AI ACTION SUMMIT

Международный научный доклад о безопасности развитого ИИ

Международный научный доклад, посвященный безопасности развитого искусственного интеллекта

Январь 2025 г.

Участники

ПРЕДСЕДАТЕЛЬ

Проф. Yoshua Bengio, Университет Монреаля / Мила — Квебекский институт искусственного интеллекта

ЭКСПЕРТНАЯ КОНСУЛЬТАТИВНАЯ ГРУППА

Международная группа была назначена правительствами 30 стран, перечисленных ниже, а также ООН, ЕС и ОЭСР.

Австралия: Bronwyn Fox, Университет Нового Южного Уэльса

Бразилия: André Carlos Ponce de Leon Ferreira de Carvalho, Институт математики и компьютерных наук, Университет Сан-Паулу

Канада: Mona Nemer, главный научный советник Канады

Чили: Raquel Pezoa Rivera, Технический университет имени Федерико Санта-Мария

Китай: Yi Zeng, Китайская академия наук

Европейский Союз: Juha Heikkilä, Европейское бюро искусственного интеллекта

Франция: Guillaume Avrin, Национальный координационный центр по искусственному интеллекту

Германия: Antonio Krüger, Немецкий исследовательский центр искусственного интеллекта

Индия: Balaraman Ravindran, Школа анализа данных и искусственного интеллекта Wadhwani, Индийский технологический институт Мадраса

Индонезия: Hammam Riza, Совместные исследования и промышленные инновации в области искусственного интеллекта (Collaborative Research and Industrial Innovation in Artificial Intelligence, KORIKA)

Ирландия: Ciarán Seoighe, Research Ireland

Израиль: Ziv Katzir, Израильское управление инноваций

Италия: Andrea Monti,, юридический эксперт заместителя государственного секретаря по цифровой трансформации, председатель Совета министров Италии

Япония: Hiroaki Kitano, Sony Group Corporation

Кения: Nusu Mwamanzi, Министерство ИКТ и цифровой экономики

Королевство Саудовская Аравия: Fahad Albalawi, Управление Саудовской Аравии по данным и искусственному интеллекту

Мексика: José Ramón López Portillo, LobsterTel

Нидерланды: Haroon Sheikh, Научный совет Нидерландов по государственной политике

Новая Зеландия: Gill Jolly,, Министерство бизнеса, инноваций и занятости

Нигерия: Olubunmi Ajala, Министерство коммуникаций, инноваций и цифровой экономики

ОЭСР: Jerry Sheehan, директор Управления по науке, технологиям и инновациям

Филиппины: Dominic Vincent Ligot, CirroLytix

Республика Корея: Kyoung Mu Lee, кафедра электротехники и вычислительной техники, Сеульский национальный университет

Руанда: Crystal Rugege, Центр четвертой промышленной революции

Сингапур: Denise Wong, Группа по инновациям и защите данных, Управление развития инфокоммуникационных медиа

Испания: Nuria Oliver, ЭЛЛИС Аликанте

Швейцария: Christian Busch, Федеральный департамент экономики, образования и исследований

Турция: Ahmet Halit Hatip, Министерство промышленности и технологий Турции

Украина: Oleksii Molchanovskyi, Экспертный комитет по развитию искусственного интеллекта в Украине

Объединенные Арабские Эмираты: Marwan Alserkal, Министерство по делам кабинета министров, Канцелярия премьер-министра

Соединенное Королевство: Chris Johnson, главный научный консультант Департамента науки, инноваций и технологий

Организация Объединенных Наций: Amandeep Singh Gill, заместитель Генерального секретаря по цифровым и новым технологиям и посланник Генерального секретаря по технологиям

Соединенные Штаты: Saif M. Khan, Министерство торговли США

НАУЧНЫЙ РУКОВОДИТЕЛЬ

Sören Mindermann, Мила — Квебекский институт искусственного интеллекта

ВЕДУЩИЙ АВТОР

Daniel Privitera, Центр KIRA

АВТОРСКАЯ ГРУППА

Tamay Besiroglu, Epoch AI

Rishi Bommasani, Стэнфордский университет

Stephen Casper, Массачусетский технологический институт

Yejin Choi, Стэнфордский университет

Philip Fox, Центр KIRA

Ben Garfinkel, Оксфордский университет

Danielle Goldfarb, Мила — Институт искусственного интеллекта в Квебеке

Hoda Heidari, Университет Карнеги — Меллона

Anson Ho, Epoch AI

Sayash Kapoor, Принстонский университет

Leila Khalatbari, Гонконгский университет науки и технологий

Shayne Longpre, Массачусетский технологический институт

Sam Manning, Центр управления ИИ

Vasilios Mavroudis, Институт Алана Тьюринга

Mantas Mazeika, Иллинойсский университет в Урбане-Шампейне

Julian Michael, Нью-Йоркский университет

Jessica Newman, Калифорнийский университет в Беркли

Kwan Yee Ng, Concordia AI

Chinasa T. Okolo, Институт Брукингса

Deborah Raji, Калифорнийский университет в Беркли

Girish Sastry, независимый автор

Elizabeth Seger (писатель широкого профиля), Demos

Theodora Skeadas, Humane Intelligence

Tobin South, Массачусетский технологический институт

Emma Strubell, Университет Карнеги — Меллона

Florian Tramèr, Швейцарская высшая техническая школа Цюриха

Lucia Velasco, Маастрихтский университет

Nicole Wheeler, Университет Бирмингема

СТАРШИЕ КОНСУЛЬТАНТЫ

Daron Acemoglu, Массачусетский технологический институт

Olubayo Adekanmbi, до прихода в EqualyzAI занимал должность старшего консультанта

