25). State of AI trust in 2026: Shifting to the agentic era. https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era

81. Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020, April). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-12).

82. Brant, A., Singh, P., Yin, X., et al. (2025). Performance of a deep learning diabetic retinopathy algorithm in India. *JAMA Network Open, 8*(3), e250984. https://doi.org/10.1001/jamanetworkopen.2025.0984

83. UNESCO. (2025). AI and the future of education: disruptions, dilemmas and directions

84. Cristia, J. et al. (2017). Technology and child development: evidence from the One Laptop per Child program. *American Economic Journal: Applied Economics*, 9(3), 295–320.

85. Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(1), Article 6. https://doi.org/10.3390/soc15010006

86. Ojija, F., Ogwu, M. C., Ally, J., John, J. P., Stephano, A., Felix, N., & Tekka, R. (2025). Artificial intelligence-driven solutions for mitigating human–wildlife conflict in biodiversity hotspots. Science Progress, 108(4), 00368504251394584.

87. Noy, S. & Zhang, W. (2023). "Experimental evidence on the productivity effects of generative artificial intelligence." Science, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586.

88. Cui, Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S. & Salz, T. (2026). "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers." Management Science. https://doi.org/10.1287/mnsc.2025.00535

89. Dell'Acqua, F., McFowland, E. III, Mollick, E., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. & Lakhani, K. R. (2023). "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Harvard Business School Working Paper 24-013. SSRN: https://ssrn.com/abstract=4573321

90. Ali, Z., Muhammad, A., Lee, N., Waqar, M., & Lee, S. W. (2025). Artificial Intelligence for Sustainable Agriculture: A Comprehensive Review of AI-Driven Technologies in Crop Production. Sustainability, 17(5), 2281. https://doi.org/10.3390/su17052281

91. Dhal, S. B., & Kar, D. (2024). Transforming Agricultural Productivity with AI-Driven Forecasting: Innovations in Food Security and Supply Chain Optimization. Forecasting, 6(4), 925–951. https://doi.org/10.3390/forecast6040046

92. Pearlman, K., Wan, W., Shah, S., & Laiteerapong, N. (2025). Use of an AI scribe and electronic health record efficiency. JAMA Network Open, 8(10), e2537000.

93. Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., Gravel Sullivan, A., Wills, G., ... & Gordon, J. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI, 2(12), AIoa2500945.

94. Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Wilson Hannay, S. B., ... & Lee, K. (2025). Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. NEJM Catalyst Innovations in Care Delivery, 6(5), CAT-25.

95. Paul A. David (1990, May). The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox. The American Economic ReviewVol. 80, No. 2, Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association, pp. 355-361

96. Brynjolfsson, E., & Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives, 14*(4), 23–48. https://doi.org/10.1257/jep.14.4.23

97. Shaw, S. D., & Nave, G. (2026). *Thinking—fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender*. SSRN. https://doi.org/10.2139/ssrn.6097646

98. Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review*, *37*(2), 45.

99. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., & Boulesteix, A.-L. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078451. https://doi.org/10.1136/bmj-2024-078451

100. Rotz, S., Duncan, E., Small, M., Botschner, J., Dara, R., Mosby, I., Reed, M., & Fraser, E. D. G. (2019). The politics of digital agricultural technologies: A preliminary review. *Sociologia Ruralis, 59*(2), 203–229. https://doi.org/10.1111/soru.12233

101. United Nations General Assembly. (2024). Global Digital Compact (Annex II to the Pact for the Future, A/RES/79/1). United Nations. https://docs.un.org/en/A/RES/79/1

102. UNESCO. (2022). K-12 AI curricula: A mapping of government-endorsed AI curricula. https://www.unesco.org/en/articles/k-12-ai-curricula-mapping-government-endorsed-ai-curricula

103. Almatrafi, O., Johri, A., & Lee, H. (2024, 2024/06/01/). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). Computers and Education Open, 6, 100173. https://doi.org/https://doi.org/10.1016/j.caeo.2024.100173

104. Ma, M., Ng, D. T. K., Liu, Z., & Wong, G. K. W. (2025, 2025/06/01/). Fostering responsible AI literacy: A systematic review of K-12 AI ethics education. Computers and Education: Artificial Intelligence, 8, 100422. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100422

105. Atias, O., & Mawasi, A. (2025, 2025/12/01/). Conceptualizing AI literacies for children and youth: A systematic review on the design of AI literacy educational programs. Computers and Education: Artificial Intelligence, 9, 100491. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100491

106. Kasirzadeh, A., & Gabriel, I. (2025, April). Characterizing AI Agents for Alignment and Governance. https://arxiv.org/abs/2504.21848

107. Wijk, H., Lin, T. R., Becker, J., Jawhar, S., Parikh, N., Broadley, T., ... & Barnes, E. (2025, October). RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts. In International Conference on Machine Learning (pp. 66772-66832). PMLR.

108. Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., ... & Weng, L. (2025, May). Mle-bench: Evaluating machine learning agents on machine learning engineering. In International Conference on Learning Representations (Vol. 2025, pp. 50466-50494).

109. Pichai, S. (2026, April 22). Cloud Next '26: Momentum and innovation at Google scale. Google Blog. https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/cloud-next-2026-sundar-pichai/

110. Cybersecurity and Infrastructure Security Agency. (2025, January 14). CISA, JCDC, government and industry partners publish AI cybersecurity collaboration playbook. U.S. Department of Homeland Security. https://www.cisa.gov/news-events/news/cisa-jcdc-government-and-industry-partners-publish-ai-cybersecurity-collaboration-playbook

111. Liu, Y., Zhao, Y., Lyu, Y., Zhang, T., Wang, H., & Lo, D. (2025). "Your AI, my shell": Demystifying prompt injection attacks on agentic AI coding editors. https://doi.org/10.48550/arXiv.2509.22040

112. Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820. https://doi.org/10.15779/Z38RV0D

113. Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2025). The liar's dividend: Can politicians claim misinformation to evade accountability?. *American Political Science Review*, *119*(1), 71-90.

114. Chan, Y.-T., et al. (2024). Assessing the article screening efficiency of artificial intelligence for systematic reviews. *Journal of Dentistry, 149*, 105259. https://doi.org/10.1016/j.jdent.2024.105259

115. Delgado-Licona, F., Alsaiari, A., Dickerson, H., Klem, P., Ghorai, A., Canty, R. B., Bennett, J. A., Jha, P., Mukhin, N., Li, J., López-Guajardo, E. A., Sadeghi, S., Bateni, F., & Abolhasani, M. (2025). Flow-driven data intensification to accelerate autonomous inorganic materials discovery. *Nature Chemical Engineering, 2*, 436–446. https://doi.org/10.1038/s44286-025-00249-z

116. Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., & Anderljung, M. (2025). Infrastructure for AI Agents. http://arxiv.org/abs/2501.10114

117. Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. AI Agents That Matter. Transactions on Machine Learning Research.

118. Zhu, L., Lu, Q., Ding, M. et al. Designing meaningful human oversight in AI. AI Ethics 6, 286 (2026). https://doi.org/10.1007/s43681-026-01147-7

119. Ferrara, E. (2024). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science, 7*, 549–569. https://doi.org/10.1007/s42001-024-00250-1

120. Djiré, A. E., Kaboré, A. K., Samhi, J., et al. (2026). *Learned or memorized? Quantifying memorization advantage in code LLMs*. In *Proceedings of the International Conference on Software Engineering*. https://arxiv.org/abs/2604.13997

121. Gowal, S., Bunel, R., Stimberg, F., Stutz, D., Ortiz-Jimenez, G., Kouridi, C., ... & Kohli, P. (2025). SynthID-Image: Image watermarking at internet scale. arXiv preprint arXiv:2510.09263.