David Dalrymple, Агентство перспективных исследований и изобретений

Thomas G. Dietterich, Университет штата Орегон

Edward W. Felten, Принстонский университет

Pascale Fung, до прихода в Meta работала старшим консультантом

Pierre-Olivier Gourinchas, Департамент исследований Международного валютного фонда

Fredrik Heintz, Линчепингский университет

Geoffrey Hinton, Университет Торонто

Nick Jennings, Университет Лафборо

Andreas Krause, Швейцарская высшая техническая школа Цюриха

Susan Leavy, Университетский колледж Дублина

Percy Liang, Стэнфордский университет

Teresa Ludermir, Федеральный университет Пернамбуку

Vidushi Marda, AI Collaborative

Helen Margetts, Оксфордский университет

John McDermid, Йоркский университет

Jane Munga, Фонд Карнеги за международный мир

Arvind Narayanan, Принстонский университет

Alondra Nelson, Институт перспективных исследований

Clara Neppel, IEEE

Alice Oh, Школа вычислительной техники KAIST

Gopal Ramchurn, Responsible AI UK

Stuart Russell, Калифорнийский университет в Беркли

Marietje Schaake, Стэнфордский университет

Bernhard Schölkopf, Институт ELLIS в Тюбингене

Dawn Song, Калифорнийский университет в Беркли

Alvaro Soto, Папский католический университет Чили

Lee Tiedrich, Университет Дьюка

Gaël Varoquaux, Inria

Andrew Yao, Институт междисциплинарных информационных наук, Университет Цинхуа

Ya-Qin Zhang, Университет Цинхуа

СЕКРЕТАРИАТ

Институт безопасности ИИ

Baran Acar

Ben Clifford

Lambrini Das

Claire Dennis

Freya Hempleman

Hannah Merchant

Rian Overy

Ben Snodin

Мила — Квебекский институт искусственного интеллекта

Jonathan Barry

Benjamin Prud'homme

БЛАГОДАРНОСТИ

Рецензенты от гражданского общества и отрасли

Гражданское общество: Институт Ады Лавлейс, Форум ИИ Новой Зеландии / Te Kāhui Atamai Iahiko o Aotearoa, Временная группа экспертов по ИИ в Австралии, Фонд Карнеги за международный мир, Центр права и инноваций / Certa Foundation, Центр управления ИИ, Meir Shamgar, судья Верховного суда / Центр цифрового права и инноваций, Eon Institute, Gradient Institute, Израильский институт демократии, Mozilla Foundation, Old Ways New, RAND, SaferAI, The Centre for Long-Term Resilience, The Future Society, Институт Алана Тьюринга, Королевское общество, Ассоциация политики искусственного интеллекта Турции.

Отрасль: Advai, Anthropic, Cohere, Deloitte Consulting, США и Deloitte LLM, Соединенное Королевство, G42, Google DeepMind, Harmony Intelligence, Hugging Face, IBM, Lelapa AI, Meta, Microsoft, Shutterstock, Zhipu.ai.

Особая благодарность

Секретариат выражает признательность за поддержку, комментарии и отзывы Angie Abdilla, Nitarshan Rajkumar, Geoffrey Irving, Shannon Vallor, Rebecca Finlay и Andrew Strait.

Любые запросы относительно данной публикации следует направлять по адресу Secretariat.AIStateofScience@dsit.gov.uk.

Запросы относительно содержания доклада следует также направлять научному руководителю.

Отказ от ответственности

Доклад не отражает точку зрения Председателя, какого-либо конкретного лица в составе авторской или консультативной группы, а также правительств, поддержавших его написание. В данном докладе обобщены существующие исследования возможностей и рисков развитого ИИ. Председатель доклада несет за него полную ответственность и контролировал его разработку от начала до конца.