122. United Nations Human Rights Council. Expert Mechanism on the Right to Development. (2024). *AI, cultural rights and the right to development* (A/HRC/EMRTD/11/CRP.2). United Nations. https://undocs.org/A/HRC/EMRTD/11/CRP.2

123. Brennan Center for Justice. (2025). *Gauging AI threat to free and fair elections*. https://www.brennancenter.org/our-work/analysis-opinion/gauging-ai-threat-free-and-fair-elections

124. Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational AI. Science, 390(6777), eaea3884. https://doi.org/10.1126/science.aea3884

125. Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, *118*(50), e2102141118.

126. Cho, J., Ahmed, S., Hilbert, M., Liu, B., & Luu, J. (2020). Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. *Journal of broadcasting & Electronic media*, *64*(2), 150-172.

127. Feezell, J. T., Wagner, J. K., & Conroy, M. (2021). Exploring the effects of algorithm-driven news sources on political behavior and polarization. *Computers in human behavior*, *116*, 106626.

128. Argyle, L. P. (2025). Political persuasion by artificial intelligence. *Science*, *390*(6777), 983-984.

129. Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., ... & Rand, D. G. (2025). Persuading voters using human–artificial intelligence dialogues. *Nature*, 1-8.

130. Myra Cheng et al. (2026). ,Sycophantic AI decreases prosocial intentions and promotes dependence.Science391,eaec8352(2026).DOI:10.1126/science.aec8352

131. Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391(6792), eaec8352. https://doi.org/10.1126/science.aec8352

132. Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., ... & Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, *13*(6), 522-530.

133. Balan, R., & Gumpel, T. P. (2025). ChatGPT Clinical Use in Mental Health Care: Scoping Review of Empirical Evidence. *JMIR Mental Health*, *12*, e81204.

134. Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". *JMIR Mental Health*, *12*(1), e85799.

135. Osler, L. (2026). Hallucinating with AI: distributed delusions and "AI psychosis". Philosophy & Technology, 39(1), 30.

136. Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., ... Kaplan, J. (2021). A general language assistant as a laboratory for alignment (arXiv:2112.00861). arXiv. https://doi.org/10.48550/arXiv.2112.00861

137. Organisation for Economic Co-operation and Development. (2024). *Facts not fakes: Tackling disinformation, strengthening information integrity*. OECD Publishing. https://doi.org/10.1787/d909ff7a-en

138. Waight, H., Yang, E., Yuan, Y., et al. (2026). State media control influences large language models. *Nature*. https://doi.org/10.1038/s41586-026-10506-7

139. Kalina Bontcheva (2024). Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities. https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf

140. Laudrain, A. (2026, April 29). When AI governs (dis)information: Five lessons for democracy. Center for Security Studies (CSS), ETH Zurich. https://css.ethz.ch/en/center/CSS-news/2026/04/when-ai-governs-disinformation-five-lessons-for-democracy.html

141. Boine, C. (2023). Emotional attachment to AI companions and European law. *MIT Case Studies in Social and Ethical Responsibilities of Computing* (Winter 2023). https://doi.org/10.21428/2c646de5.db67ec7f

142. Department of Enterprise, Tourism and Employment. (2026, February 4). General scheme of the Regulation of Artificial Intelligence Bill 2026. Government of Ireland. https://www.gov.ie/en/department-of-enterprise-tourism-and-employment/publications/general-scheme-of-the-regulation-of-artificial-intelligence-bill-2026

143. United States Congress. Senate. (2025). *S. 3062—GUARD Act: Guidelines for User Age-verification and Responsible Dialogue Act of 2025* (119th Congress). Congress.gov. https://www.congress.gov/bill/119th-congress/senate-bill/3062/text

144. European Commission. (2026, April 29). *Blueprint for an age verification solution to help protect minors online*. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/factpages/blueprint-age-verification-solution-help-protect-minors-online

145. Reuters. (2026, April 24). *From Australia to Europe, countries move to curb children's social media access*. https://www.reuters.com/legal/government/australia-europe-countries-move-curb-childrens-social-media-access-2026-04-24/

146. Morrin H, Nicholls L, Levin M et al. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. The Lancet Psychiatry, 2026; 13, 522-530

147. Examining the harm of AI chatbots, Hearing before the Subcomm. on Crime and Counterterrorism of the S. Comm. on the Judiciary, 119th Cong. (2025) (testimony of Megan Garcia). https://www.judiciary.senate.gov/download/2025-09-16-pm_-testimony_-garciapdf

148. Solove, D. J. (2025). Artificial intelligence and privacy. *Florida Law Review, 77*. https://scholarship.law.ufl.edu/flr/vol77/iss1/1

149. Office of the United Nations High Commissioner for Human Rights. (2025). *The right to privacy in the digital age* (UN Doc. A/HRC/60/45). https://www.ohchr.org/en/documents/thematic-reports/ahrc6045-right-privacy-digital-age-reportoffice-united-nations-high

150. Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics, 143*(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

151. UNESCO, IRCAI, & University College London. (2024). *Challenging systematic prejudices: An investigation into bias against women and girls in large language models*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000388971

152. Shah, S. S. (2024). Gender bias in artificial intelligence: Empowering women through digital literacy. *Premier Journal of Artificial Intelligence, 1*, 1000088. https://doi.org/10.70389/PJAI.1000088

153. Haider, S. A., Borna, S., Gomez-Cabello, C. A., et al. (2026). The algorithmic divide: A systematic review on AI-driven racial disparities in healthcare. *Journal of Racial and Ethnic Health Disparities, 13*, 188–217. https://doi.org/10.1007/s40615-024-02237-0

154. International Telecommunication Union. (2025). *Joint statement on AI and the rights of the child*. International Telecommunication Union. https://www.itu.int/hub/publication/d-str-cyb_joint-2025

155. Johnson, A. K., Winther, D. K., & Bhargava, A. (2026). *Artificial intelligence and child sexual exploitation and abuse: Emerging risks and implications for children's rights* (Issue brief). United Nations Children's Fund (UNICEF). https://www.unicef.org/media/178571/file/UNICEF%20AI%20CSEA%20Brief_FINAL3.pdf

156. Thiel, D. (2023). Identifying and Eliminating CSAM in Generative ML Training Data and Models. Stanford Digital Repository. Available at https://purl.stanford.edu/kh752sm9123

157. Internet Watch Foundation. (2026). Harm without limits: AI child sexual abuse material through the eyes of our Analysts. https://www.iwf.org.uk/media/hl1nvdti/iwf-ai-csam-report-2026.pdf

158. Goodacre, E.J. and Gibson, J.L. (2026) AI in the early years: Examining the implications of GenAI toys for young children. Cambridge: University of Cambridge (unpublished report, available via Apollo repository).

159. Kurian, N. (2025). Developmentally aligned AI: a framework for translating the science of child development into AI design. AI, Brain and Child, 1(1). https://doi.org/10.1007/s44436-025-00009-z

160. Livingstone, S., & Sylwander, K. R. (2025). Conceptualizing age-appropriate social media to support children's digital futures. British Journal of Developmental Psychology. https://doi.org/10.1111/bjdp.70006

161. Xiao, W., & Gonçalves, A. (2025). Intelligent toys, complex questions: A literature review of artificial intelligence in children's toys and devices. Big Data & Society, 12(4), 20539517251389860.

162. Grimmelikhuijsen, S. (2022). Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. Public Administration Review, 82(4), 706–718. https://doi.org/10.1111/puar.13483

163. Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review Online, 94, 15–55. https://ssrn.com/abstract=3333423

164. Mantelero, A. (2022). Human rights impact assessment and AI. In Beyond data: Human rights, ethical and social impact assessment in AI (pp. 45-91). The Hague: TMC Asser Press.

165. Mantelero, A., & Esposito, M. S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. Computer Law and Security Review, 41. https://doi.org/10.1016/j.clsr.2021.105561

166. United Nations Educational, Scientific and Cultural Organization. (2025). *How should children's rights be integrated into AI governance?* https://www.unesco.org/en/articles/how-should-childrens-rights-be-integrated-ai-governance

167. Livingstone, S. & Pothong, K. (2025). Child Rights Impact Assessment: A Policy Tool for aRights-Respecting Digital Environment. Policy & Internet.

168. Denain, J.-S., & Barry, A. (2026, April 16). *Have AI capabilities accelerated?* Epoch AI. https://epoch.ai/blog/have-ai-capabilities-accelerated

169. Model Evaluation & Threat Research (METR). (2026, May 8). *Task-completion time horizons of frontier AI models*. https://metr.org/time-horizons/

170. Juniewicz, I. (2026, February 26). *Hyperscaler capex has quadrupled since GPT-4's release*. Epoch AI. https://epoch.ai/data-insights/hyperscaler-capex-trend

171. Epoch AI. (2026, May 28). *Data on AI companies.* https://epoch.ai/data/ai-companies?view=graph&tab=revenue&showTotal=true

172. Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5), 612-634.

173. AI Alliance Network. (2025). AI Horizons: What will AI technologies look like in 10 years? A Research Project. November 2025, p. 83. Based on 21 foresight sessions and 32 in-depth interviews with over 270 AI researchers from 36 countries.

174. Gommers, J., Hernström, V., Josefsson, V., Sartor, H., Schmidt, D., Hjelmgren, A., ... & Lång, K. (2026). Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. The Lancet, 407(10527), 505-514.

175. Reichstein, M., et al. (2025). Early warning of complex climate risk with integrated artificial intelligence. *Nature Communications, 16*, 2564. https://www.nature.com/articles/s41467-025-57640-w

176. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15(1), 17458.

177. Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. npj Science of Learning, 10(1), 29.

178. Delgado-Licona, F., Alsaiari, A., Dickerson, H., Klem, P., Ghorai, A., Canty, R. B., ... & Abolhasani, M. (2025). Flow-driven data intensification to accelerate autonomous inorganic materials discovery. Nature Chemical Engineering, 2(7), 436-446.

179. Rotz, S., Duncan, E., Small, M., Botschner, J., Dara, R., Mosby, I., ... & Fraser, E. D. (2019). The politics of digital agricultural technologies: a preliminary review. Sociologia ruralis, 59(2), 203-229.

180. Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., Gravel Sullivan, A., Wills, G., ... & Gordon, J. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI, 2(12), AIoa2500945.

181. Chen, M., Wu, Y., Ma, J., Jia, X., Gao, C., Zhao, F., & Qiao, Y. (2026). Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. npj Digital Medicine.

182. Tao, X., Zhou, S., Ding, K., Li, S., Li, Y., Wu, B., ... & Han, S. (2026). An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial. Nature Medicine, 1-9.

183. Costa-Gomes, B., Tolmachev, P., Taysom, E., Sounderajah, V., Richardson, H., Schoenegger, P., ... & King, D. (2026). Public use of a generalist LLM chatbot for health queries. Nature Health, 1-8.

184. Scientific Computing World. (2025, August 5). *AI health assistant displays high diagnostic accuracy in tests*. https://www.scientific-computing.com/article/sberhealths-gigachat-powered-ai-health-assistant-displays-high-diagnostic-accuracy-tests

185. MASTEL, P. M., de Dieu Nyandwi, J., Rutunda, S., & Kabanda, K. (2025). Mbaza RBC: Deploying and evaluation of an LLM powered Chatbot for Community Health Workers in Rwanda. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.

186. Mateen, B. A., Williams, G., Korom, R., Mwaniki, P., Emmanuel-Fabula, M., & Agweyu, A. (2026). Learning Effects from A GenAI-based Clinical Decision Support System in Primary Healthcare. *medRxiv*, 2026-05.

187. OECD (2025), Results from TALIS 2024: The State of Teaching, TALIS, OECD Publishing, Paris, https://doi.org/10.1787/90df6235-en.

188. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en .

189. Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K–12 education. *npj Science of Learning, 10*(1), 29. https://www.nature.com/articles/s41539-025-00320-7

190. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en.

191. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en.

192. Vodafone Foundation.(2025) AI in European Schools: A European Report --- comparing seven countries

193. World Food Programme. (2024). *HungerMap LIVE: Global hunger monitoring*. https://hungermap.wfp.org/

194. Yakov and Partners. (2024). *Artificial intelligence in Russia's agricultural sector: Hype or real money?* https://yakovpartners.com/publications/ai-in-agriculture/

195. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

196. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports, 15*, 17458. https://doi.org/10.1038/s41598-025-97652-6

197. Shneiderman, B. (2022). Human-Centered AI. Oxford University Press.

198. Sparling, T. M., Offner, C., Deeney, M., Denton, P., Bash, K., Juel, R., ... & Kadiyala, S. (2024). Intersections of climate change with food systems, nutrition, and health: an overview and evidence map. Advances in Nutrition, 15(9), 100274.

199. Reichstein, M., et al. (2025). Early warning of complex climate risk with integrated artificial intelligence. Nature Communications, 16, 2564. https://doi.org/10.1038/s41467-025-57640-w

200. Becker-Reshef, I., Justice, C., Barker, B., Humber, M., Rembold, F., Bonifacio, R., Zappacosta, M., Budde, M., Magadzire, T., Shitote, C., Pound, J., Constantino, A., Nakalembe, C., Mwangi, K., Sobue, S., Newby, T., Whitcraft, A., Jarvis, I., & Verdin, J. (2020). Strengthening agricultural decisions in countries at risk of food insecurity: The GEOGLAM Crop Monitor for Early Warning. *Remote Sensing of Environment, 237*, 111553. https://doi.org/10.1016/j.rse.2019.111553

201. Food and Agriculture Organization of the United Nations. (2025). *Evidence in action: How anticipatory cash transfers reduce humanitarian needs and strengthen resilience in Somalia*. https://www.fao.org/agrifood-economics/publications/detail/en/c/1756210/

202. World Food Programme. (2025). *WFP's evidence base on anticipatory action 2015–2024*. https://www.wfp.org/publications/wfps-evidence-base-anticipatory-action-2015-2024

203. Nakalembe, C., Kerner, H. R., Zvonkov, I., Humber, M., Galvez, A. S., Venturini, S., & Becker Reshef, I. (2025). A framework for EO based National Agricultural Monitoring for the African context. npj Sustainable Agriculture, 3, 45. https://www.nature.com/articles/s44264-025-00083-z

204. Nowak, A. C., et al. (2024). Opportunities to strengthen Africa's efforts to track national level climate adaptation. Nature Climate Change, 14, 876–882.https://www.nature.com/articles/s41558-024-02054-7

205. Karger, E., Kuusela, O., Abaluck, J., Bryan, K. A., Halperin, B., Jones, T. R., et al. (2026). *Forecasting the economic effects of AI* (NBER Working Paper No. w35046). National Bureau of Economic Research. https://doi.org/10.3386/w35046

206. Goldman Sachs (2023). Generative AI Could Raise Global GDP by 7 Percent. https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent

207. Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: a review. Artificial intelligence review, 55(1), 589-656.

208. United Nations News. (2026, February 18). *Artists face steep income decline due to AI, UNESCO finds*. United Nations. https://news.un.org/en/story/2026/02/1166989

209. Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. American Economic Journal: Macroeconomics, 13(1), 333-372.

210. Gmyrek, P., Winkler, H., & Garganta, S. (2024). Buffer or Bottleneck? Employment Exposure to Generative AI and the Digital Divide in Latin America (ILO Working Paper 121 / World Bank Policy Research Working Paper 10863). International Labour Organization & World Bank.

211. Autor, D., Chin, C., Salomons, A., & Seegmiller, B. (2024). New frontiers: The origins and content of new work, 1940–2018. Quarterly Journal of Economics, 139(3), 1399–1465.