Номер серии исследований: DSIT 2025/001

Об этом докладе

- **Это первый международный научный доклад о безопасности развитого ИИ.** После публикации промежуточной редакции доклада в мае 2024 года в подготовке полного первого доклада продолжила работу разносторонняя группа из 96 экспертов в области искусственного интеллекта (ИИ), включая международную экспертную консультативную группу, назначенную правительствами 30 стран, Организацией экономического сотрудничества и развития (ОЭСР), Европейским союзом (ЕС) и Организацией Объединенных Наций (ООН). Целью доклада является предоставление научной информации, которая будет способствовать принятию обоснованных политик. Он не рекомендует какие-либо конкретные политики.
- **Доклад является результатом труда независимых экспертов.** Независимые эксперты, работавшие под руководством председателя и совместно подготовившие этот доклад, имели полную свободу действий в отношении его содержания.
- **Хотя в этом докладе рассматриваются риски и безопасность ИИ, он также обеспечивает множество потенциальных преимуществ для людей, бизнеса и общества.** Существует множество типов ИИ, каждый из которых имеет свои преимущества и риски. В большинстве случаев и в большинстве применениях ИИ помогает отдельным людям и организациям работать более эффективно. Однако люди во всем мире смогут в полной мере и безопасно воспользоваться многочисленными потенциальными преимуществами ИИ только в том случае, если будут надлежащим образом управлять рисками, связанными с ним. В настоящем докладе основное внимание уделяется выявлению этих рисков и оценке методов их снижения. Целью исследования не является всеобъемлющая оценка всех возможных социальных влияний ИИ, включая его многочисленные потенциальные преимущества.
- **Основное внимание в докладе уделяется ИИ общего назначения.** В докладе основное внимание уделяется типу ИИ, который особенно быстро развивался в последние годы, и связанные с ним риски меньше изучены и хуже понимаются — ИИ общего назначения или ИИ, который может выполнять широкий спектр задач. Анализ, выполненный в этом докладе, сосредоточен на самых развитых системах ИИ общего назначения на момент написания статьи, а также на будущих системах, которые могут оказаться еще более эффективными.
- **В докладе обобщены научные данные по трем основным вопросам:** Что может делать ИИ общего назначения? Каковы риски, связанные с ИИ общего назначения? И какие существуют техники смягчения последствий этих рисков?
- **Ставки очень высоки.** Мы, эксперты, принявшие участие в подготовке этого доклада, по-прежнему не имеем единого мнения по ряду вопросов, как второстепенных, так и наиболее важных, касающихся возможностей ИИ общего назначения, рисков и мер по снижению рисков. Но мы считаем этот доклад необходим для улучшения нашего коллективного понимания данной технологии и ее потенциальных рисков. Мы надеемся, что этот доклад поможет международному сообществу прийти к большему консенсусу в отношении ИИ общего назначения и более эффективно снижать его риски, чтобы люди могли безопасно пользоваться его многочисленными потенциальными преимуществами. Ставки высоки. Мы с нетерпением ожидаем продолжения нашей совместной работы.

Обновленная информация о последних достижениях в области ИИ после окончания написания настоящего доклада: примечание председателя

В период между завершением написания настоящего доклада (5 декабря 2024 г.) и его публикацией в январе 2025 г. произошло важное событие. Компания OpenAI, занимающаяся разработкой ИИ, поделилась первыми результатами испытаний новой модели ИИ, *o3*. Они указывают на значительно более высокую производительность, чем у любой предыдущей модели, в ряде самых сложных тестов в области программирования, абстрактного мышления и научных умозаключений. В некоторых из этих тестов o3 превосходит многих (но не всех) людей-экспертов. Кроме того, она совершила прорыв в ключевом тесте на абстрактное мышление, что многие эксперты, включая меня, до недавнего времени считали недостижимым. Однако на момент написания доклада общедоступной информации о реальных возможностях модели, особенно в плане решения многовариантных задач, не было.