212. The Conversation. (2026, March 18). *Tech companies are blaming massive layoffs on AI. What's really going on?* https://theconversation.com/tech-companies-are-blaming-massive-layoffs-on-ai-whats-really-going-on-278314

213. Society for Human Resource Management. (2026, May). *The AI layoffs narrative: Real transformation, or scapegoat?* https://www.shrm.org/topics-tools/news/technology/ai-layoffs-transformation-scapegoat

214. OpenAI. (2026, February 27). Company communication accompanying a US$110 billion private funding round [Company communication]

215. Rodrik, D., & Sabel, C. (2020). Building a Good Jobs Economy (HKS Faculty Research Working Paper RWP20-001). Harvard Kennedy School.

216. Acemoglu, D. (2025). The simple macroeconomics of AI. Economic Policy, 40(121), 13–58. Acemoglu's headline figure is a TFP gain of no more than 0.71% over ten years.

217. Korinek, A., & Suh, D. (2024). Scenarios for the Transition to AGI (NBER Working Paper No. 32255). In their full-automation scenario, output rises sharply while wages collapse once automation crosses a critical threshold.

218. Karger, E., Kuusela, O., Abaluck, J., Bryan, K., Halperin, B., Jones, T., et al. (2026). Forecasting the Economic Effects of AI. Federal Reserve Bank of Chicago and Forecasting Research Institute, March 2026.

219. Jones, C. I., & Tonetti, C. (2026). Past Automation and Future AI: How Weak Links Tame the Growth Explosion. Working paper presented at the Bendheim Center for Finance, Princeton University, March 2026.

220. Trammell, P., & Korinek, A. (2025). Economic Growth under Transformative AI (NBER Working Paper No. 31815, revised September 2025).

221. OECD (2024). Artificial Intelligence, Data and Competition (OECD Artificial Intelligence Papers No. 18). OECD Publishing, Paris. https://doi.org/10.1787/e7e88884-en

222. Acemoglu, D., & Restrepo, P. (2026). Automation and rent dissipation: Implications for wages, inequality, and productivity. Quarterly Journal of Economics, 141(2), 1521–1579.

223. Liu, Y., Zhao, Y., Lyu, Y., Zhang, T., Wang, H., & Lo, D. (2025). "Your AI, my shell": Demystifying prompt injection attacks on agentic AI coding editors. arXiv. https://doi.org/10.48550/arXiv.2509.22040

224. Regilme, S. S. F. (2024). Artificial Intelligence Colonialism: Environmental Damage, Labor Exploitation, and Human Rights Crises in the Global South. SAIS Review of International Affairs, 44(2), 75–92. https://muse.jhu.edu/pub/1/article/950958

225. Barnett-Itzhaki, Z. (2026). The water footprint of artificial intelligence: Emerging solutions and governance imperatives. Water Research, 299, 125866. https://doi.org/10.1016/j.watres.2026.125866

226. International Energy Agency. (2025). Global Critical Minerals Outlook 2025 – Analysis (p. 312). https://iea.blob.core.windows.net/assets/ef5e9b70-3374-4caa-ba9d-19c72253bfc4/GlobalCriticalMineralsOutlook2025.pdf

227. Zhu, L., & Lu, Q. (2026). Verifiability-First AI Engineering in the Era of AIware: A Conceptual Framework, Design Principles, and Architectural Patterns for Scalable Verification. Design Principles, and Architectural Patterns for Scalable Verification (January 07, 2026).

228. UN Scientific Advisory Board. (2026, March 19). AI deception: Brief of the Scientific Advisory Board. United Nations. https://www.un.org/scientific-advisory-board/en/ai-deception

229. Bengio, Y., Clare, S., Prunkl, C., Rismani, S., Andriushchenko, M., Bucknall, B., ... & Zhu, L. (2025). International AI Safety Report 2025: First Key Update: Capabilities and Risk Implications. *arXiv preprint arXiv:2510.13653*.

230. United Nations Secretary-General. (2024). *Human rights due diligence guidance on digital technology use*. https://www.ohchr.org/sites/default/files/2024-08/digital-technology-use-guidance-sg-1-en.pdf

231. AI Equality Toolbox. (2025). *Human rights impact assessment methodology*. https://aiequalitytoolbox.com/tools/hria-workbook/

232. Baldé, C. P., Kuehr, R., Yamamoto, T., McDonald, R., D'Angelo, E., Althaf, S., Bel, G., Deubzer, O., Fernandez-Cubillo, E., Forti, V., Gray, V., Herat, S., Honda, S., Iattoni, G., Khetriwal, D. S., Luda di Cortemiglia, V., Lobuntsova, Y., Nnorom, I., Pralat, N., & Wagner, M. (2024). *The global e-waste monitor 2024: Electronic waste rising five times faster than documented e-waste recycling*. United Nations Institute for Training and Research & International Telecommunication Union. https://ewastemonitor.info/the-global-e-waste-monitor-2024/

233. International Panel on the Information Environment. (2024). *Expert survey on the global information environment 2024: Searching for solutions (SFP2024-1)*. https://www.ipie.info/research/sfp2024-1

234. Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining, 13*(1), 30. https://doi.org/10.1007/s13278-023-01028-5

235. James, K., Perveen, S., & Jacobson, B. (2025). *Drained by data: The cumulative impact of data centers on regional water stress*. Ceres. https://www.ceres.org/resources/reports/drained-by-data-the-cumulative-impact-of-data-centers-on-regional-water-stres

236. Office of the United Nations High Commissioner for Human Rights. (2024). *Taxonomy of generative AI-related human rights harms*. https://www.ohchr.org/sites/default/files/documents/issues/expression/statements/2025-10-24-joint-declaration-artificial-intelligence.pdf

237. Patel, A. (2025, May 19). *Freedom of expression, artificial intelligence and elections*. United Nations Educational, Scientific and Cultural Organization (UNESCO) & United Nations Development Programme (UNDP). https://www.undp.org/publications/freedom-expression-artificial-intelligence-and-elections

238. Scharfbillig, M., Lewandowsky, S., Altay, S., Van Alstyne, M., Kozyreva, A., Hertwig, R., Lorenz-Spreen, P., DiResta, R., Valenzuela, S., Egidy, S., Quattrociocchi, W., & Orben, A. (2026). *Fractured reality: How democracy can win the global struggle over the information space* (JRC144603). Publications Office of the European Union. https://doi.org/10.2760/9358883

239. Observatory on Information and Democracy. (2024). *Information ecosystems and troubled democracy: A global synthesis of the state of knowledge on news, media, AI, and data governance*. https://observatory.informationdemocracy.org/report/information-ecosystem-and-troubled-democracy/

240. Solove, D. J. (2025). *Artificial intelligence and privacy*. *Florida Law Review, 77*. https://scholarship.law.ufl.edu/flr/vol77/iss1/1

241. Office of the United Nations High Commissioner for Human Rights (2025). The right to privacy in the digital age. URL https://www.ohchr.org/en/documents/thematic-reports/ahrc6045-right-privacy-digital-age-reportoffice-united-nations-high. UN Doc. A/HRC/60/45.

242. Council of Europe. Steering Committee on Media and Information Society. (2025). *Guidance note on the implications of generative artificial intelligence for freedom of expression* (CDMSI(2025)15rev). https://rm.coe.int/cdmsi-2025-15rev-guidance-note-on-the-implications-of-generative-artif/488029df80

243. European Union. (2024, June 13). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. *Official Journal of the European Union, L series*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

244. Expert Mechanism on the Right to Development. (2024). *Artificial intelligence, cultural rights and the right to development* (A/HRC/EMRTD/13/CRP.1). United Nations Human Rights Council. https://www.ohchr.org/sites/default/files/documents/issues/development/emd/session13/a-hrc-emrtd-13-crp-1.pdf

245. Council of Europe, Steering Committee on Media and Information Society. (2025). *Guidance note on the implications of generative artificial intelligence for freedom of expression* (CDMSI(2025)15rev). https://rm.coe.int/cdmsi-2025-15rev

246. Waight, H., Yang, E., Yuan, Y., et al. (2026). State media control influences large language models. *Nature*. https://doi.org/10.1038/s41586-026-10506-7

247. Li, P., Yang, J., Islam, M. A., & Ren, S. (2025). Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models. *Communications of the ACM, 68*(7), 54–61. https://doi.org/10.1145/3724499

248. Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., Goodman, S., ... & Manyika, J. (2025). Measuring the environmental impact of delivering AI at Google Scale. arXiv preprint arXiv:2508.15734.

249. Arsenault, A. C., & Kreps, S. (2026). Whose voice counts? The role of large language models in public commenting. *Big Data & Society, 13*(1). https://doi.org/10.1177/20539517261419341

250. Alslaity, A., Chan, G., & Orji, R. (2023). A panoramic view of personalization based on individual differences in persuasive and behavior change interventions. *Frontiers in Artificial Intelligence, 6*, 1125191. https://doi.org/10.3389/frai.2023.1125191

251. Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science, 385*, eadq1814. https://doi.org/10.1126/science.adq1814

252. Boissin, E., Costello, T. H., Spinoza-Martín, D., Rand, D. G., & Pennycook, G. (2025). Dialogues with large language models reduce conspiracy beliefs even when the AI is perceived as human. *PNAS Nexus, 4*(11), pgaf325. https://doi.org/10.1093/pnasnexus/pgaf325

253. Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., ... & Kunst, J. R. (2026). How malicious AI swarms can threaten democracy. Science, 391(6783), 354-357.

254. Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational AI. Science, 390(6777), eaea3884. https://doi.org/10.1126/science.aea3884

255. Germano, F., Gómez, V., & Sobbrio, F. (2025). *Ranking for engagement: How social media algorithms fuel misinformation and polarization*. Barcelona School of Economics, Working Paper No. 1501. https://bw.bse.eu/wp-content/uploads/2025/07/1501.pdf

256. Council of Europe CDMSI. (2025). Guidance Note on Generative AI and Freedom of Expression. CDMSI(2025) https://rm.coe.int/cdmsi-2025-15rev-guidance-note-on-the-implications-of-generative-artif/488029df80

257. Buyl, M., Rogiers, A., Noels, S., Bied, G., Dominguez-Catena, I., Heiter, E., ... & De Bie, T. (2026). Large language models reflect the ideology of their creators. *npj Artificial Intelligence*, *2*(1), 7.

258. Wang, P., Zhang, L.-Y., Tzachor, A., & Chen, W.-Q. (2024). E-waste challenges of generative artificial intelligence. Nature Computational Science, 4, 818–823. https://doi.org/10.1038/s43588-024-00700-3

259. Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., ... & Kunst, J. R. (2026). How malicious AI swarms can threaten democracy. Science, 391(6783), 354-357.

260. Council of Europe. (2024). *Council of Europe framework convention on artificial intelligence and human rights, democracy and the rule of law* (Council of Europe Treaty Series No. 225). https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

261. Observatory on Information and Democracy. (2024). *Information ecosystems and troubled democracy: A global synthesis of the state of knowledge on news, media, AI, and data governance*. https://observatory.informationdemocracy.org/report/information-ecosystem-and-troubled-democracy/

262. Council of Europe, Steering Committee on Media and Information Society. (2025). *Guidance note on the implications of generative artificial intelligence for freedom of expression* (CDMSI(2025)15rev). https://rm.coe.int/cdmsi-2025-15rev

263. OECD. (2024). Facts not fakes: Tackling disinformation, strengthening information integrity. OECD Publishing. https://doi.org/10.1787/d909ff7a-en

264. United Nations General Assembly. (2017). *Promotion and protection of human rights: Human rights questions, including alternative approaches for improving the effective enjoyment of human rights and fundamental freedoms: Report of the Third Committee, 72nd session* (A/72/...). United Nations Digital Library. http://digitallibrary.un.org/record/1326669

265. Penney, J. W. (2025). *Chilling effects: Repression, conformity, and power in the digital age*. Cambridge University Press. https://doi.org/10.1017/9781108918022

266. UN Women. (2022). Tipping point: The chilling escalation of online violence against women in the public sphere. UN Women. https://www.unwomen.org/sites/default/files/2025-12/tipping-point-the-chilling-escalation-of-violence-against-women-in-the-public-sphere-in-the-age-of-ai-en.pdf

267. United Nations Educational, Scientific and Cultural Organization. (2024). *Challenging systematic prejudices: An investigation into bias against women and girls in large language models*. https://unesdoc.unesco.org/ark:/48223/pf0000388971

268. Chowdhury, R., & Lakshmi, D. (2023). “Your opinion doesn’t matter, anyway”: Exposing technology-facilitated gender-based violence in an era of generative AI (2nd ed.). UNESCO.

269. United Nations Educational, Scientific and Cultural Organization. (2024, March 7). *Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes*. https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes

270. Fung, P. (2019, June 30). *This is why AI has a gender problem*. World Economic Forum. https://www.weforum.org/stories/2019/06/this-is-why-ai-has-a-gender-problem/

271. United Nations Conference on Trade and Development. (2025). *Technology and innovation report 2025: The AI divide*. https://unctad.org/publication/technology-and-innovation-report-2025

272. Ahmed, N., & Wahed, M. (2020). The De-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*.

273. Coeckelbergh, M. (2026) Technofascism: AI, Big Tech, and the New Authoritarianism. AI & Society https://doi.org/10.1007/s00146-026-02862-9

274. Varoufakis, Y. (2024). *Technofeudalism: What killed capitalism*. Melville House.

275. Kalluri, P. R., Agnew, W., Cheng, M., Owens, K., Soldaini, L., & Birhane, A. (2025). Computer-vision research powers surveillance technology. Nature, 643(8070), 73-79. https://www.nature.com/articles/s41586-025-08972-6

276. Fola-Rose, A., Solomon, E., Bryant, K., & Woubie, A. (2024, August). A systematic review of facial recognition methods: Advancements, applications, and ethical dilemmas. In *Proceedings of the 2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI 2024)* (pp. 314–319). IEEE. https://doi.org/10.1109/IRI62200.2024.00070

277. Fussey, P., & Murray, D. (2025). Facial Recognition Surveillance: Policing and Human Rights in the Age of Artificial Intelligence. Oxford University Press.

278. Office of the United Nations High Commissioner for Human Rights. (2024). *Mapping report: Human rights and new and emerging digital technologies* (A/HRC/56/45). https://www.ohchr.org/en/documents/reports/mapping-report-human-rights-and-new-and-emerging-digital-technologies

279. Secretary-General. (2024). Human rights in the administration of justice (A/79/296). United Nations. https://digitallibrary.un.org

280. American Civil Liberties Union, More than a Dozen Wrongful Arrests Due to Police Reliance on Facial Recognition Technology (2025), https://www.aclu.org/news/privacy-technology/more-than-a-dozen-wrongful-arrests-due-to-police-reliance-on-facial-recognition-technology: Documentation of at least 14 publicly known wrongful arrests in the United States attributed to police use of facial recognition; in nearly all cases the persons wrongfully arrested were Black.

281. Saxena, D., & Guha, S. (2024). Algorithmic harms in child welfare: Uncertainties in practice, organization, and street-level decision-making. *ACM Journal on Responsible Computing, 1*(1), Article 2, 1–32. https://doi.org/10.1145/3616473

282. German Marshall Fund. (2024). Spitting Images: Tracking Deepfakes and Generative AI in Elections.

283. International Institute for Democracy and Electoral Assistance. (2024). *The 2024 global elections super-cycle*. https://www.idea.int/initiatives/the-2024-global-elections-supercycle

284. Recorded Future. (2024). 2024 Deepfakes and Election Disinformation Report: Key Findings and Mitigation Strategies.

285. Associated Press. (2025, June 13). New Hampshire jury acquits consultant behind AI robocalls mimicking Biden on all charges.