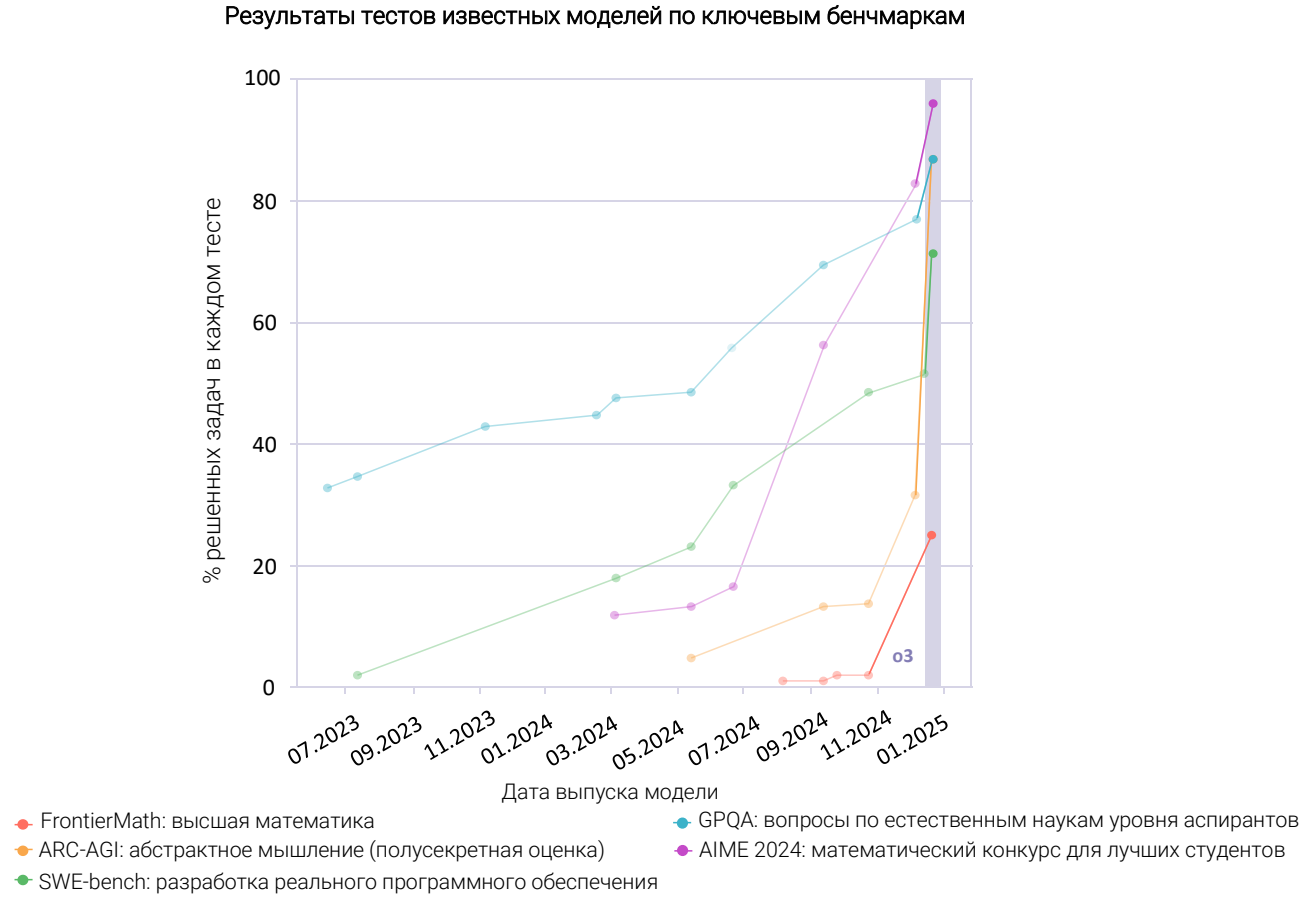

Рисунок 0.1. *Результаты тестов известных моделей ИИ общего назначения с июня 2023 года по декабрь 2024 года. Модель o3 продемонстрировала значительное улучшение производительности по сравнению с предыдущим уровнем (затемненная область). Эти бенчмарки являются одними из самых сложных тестов в области программирования, абстрактного мышления и научных умозаключений. Для невыпущенной модели o3 указана дата анонса; для остальных моделей указана дата выпуска. Некоторые из последних моделей ИИ, включая o3, выиграли от улучшенной кодогенерации и большего количества вычислений во время тестирования. Источники: Anthropic, 2024; Chollet, 2024; Chollet et al., 2025; Epoch AI, 2024; Glazer et al. 2024; OpenAI, 2024a; OpenAI, 2024b; Jimenez et al., 2024; Jimenez et al., 2025.*

Результаты o3 свидетельствуют о том, что темпы развития возможностей ИИ могут оставаться высокими или даже ускоряться. В частности, они предполагают, что предоставление моделям большей вычислительной мощности для решения заданной задачи («масштабирование вывода») может помочь преодолеть предыдущие ограничения. Как правило, масштабирование вывода делает модели более дорогими в использовании. Но, как показала другая известная модель, *R1*, выпущенная компанией DeepSeek в январе 2025 года, исследователи успешно работают над снижением этих затрат. В целом масштабирование вывода может позволить разработчикам ИИ добиться дальнейшего прогресса в будущем. Результаты o3 также подчеркивают необходимость лучшего понимания того, как растущее использование ИИ разработчиками может повлиять на скорость самого дальнейшего развития ИИ.

Выявленные o3 тенденции могут иметь серьезные последствия для рисков, связанных с ИИ. Достижения в науке и возможности программирования ранее уже давали все больше доказательств таких рисков, как кибератаки и биологические атаки. Результаты o3 также применимы для оценки потенциального воздействия на рынок труда, риска потери контроля и, кроме того, потребления энергии. Возможности o3 также могут быть использованы для защиты от неполадок и злонамеренного использования. В целом оценки рисков в этом докладе следует рассматривать с учетом того, что с момента написания доклада возможности ИИ расширились. Однако до сих пор нет никаких доказательств воздействия o3 на реальный мир, а также информации, подтверждающей или исключающей серьезные новые и/или непосредственные риски.

Улучшение возможностей, выявленное на основе результатов o3, и наше ограниченное понимание последствий для рисков ИИ подчеркивают ключевую проблему для разработчиков политики, которую определяет этот доклад: им часто придется взвешивать потенциальные выгоды и риски предстоящих достижений ИИ, не имея в наличии большого объема научных доказательств. Тем не менее, получение данных о последствиях для безопасности и защиты, которые подразумевают тенденции, заложенные в o3, станет первоочередной задачей исследований в области ИИ в ближайшие недели и месяцы.

Основные выводы доклада

- **Возможности ИИ общего назначения, которому посвящен этот доклад, стремительно выросли за последние годы и еще больше развились в последние месяцы.** Несколько лет назад лучшие большие языковые модели (БЯМ) редко могли создать связный абзац текста. Сегодня ИИ общего назначения может писать компьютерные программы, создавать фотореалистичные изображения по запросу пользователя и участвовать в длительных открытых беседах. С момента публикации промежуточного доклада (май 2024 г.) новые модели продемонстрировали заметно более высокую эффективность в тестах на научное мышление и программирование.
- **Многие компании сейчас инвестируют в разработку агентов ИИ общего назначения, рассматривая их как потенциальное направление для дальнейшего развития.** Агенты ИИ — это системы ИИ общего назначения, которые могут действовать автономно, планировать действия и делегировать полномочия для достижения целей практически без человеческого надзора. Например, в отличие от существующих систем, улучшенные агенты ИИ смогут использовать компьютеры для выполнения более длительных проектов, что создаст не только дополнительные преимущества, но и дополнительные риски.
- **Дальнейшее развитие возможностей в ближайшие месяцы и годы может быть как медленным, так и чрезвычайно быстрым.**[†] Прогресс будет зависеть от того, смогут ли компании быстро задействовать еще больше данных и вычислительных мощностей для обучения новых моделей, и позволит ли такое «масштабирование» моделей преодолеть их текущие ограничения. Недавние исследования показали, что быстрое масштабирование моделей может оставаться физически осуществимым по крайней мере в течение нескольких лет. Однако для существенного расширения возможностей могут потребоваться и другие факторы, например новые научные открытия, которые трудно предсказать, или успех нового подхода к масштабированию, недавно введенного компаниями.
- **Некоторые виды вреда от ИИ общего назначения уже хорошо известны.** К ним относятся мошенничество, интимные изображения, полученные без согласия (NCII), и материалы, содержащие сцены сексуального насилия над детьми (CSAM), результаты моделирования, предвзятые по отношению к определенным группам лиц или определенным мнениям, проблемы надежности и нарушения конфиденциальности. Исследователи разработали техники смягчения последствий этих проблем, но до сих пор ни одна комбинация методов не позволяет решить их полностью. После публикации промежуточного доклада (май 2024 г.) появились новые доказательства дискриминации, связанной с системами ИИ общего назначения, свидетельствующие о более тонких формах предвзятости.
- **По мере того как возможности ИИ общего назначения становятся все более широкими, постепенно появляются свидетельства дополнительных рисков.** К ним относятся такие риски, как масштабное влияние на рынок труда, поддерживаемый ИИ взлом или биологические атаки с использованием ИИ, а также потеря обществом контроля над ИИ общего назначения. Эксперты по-разному интерпретируют имеющиеся данные об этих рисках: некоторые полагают, что подобные риски появятся лишь через десятилетия, в то время как другие считают, что ИИ общего назначения может нанести ощутимый вред в масштабах общества уже в ближайшие годы. Недавние достижения в расширении возможностей ИИ общего назначения, особенно в тестах на научное мышление и программирование, дали новые доказательства

[†] Ознакомьтесь с обновленной информацией от Председателя о последних достижениях в области ИИ после написания настоящего доклада.

потенциальных рисков, таких как взлом и биологические атаки с использованием ИИ, что побудило одну крупную компанию в сфере ИИ повысить оценку биологического риска ее лучшей модели с «низкого» до «среднего».