286. Federal Communications Commission. (2024, September). $6 million fine against Steven Kramer for AI-generated robocalls.

287. Global Witness. (2024). What Happened on TikTok Around the Romanian Elections?

288. IFES. (2024). The Romanian 2024 Election Annulment: Addressing Emerging Threats to Electoral Integrity.

289. Associated Press. (2024). Election disinformation takes a big leap with AI being used to deceive worldwide.

290. United Nations. (1966). International Covenant on Civil and Political Rights. Articles 17, 18, 19 and 25.

291. Council of Europe. (1950). Convention for the Protection of Human Rights and Fundamental Freedoms. Articles 8, 9 and 10; Protocol No. 1, Article 3.

292. EQUATE Language AI Readiness Index (https://equate.vercel.app/en)

293. Han, W., Zhang, Y., Chen, Z., Liu, B., Lin, H., Zhang, B., ... & Zheng, Y. (2025). MuBench: Assessment of Multilingual Capabilities of Large Language Models Across 61 Languages. arXiv preprint arXiv:2506.19468.

294. Lissak, S., Calderon, N., Shenkman, G., Ophir, Y., Fruchter, E., Klomek, A. B., & Reichart, R. (2024, June). The colorful future of llms: Evaluating and improving llms as emotional supporters for queer youth. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 2040-2079).

295. Gamboa, L. C. L., Feng, Y., & Lee, M. (2025, November). Social Bias in Multilingual Language Models: A Survey. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 27845-27868).

296. Choudhury, M., Sitaram, S., Vashistha, A., et al. (2026). *AI for the Global South: 12 critical research questions for the next decade*. AI for the Global South (AI4GS), Mohamed bin Zayed University of Artificial Intelligence. https://ai4gs.github.io/

297. Bartl, M., Mandal, A., Leavy, S., & Little, S. (2025). Gender bias in natural language processing and computer vision: A comparative survey. ACM Computing Surveys, 57(6), 1-36. https://dl.acm.org/doi/pdf/10.1145/3700438

298. Robb, M. B., & Mann, S. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*. Common Sense Media. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf

299. U.S. PIRG Education Fund. (2025). AI comes to playtime: Artificial companions, real risks. U.S. PIRG Education Fund. https://pirg.org/edfund/wp-content/uploads/2025/12/AI-Comes-to-Playtime-Artifical-companions-real-risks.pdf

300. Staksrud, E., Mascheroni, G., Milosevic, T., Ní Bhroin, N., Ólafsson, K., Şengül-İnal, G., & Stoilova, M. (2026). European children's use and understanding of generative AI. EU Kids Online V.

301. Committee on the Rights of the Child. (2021). *General comment No. 25 (2021) on children's rights in relation to the digital environment* (CRC/C/GC/25). United Nations. https://digitallibrary.un.org/record/3906061

302. UNICEF (2025). Guidance on AI and Children: Updated guidance for governments and businesses to create AI policies and systems that uphold children's rights. https://www.unicef.org/innocenti/media/11991/file/UNICEF-Innocenti-Guidance-on-AI-and-Children-3-2025.pdf

303. UNESCO (2025). How should children's rights be integrated into AI governance? https://www.unesco.org/en/articles/how-should-childrens-rights-be-integrated-ai-governance

304. Grossman, S., Pfefferkorn, R., & Liu, S. (2025). AI-Generated Child Sexual Abuse Material: Insights from Educators, Platforms, Law Enforcement, Legislators, and Victims. Version 1. Stanford Digital Repository. Available at https://purl.stanford.edu/mn692xc5736/version/1. https://doi.org/10.25740/mn692xc5736.

305. Livingstone, S., Atabey, A., Stoilova, M., & Sylwander, K. R. (2025). How does, and how could, generative AI respect and enable children's rights? In N. Ni Loideain (Ed.), *AI and power: Regulation and rights*. University of London Press.

306. European Parliamentary Research Service. (2024). Children and deepfakes. European Parliament. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2025)775855

307. United Nations Children's Fund (UNICEF). (2026). *Artificial intelligence and child sexual abuse and exploitation* [Issue brief]. https://www.unicef.org/reports/artificial-intelligence-and-child-sexual-abuse-and-exploitation

308. Ozcan, B., Sperati, V., Giocondo, F., Schembri, M., & Baldassarre, G. (2022, June). Interactive soft toys to support social engagement through sensory-motor plays in early intervention of kids with special needs. In Proceedings of the 21st Annual ACM Interaction Design and Children Conference (pp. 625-628).

309. Goodacre, E., & Gibson, J. (2026). AI in the Early Years: Examining the implications of GenAI toys for young children. https://www.cam.ac.uk/stories/ai-toys-study-play

310. British Standards Institution. (2026, May). *Half of children have AI toys despite safety concerns*. https://www.bsigroup.com/en-GB/insights-and-media/media-centre/press-releases/2026/may/half-of-children-have-ai-toys-as-parents-allow-widespread-use-despite-safety-concerns-and-gaps-in-guidance-parents/

311. Goodacre, E., & Gibson, J. (2026). AI in the Early Years: Examining the implications of GenAI toys for young children. https://doi.org/10.17863/CAM.126270

312. Chou, C. Y., Chan, T. W., Chen, Z. H., Liao, C. Y., Shih, J. L., Wu, Y. T., & Hung, H. C. (2025). Defining AI companions: a research agenda--from artificial companions for learning to general artificial companions for Global Harwell. *Research & Practice in Technology Enhanced Learning*, *20*.

313. Ho, J. Q., Hu, M., Chen, T. X., & Hartanto, A. (2025). Potential and pitfalls of romantic artificial intelligence companions: A systematic review. Computers in Human Behavior Reports, 19, 100715.

314. Hollanek, T., & Sobey, A. (2025). AI companions for health and mental wellbeing: opportunities, risks and policy implications. Leverhulme Centre for the Future of Intelligence

315. De Freitas, J., Oguz-Uguralp, Z., & Kaan-Uguralp, A. (2025). Emotional manipulation by AI companions. arXiv preprint arXiv:2508.19258.

316. Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). The rise of AI companions: how human-chatbot relationships influence well-being. arXiv preprint arXiv:2506.12605.

317. Dewitte, P. (2024). Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. Computer Law & Security Review, 54, 106019.

318. Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., & Lee, Y. C. (2025, April). The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In Proceedings of the 2025 CHI conference on human factors in computing systems (pp. 1-17).

319. De Freitas, J., & Cohen, I. G. (2024). The health risks of generative AI-based wellness apps. Nature medicine, 30(5), 1269-1275.

320. Radesky, J., Bragg, M. A., & Hiniker, A. (2026). Risks and Consequences of Children's Use of Social AI—A Framework. JAMA Pediatrics. https://doi.org/10.1001/jamapediatrics.2026.1349

321. Muldoon, J., & Parke, J. J. (2025). Cruel companionship: How AI companions exploit loneliness and commodify intimacy. new media & society, 14614448251395192.

322. Jacobs, K. A. (2024). Digital loneliness—changes of social recognition through AI companions. Frontiers in Digital Health, 6, 1281037.

323. Ho, J. Q., Hu, M., Chen, T. X., & Hartanto, A. (2025). Potential and pitfalls of romantic Artificial Intelligence (AI) companions: A systematic review. Computers in Human Behavior Reports, 19, 100715.

324. Hollanek, T., & Sobey, A. (2025). AI companions for health and mental wellbeing: opportunities, risks and policy implications.

325. Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). The rise of AI companions: how human-chatbot relationships influence well-being. arXiv preprint arXiv:2506.12605.