- Методы управления рисками только формируются, но прогресс возможен. Существуют различные технические методы оценки и снижения рисков, связанных с ИИ общего назначения, которыми могут пользоваться разработчики и которые могут использовать регулирующие органы, но все они имеют ограничения. Например, существующие методы интерпретации, позволяющие объяснить, почему модель ИИ общего назначения выдала тот или иной результат, остаются крайне ограниченными. Однако исследователи добились определенного прогресса в устранении этих ограничений. Кроме того, исследователи и разработчики политики все чаще пытаются стандартизировать подходы к управлению рисками и координировать их на международном уровне.
- Темпы и непредсказуемость прогресса в области ИИ общего назначения создают для разработчиков политики «дилемму доказательств».[†] Учитывая порой быстрый и неожиданный прогресс, разработчикам политики часто приходится взвешивать потенциальные выгоды и риски, связанные с предстоящим развитием ИИ, не имея в наличии большого объема научных данных. При этом они сталкиваются с дилеммой. С одной стороны, упреждающие меры по смягчению последствий риска, основанные на ограниченных доказательствах, могут оказаться неэффективными или нецелесообразными. С другой стороны, ожидание более веских доказательств приближающегося риска может лишить общество возможности подготовиться или даже сделает невозможным смягчение последствий, например, в случае резкого скачка возможностей ИИ и связанных с ними рисков. Компании и правительства разрабатывают системы раннего предупреждения и структуры управления рисками, которые могут снизить остроту дилеммы. Некоторые из них инициируют конкретные меры по смягчению последствий при появлении новых доказательств рисков, другие же требуют от разработчиков предоставлять доказательства безопасности перед выпуском новой модели.
- По общему мнению исследователей, были бы полезны достижения в следующих вопросах: Насколько быстро будут развиваться возможности ИИ общего назначения в ближайшие годы и как исследователи могут надежно измерять этот прогресс? Каковы целесообразные пороги риска для принятия мер по смягчению последствий? Как разработчики политики могут наилучшим образом получить доступ к информации об ИИ общего назначения, имеющей отношение к общественной безопасности? Как исследователи, технологические компании и правительства могут достоверно оценить риски разработки и развертывания ИИ общего назначения? Как работают внутренние механизмы моделей ИИ общего назначения? Как можно спроектировать ИИ общего назначения, который будет работать надежно?
- ИИ — это не что-то, данное нам свыше: его развитие определяют люди своими действиями. Будущее технологии ИИ общего назначения неопределенно, и даже в ближайшее время возможны самые разные варианты развития событий, включая как весьма позитивные, так и весьма негативные. Такая неопределенность может создать ощущение фатализма и заставить нас воспринимать ИИ как нечто, развивающееся само по себе. Но именно от решений общества и правительств о том, как преодолеть эту неопределенность, зависит путь, по которому мы пойдем. Цель данного доклада — способствовать конструктивному и основанному на фактах обсуждению этих решений.

[†] Ознакомьтесь с обновленной информацией от Председателя о последних достижениях в области ИИ после написания настоящего доклада.

Краткое резюме

Цель этого доклада

В этом докладе обобщено научное понимание ИИ общего назначения — ИИ, способного решать широкий спектр задач, — уделяя особое внимание к пониманию его рисков и управлению ими.

В этом докладе обобщены научные данные о безопасности ИИ общего назначения. Цель данного доклада — помочь сформировать единое международное понимание рисков, связанных с развитым ИИ, и способов их снижения. Для достижения этой цели в настоящем докладе основное внимание уделяется ИИ общего назначения — ИИ, способному решать широкий спектр задач, — поскольку этот тип ИИ особенно быстро развивался в последние годы и широко используется технологическими компаниями для различных бытовых и деловых целей. В докладе обобщено научное понимание ИИ общего назначения, уделяя особое внимание пониманию его рисков и управлению ими.

На фоне быстрого прогресса ИИ исследования в области ИИ общего назначения в настоящий момент находятся на этапе научных открытий и во многих случаях еще не дали окончательно сформировавшихся научных знаний. В докладе представлен краткий обзор текущего научного понимания ИИ общего назначения и связанных с ним рисков. Это включает в себя выявление областей научного консенсуса и областей, где существуют различные мнения или пробелы в современных научных знаниях.

Люди во всем мире смогут в полной мере и безопасно воспользоваться потенциальными преимуществами ИИ общего назначения только в том случае, если его риски будут надлежащим образом управляться. В настоящем докладе основное внимание уделяется выявлению этих рисков и технических методов их оценки и смягчения последствий, включая способы, подразумевающие смягчение последствий рисков с помощью самого ИИ общего назначения. Что очень важно, доклад не ставит своей целью всестороннюю оценку всех возможных социальных влияний ИИ общего назначения — текущие и потенциальные будущие выгоды ИИ общего назначения, хотя они и обширны, выходят за рамки данного доклада. Для разработки комплексной политики необходимо учитывать как потенциальные преимущества ИИ общего назначения, так и риски, рассматриваемые в этом докладе. Также необходимо учитывать, что другие типы ИИ имеют профили риска/пользы, которые отличаются от профилей риска/пользы для существующего ИИ общего назначения.

В трех основных разделах доклада обобщены научные данные по трем ключевым вопросам: Что может делать ИИ общего назначения? Каковы риски, связанные с ИИ общего назначения? Какие существуют техники смягчения последствий этих рисков?

Раздел 1. Возможности ИИ общего назначения: что может делать ИИ общего назначения сейчас и на что он будет способен в будущем?

Возможности ИИ общего назначения стремительно развивались в последние годы, и дальнейшее развитие может быть как медленным, так и чрезвычайно быстрым.

Возможности ИИ являются ключевым фактором многих рисков, которые он создает, и, согласно многим показателям, возможности ИИ общего назначения быстро развиваются. Пять лет назад ведущие языковые модели ИИ общего назначения редко могли создать связный абзац текста. Сегодня некоторые модели ИИ общего назначения могут участвовать в беседах на самые разные темы, писать компьютерные программы или создавать реалистичные короткие видеоролики на основе описания. Однако достоверно оценить и описать возможности ИИ общего назначения технически сложно.

В последние годы разработчики ИИ быстро улучшали возможности ИИ общего назначения, в основном за счет «масштабирования».[†] Они постоянно увеличивают ресурсы, используемые для обучения новых моделей (это часто называют «масштабированием»), и совершенствуют существующие подходы для более эффективного использования этих ресурсов. Например, согласно последним оценкам, современные модели ИИ демонстрируют ежегодный рост вычислительных ресурсов, используемых для обучения, примерно в 4 раза, а размер обучающих наборов данных — в 2,5 раза.

Темпы развития возможностей ИИ общего назначения будут иметь существенные последствия для управления возникающими рисками, но эксперты расходятся во мнениях относительно того, чего ожидать в ближайшие месяцы и годы. Эксперты по-разному оценивают возможность медленного, быстрого или чрезвычайно быстрого развития возможностей ИИ общего назначения.

Эксперты расходятся во мнениях относительно темпов будущего прогресса из-за разных взглядов на перспективы дальнейшего «масштабирования», а компании изучают дополнительный, новый тип масштабирования, который может еще больше ускорить развитие возможностей. Хотя масштабирование часто позволяло преодолеть ограничения предшествующих систем, эксперты расходятся во мнениях относительно его потенциала в преодолении имеющихся ограничений современных систем, таких как ненадежность при работе в физическом мире и при решении сложных задач на компьютерах. В последние месяцы новый тип масштабирования продемонстрировал потенциал для дальнейшего расширения возможностей. Вместо того чтобы просто масштабировать ресурсы, используемые для обучения моделей, компании, занимающиеся ИИ, все больше интересуются «масштабированием логического вывода» — позволяя уже обученной модели использовать больше вычислительных мощностей для решения заданной задачи, например, для улучшения ее собственного решения или для написания так называемых «цепочек рассуждений», которые разбивают задачу на более простые шаги.

[†] Ознакомьтесь с обновленной информацией от Председателя о последних достижениях в области ИИ после написания настоящего доклада.

Несколько ведущих компаний, разрабатывающих ИИ общего назначения, делают ставку на «масштабирование», чтобы повышать производительность и далее. Если сохранятся последние тенденции, то к концу 2026 года некоторые модели ИИ общего назначения будут обучаться с использованием примерно в 100 раз большего объема тренировочных вычислительных ресурсов, чем самые ресурсоемкие модели 2023 года, а к 2030 году этот показатель вырастет до 10 000 раз в сочетании с алгоритмами, которые обеспечивают более высокие возможности для заданного объема доступных вычислений. Помимо этого потенциального масштабирования обучающих ресурсов, недавние тенденции, такие как масштабирование логических выводов и использование моделей для генерации обучающих данных, могут означать, что в целом будет использоваться еще больше вычислительных ресурсов. Однако существуют потенциальные препятствия для дальнейшего быстрого увеличения объемов данных и вычислительных ресурсов, такие как доступность данных, микросхем ИИ, капитала и местных энергетических мощностей. Компании, разрабатывающие ИИ общего назначения, работают над устранением этих потенциальных препятствий.

С момента публикации промежуточного отчета (май 2024 г.) ИИ общего назначения достиг экспертного уровня в некоторых тестах и соревнованиях по научному мышлению и программированию, и компании прилагают большие усилия для разработки автономных агентов ИИ. Достижения в области науки и программирования были обусловлены методами масштабирования выводов, такими как написание длинных «цепочек рассуждений». Новые исследования показывают, что дальнейшее масштабирование таких подходов, например предоставление моделям возможности анализировать задачи путем составления еще более длинных цепочек рассуждений, чем современные модели, может привести к дальнейшему прогрессу в областях, где рассуждения имеют большее значение, таких как наука, разработка программного обеспечения и планирование. В дополнение к этой тенденции компании прилагают большие усилия для разработки более совершенных агентов ИИ общего назначения, которые могут планировать и действовать автономно для достижения поставленной цели. Наконец, рыночная стоимость использования ИИ общего назначения заданного уровня возможностей резко упала, что делает эту технологию более широкодоступной и широко используемой.

В настоящем докладе основное внимание уделяется техническим аспектам развития ИИ, однако то, насколько быстро будет развиваться ИИ общего назначения, — это не только технический вопрос. Темпы будущего прогресса также будут зависеть от нетехнических факторов, включая, возможно, подходы правительств к регулированию ИИ. В этом докладе не обсуждается, как различные подходы к регулированию могут повлиять на скорость разработки и внедрения ИИ общего назначения.

Раздел 2. Риски: каковы риски, связанные с ИИ общего назначения?

Некоторые виды вреда от ИИ общего назначения уже хорошо известны. По мере того как возможности ИИ общего назначения становятся все более широкими, постепенно появляются свидетельства дополнительных рисков.

В этом докладе риски ИИ общего назначения разделены на три категории: риски злонамеренного использования, риски из-за неисправностей и системные риски. Каждая из этих категорий содержит риски, которые уже материализовались, а также риски, которые могут материализовались в ближайшие несколько лет.