326. Dewitte, P. (2024). Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. Computer Law & Security Review, 54, 106019.

327. Robb, M. B., & Mann, S. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*. Common Sense Media. https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions

328. Rousmaniere, T., Zhang, Y., Li, X., & Shah, S. (2025). Large language models as mental health resources: Patterns of use in the United States. *Practice Innovations*. https://doi.org/10.1037/pri0000292

329. Callahan, C., Tanner, L., Coe, C., Davis, M., Glover, J., Bernstein, E., ... & Kunkle, S. (2026). Real-World Use of a Mental Health AI Companion: Multiple Methods Study. JMIR Formative Research, 10, e86904.

330. Associated Press. (2026, June 1). *Chatbot AI lawsuit alleges links to teen suicide*. https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0

331. Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". JMIR Mental Health, 12(1), e85799.

332. Green, H. H. (2026, March 14). *New study raises concerns about AI chatbots fueling delusional thinking*. The Guardian. https://www.theguardian.com/technology/2026/mar/14/ai-chatbots-psychosis

333. Ministère du Travail, de la Santé, des Solidarités et des Familles. (2025, March 24). *La santé mentale, grande cause nationale 2025*. Gouvernement de la France. https://solidarites.gouv.fr/la-sante-mentale-grande-cause-nationale-2025

334. American Psychiatric Association. (n.d.). *Applications of artificial intelligence in mental health care*. https://www.psychiatry.org/psychiatrists/practice/artificial-intelligence/applications

335. U.S. Food and Drug Administration. (2025). *Executive summary for the Digital Health Advisory Committee meeting: Generative artificial intelligence–enabled digital mental health medical devices*. https://www.fda.gov/media/189391/download

336. Hollis, A., & McKeown, G. (2024, September). Empathic AI for autism: Potential and pitfalls of empathic social chatbots in addressing loneliness. In 24th ACM International Conference on Intelligent Virtual Agents: CONNECT, A Workshop on Connecting Interdisciplinary Research on Connections With and Through Technology: IVA 2024.

337. Sharma, D., Meshkat, S., Perivolaris, A., Kamaleddin, M. A., Teferra, B. G., Rueda, A., ... & Bhat, V. (2026). Reimagining psychiatric care with agentic AI: promise, challenges, and a roadmap forward. npj Digital Medicine.

338. Straw, I., & Callison-Burch, C. (2020). Artificial Intelligence in mental health and the biases of language based models. PloS one, 15(12), e0240376.

339. Garg, M. (2024). Towards mental health analysis in social media for low-resourced languages. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(3), 1-22.

340. Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-T., Jiao, W., & Lyu, M. R. (2024). All languages matter: On the multilingual safety of LLMs. *Findings of the Association for Computational Linguistics: ACL 2024*, 5865–5877. https://doi.org/10.18653/v1/2024.findings-acl.349

341. Nigatu, H. H., Mehandru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). *Viability of machine translation for healthcare in low-resourced languages*. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), 10584–10598. https://aclanthology.org/2025.emnlp-main.535/

342. Fu, Y. V., Ramachandran, G. K., Park, N., Lybarger, K., Xia, F., Uzuner, Ö., & Yetisgen, M. (2025). BioMistral-NLU: Towards more generalizable medical language understanding through instruction tuning. *AMIA Joint Summits on Translational Science Proceedings, 2025*, 149–158. https://pubmed.ncbi.nlm.nih.gov/40502228/

343. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., & Dufour, R. (2024). BioMistral: A collection of open-source pretrained large language models for medical domains. *Findings of the Association for Computational Linguistics: ACL 2024*, 5848–5864. https://aclanthology.org/2024.findings-acl.348/

344. Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., & Xie, W. (2024). Towards building multilingual language models for medicine. *Nature Communications, 15*, 8384. https://doi.org/10.1038/s41467-024-52417-z

345. Nwabufo, J., Ogueji, K., Adelani, D. I., Alabi, J., et al. (2025). Healthcare NLP for African Languages: Current State and Challenges. Proceedings of the AfricaNLP Workshop (AfricaNLP 2025). https://aclanthology.org/2025.africanlp-1.32/

346. Okafor, U. (2025). Multilingual NLP for African Healthcare: Bias, Translation, and Explainability Challenges. Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), 221–229. https://aclanthology.org/2025.africanlp-1.32/

347. Skianis, K., Doğruöz, A. S., & Pavlopoulos, J. (2024). Leveraging LLMs for translating and classifying mental health data. In J. Sälevä & A. Owodunni (Eds.), Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024) (pp. 236–241). https://doi.org/10.18653/v1/2024.mrl-1.20

348. Cronin, A., Kelly, A., Wrona, M., O'Donnell, P., Hassan, A., Myles, T., Fallon, T., & MacFarlane, A. (2025). The patient-safety implications of AI-based communication with migrants in general practice: a scoping review. BJGP Open, 9(4), BJGPO.2025.0107. https://bjgpopen.org/content/9/4/BJGPO.2025.0107

349. House of Lords Public Services Committee. (2025). *Lost in translation? Interpreting services in the courts* (2nd Report of Session 2024–26, HL Paper 87). https://committees.parliament.uk/publications/44602/documents/221328/default/

350. Nigatu, H. H., Mehandru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). Viability of machine translation for healthcare in low-resourced languages. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 10584–10598. https://aclanthology.org/2025.emnlp-main.535/

351. MIT AI Risk Initiative. (2025). *AI risk mitigation database and draft taxonomy*. https://airisk.mit.edu/ai-riskmitigations

352. Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., ... & Manyika, J. (2024). The ethics of advanced AI assistants. *arXiv preprint arXiv:2404.16244*.

353. Staufer, L., Feng, K., Wei, K., et al. (2026). *The 2025 AI agent index: Documenting technical and safety features of deployed agentic AI systems*. https://doi.org/10.48550/arXiv.2602.17753

354. Weidinger, L., Raji, I. D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., ... & Isaac, W. (2025). Toward an evaluation science for generative ai systems. *arXiv preprint arXiv:2503.05336*.

355. Rabanser, S., Kapoor, S., Kirgis, P., et al. (2026). *Towards a science of AI agent reliability*. https://doi.org/10.48550/arXiv.2602.16666

356. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). Generative AI can harm learning. *The Wharton School Research Paper*.

357. Shen, J. H., & Tamkin, A. (2026). How AI impacts skill formation. *arXiv preprint arXiv:2601.20245*.

358. Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., ... & Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, *10*(10), 896-903.

359. Epoch AI. (2026). *Data on AI models*. https://epoch.ai/data/ai-models

360. Feng, K. J., McDonald, D. W., & Zhang, A. X. (2025). Levels of autonomy for ai agents. *arXiv preprint arXiv:2506.12469*.

361. Fink, M. (2025). *Operationalizing meaningful human oversight under Article 14 of the EU AI Act*. In *AI Act commentary: A thematic analysis* (forthcoming). Hart-Bloomsbury.

362. Shapira, N., Wendler, C., Yen, A., Sarti, G., Pal, K., Floody, O., ... & Bau, D. (2026). Agents of chaos. *arXiv preprint arXiv:2602.20021*.

363. Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., ... & Rahwan, I. (2025). Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*.

364. Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015). *Corrigibility*. Machine Intelligence Research Institute. https://intelligence.org/files/Corrigibility.pdf.

365. Zeng, Y., Lu, E., Guo, X., Huangfu, C., Xie, J., Chen, Y., ... & Younas, A. (2025). AI Governance InternationaL Evaluation Index (AGILE Index) 2025. *arXiv preprint arXiv:2507.11546*.

366. Bommasani, R., Kapoor, S., Klyman, K., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2023, December 13). *Considerations for governing open foundation models*. Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/policy/issue-brief-considerations-governing-open-foundation-mode

367. Tzachor, A., Devare, M., Richards, C., Pypers, P., Ghosh, A., Koo, J., ... & King, B. (2023). Large language models and agricultural extension services. *Nature food*, *4*(11), 941-948.

368. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature, 616*, 259–265. https://doi.org/10.1038/s41586-023-05881-4

369. Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2024). *The foundation model transparency index v1.1*. arXiv:2407.12929. https://arxiv.org/abs/2407.12929

370. BigScience Workshop, Le Scao, T., Fan, A., et al. (2022). *BLOOM: A 176B-parameter open-access multilingual language model*. arXiv. https://doi.org/10.48550/arXiv.2211.05100

371. Touvron, H., Lavril, T., Izacard, G., et al. (2023). *LLaMA: Open and efficient foundation language models*. arXiv. https://arxiv.org/abs/2302.13971

372. Qwen Team. (2024). *QwenLM*. GitHub. https://github.com/QwenLM/Qwen

373. Qwen Team. (2024). *Qwen2 technical report*. arXiv. https://arxiv.org/abs/2407.10671

374. DeepSeek-AI. (2024). *DeepSeek-V3 technical report*. arXiv. https://doi.org/10.48550/arXiv.2412.19437

375. DeepSeek-AI. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv. https://arxiv.org/abs/2501.12948

376. Jiang, A. Q., et al. (2023). *Mistral 7B*. arXiv. https://arxiv.org/abs/2310.06825. Mistral AI.

377. Technology Innovation Institute. (2023). *The Falcon series of open language models*. arXiv. https://arxiv.org/abs/2311.16867

378. SberDevices. (2026, March). *GigaChat-3.1: Большое обновление больших моделей*. Habr. https://habr.com/ru/companies/sberbank/articles/1014146/

379. Yandex. (2025, February 25). *YandexGPT 5 — в Алисе, облаке и опенсорсе*. Habr. https://habr.com/ru/companies/yandex/articles/885218/

380. Sarvam AI. (2025). *Sarvam AI models on Hugging Face*. https://huggingface.co/sarvamai

381. SB Intuitions. (2025). *Sarashina models on Hugging Face*. https://huggingface.co/sbintuitions

382. Naver Cloud. (2024). *HyperCLOVA X technical report*. arXiv. https://arxiv.org/abs/2404.01954

383. Seger, E., Dreksler, N., Moulange, R., et al. (2023). *Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives*. Centre for the Governance of AI. https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

384. Anthropic. (2025). *Disrupting the first reported AI-orchestrated cyber espionage campaign*. https://www.anthropic.com/news/disrupting-AI-espionage

385. Partnership on AI. (2023). *PAI's guidance for safe foundation model deployment*. https://partnershiponai.org/modeldeployment/

386. International Energy Agency. (2025). *Energy and AI*. International Energy Agency. https://www.iea.org/reports/energy-and-ai

Независимая международная научная группа по искусственному интеллекту

Состав и мандат

Независимая международная научная группа по искусственному интеллекту была учреждена в рамках Организации Объединенных Наций в соответствии с резолюцией 79/325 Генеральной Ассамблеи и обязательствами, предусмотренными в Глобальном цифровом договоре и Пакте во имя будущего.

В состав Группы входят 40 независимых экспертов, назначаемых Генеральной Ассамблеей на трехлетний срок на основе их выдающихся экспертных знаний в области ИИ и смежных областях; состав Группы гендерно сбалансирован, и в нее входят представители всех пяти региональных групп государств-членов, занимающиеся различными дисциплинами, в том числе основными техническими аспектами ИИ, развитием прикладного ИИ, безопасностью и инфраструктурой, а также вопросами политики, этики и последствий использования ИИ.

Группе поручено публиковать результаты проведенных с использованием фактических данных научных оценок, обобщая и анализируя существующие исследования, касающиеся возможностей и рисков, связанных с ИИ, и его воздействия, в виде одного ежегодного сводного доклада, имеющего отношение к политике, но не носящего предписывающего характера, и содержащего тематическую справочную информацию, которую Группа сочтет необходимой. В своей научной работе Группа должна руководствоваться принципами независимости, научной достоверности и строгости, междисциплинарности и инклюзивного участия.

Группе также поручено представлять свой ежегодный сводный доклад в рамках Глобального диалога Организации Объединенных Наций по вопросам управления искусственным интеллектом. Предоставляя информацию для Глобального диалога и ведения более широких международных процессов, Группа дает мировому сообществу возможность предвидеть формирующиеся вызовы, принимать более обоснованные решения в области управления и обеспечивать равную информированность сторон, ответственных за выработку политики, во всем мире.

Деятельность Группы поддерживает секретариат Группы, работу которого координирует Управление цифровых и новейших технологий Организации Объединенных Наций.

Процесс работы над настоящим предварительным докладом

После назначения членов Группы в феврале 2026 года ее первое заседание состоялось в марте 2026 года, и на протяжении трех месяцев, прошедших после него, Группа интенсивно работала над подготовкой настоящего доклада. Этот интенсивный процесс научного обмена мнениями и коллективного анализа включал трехдневное очное пленарное заседание и более 60 виртуальных заседаний Группы, проведенных под руководством ее избранных сопредседателей.

Данный предварительный доклад, который можно принять за отправную точку, закладывает крайне важную основу для проведения более широких консультаций с внешними экспертами и для подготовки последующих бюллетеней с тематической справочной информацией и ежегодных докладов.

Научная независимость Группы

Члены Группы выступают в своем личном качестве экспертов, ведущих свою научную деятельность независимым образом. В связи с назначением в качестве экспертов Организации Объединенных Наций в командировке каждый из них обязался и обещал не запрашивать и не принимать указания относительно выполнения своих обязанностей от какого бы то ни было правительства или из иного источника. Положения, регулирующие статус, основные права и обязанности экспертов в командировках (ST/SGB/2002/9) также содержат положения, касающиеся их поведения и подотчетности.

Donors

The Panel Secretariat gratefully acknowledges the financial and in-kind contributions of the following governments and partners, without whom the Panel would not have been able to carry out its responsibilities:

Government of Germany
Government of Japan
Government of Spain
Omidyar Network Fund

Panel Secretariat

Coordinator

- Amandeep Singh Gill, United Nations Under-Secretary-General for Digital and Emerging Technologies

Secretariat Coordination & Support

- Quintin Chou-Lambert, United Nations Office for Digital and Emerging Technologies (UN ODET)
- Rebakah Hayoung Woo, UN ODET
- Peppi Väänänen, UN ODET

Fundraising & Logistics

- Sebastian Frank, UN ODET
- Antonieta Loaiza, UN ODET

Communications

- Karoline Hassfurter, UN ODET
- Brian Shung Seun Lau, UN ODET
- Anamika Madhuraj, UN ODET

Editing and Drafting Support*

- Jiaee Cheong, United Nations University (UNU)
- Kevin Kohler, UNU
- Max Springer, UNU

Rapporteurs

- Wernhard Berger, United Nations Industrial Development Organization
- Jin Cui, International Telecommunication Union (ITU)
- Tim Engelhardt, Office of the United Nations High Commissioner for Human Rights (UN OHCHR)
- Andrew Morritt, United Nations Department of Peace Operations
- Prateek Sibal, United Nations Educational, Scientific and Cultural Organization (UNESCO)
- Mariagrazia Squicciarini, UNESCO
- Oleksandra Vereschak, UNESCO
- Ana Gabriela Fernandez Vergara, ITU
- Li Zhou, UN OHCHR

* To ensure scientific independence of the Panel's work, Editing and Drafting Support personnel report to the Panel on substantive content/language of written outputs, while complying with coordination requirements, timelines, and sharing of produced content as part of the Panel Secretariat. For administrative purposes, they report to United Nations University.

The following United Nations Secretariat entities also supported the Panel Secretariat in the preparation of the report: Department for General Assembly and Conference Management, Department of Global Communications, and United Nations Geospatial (Office of Information and Communications Technology).