Риски из-за злонамеренного использования: злоумышленники могут использовать ИИ общего назначения для причинения вреда отдельным лицам, организациям или обществу. Формы злонамеренного использования включают в себя:

- Нанесение вреда отдельным лицам посредством фальшивого контента. В настоящее время злоумышленники могут использовать ИИ общего назначения для создания фальшивого контента, который целенаправленно вредит отдельным лицам. К таким видам вредоносного использования относятся создаваемая без согласия дипфейк-порнография и CSAM-контент, сгенерированный ИИ, финансовое мошенничество с использованием имитации голоса, шантаж с целью вымогательства, подрыв личной и профессиональной репутации, а также психологическое насилие. Однако сообщения о происшествиях, связанных с нанесением вреда фальшивым контентом, сгенерированным ИИ, встречаются часто, однако достоверная статистика о частоте таких происшествий по-прежнему отсутствует.
- Манипулирование общественным мнением. ИИ общего назначения упрощает создание убедительного контента в больших масштабах. Этим могут воспользоваться те, кто стремится манипулировать общественным мнением, например, чтобы повлиять на политические результаты. Однако данные о том, насколько распространены и эффективны такие усилия, остаются ограниченными. Технические контрмеры, такие как наложение водяных знаков на контент, хотя и полезны, обычно могут быть обойдены злоумышленниками средней квалификации.
- Киберпреступления. ИИ общего назначения может облегчить или ускорить проведение кибератак злоумышленниками разного уровня квалификации. Современные системы продемонстрировали свои возможности в задачах кибербезопасности низкой и средней сложности, а финансируемые государством организации активно изучают возможности ИИ для анализа целевых систем. Новые исследования подтвердили, что возможности ИИ общего назначения, связанные с киберпреступностью, значительно расширяются, но пока неясно, повлияет ли это на баланс между хакерами и защитниками.
- Биологические и химические атаки. Современные системы ИИ общего назначения продемонстрировали определенную способность предоставлять инструкции и рекомендации по устранению проблем при воспроизведении известных видов биологического и химического оружия, а также облегчать разработку новых токсичных соединений. В новых экспериментах, в которых проверялась способность генерировать

планы по производству биологического оружия, система ИИ общего назначения иногда показывала лучшие результаты, чем эксперты-люди, имеющие доступ к Интернету. В ответ на это одна компания, занимающаяся разработкой ИИ, повысила оценку биологического риска для своей лучшей модели с «низкого» до «среднего». Однако реальные попытки разработать такое оружие по-прежнему требуют значительных дополнительных ресурсов и опыта. Всеобъемлющая оценка биологического и химического риска затруднена, поскольку большая часть соответствующих исследований засекречена.

С момента публикации промежуточного доклада (май 2024 г.) возможности ИИ общего назначения в областях, подверженных злонамеренному использованию, стали более широкими. Например, недавно исследователи создали системы ИИ общего назначения, которые смогли самостоятельно найти и использовать некоторые уязвимости в системе кибербезопасности, а также при помощи человека обнаружить ранее неизвестную уязвимость в широко используемом программном обеспечении. Также расширились возможности ИИ общего назначения, связанные с рассуждениями и интеграцией различных типов данных, которые могут помочь в исследованиях патогенов или в других областях двойного назначения.

Риски из-за неисправностей. ИИ общего назначения также может причинить непреднамеренный вред. Даже если пользователи не намерены причинять вред, могут возникнуть серьезные риски из-за неисправностей ИИ общего назначения. К таким неисправностям относятся:

- Проблемы надежности. Современный ИИ общего назначения может быть ненадежным, что может привести к причинению вреда. Например, если пользователи обращаются к системе ИИ общего назначения за медицинской или юридической консультацией, система может выдать ответ, содержащий ложные сведения. Пользователи часто не знают об ограничениях продукта на основе ИИ, например, из-за ограниченной «грамотности в области ИИ», вводящей в заблуждение рекламы или недопонимания. Известно несколько случаев причинения вреда из-за проблем надежности, но по-прежнему недостаточно данных о том, насколько широко распространены различные формы этой проблемы.
- Предвзятость. Системы ИИ общего назначения могут усиливать социальную и политическую предвзятость, нанося конкретный вред. Они часто проявляют предвзятость по отношению к расе, полу, культуре, возрасту, инвалидности, политическим взглядам или другим аспектам человеческой идентичности. Это может привести к дискриминационным результатам, включая неравное распределение ресурсов, усиление стереотипов и систематическое игнорирование недостаточно представленных групп или точек зрения. Технические подходы к снижению предвзятости и дискриминации в системах ИИ общего назначения развиваются, но при этом приходится идти на компромиссы между снижением предвзятости и конкурирующими целями, такими как точность и конфиденциальность, а также решать другие проблемы.
- Потеря контроля. Сценарии «потери контроля» — это гипотетические будущие сценарии, в которых одна или несколько систем ИИ общего назначения выходят из-под чьего-либо контроля, и нет четкого пути к восстановлению контроля. По общему мнению, существующий ИИ общего назначения не обладает достаточными возможностями, чтобы представлять подобный риск. Однако мнения экспертов о вероятности потери контроля в течение следующих нескольких лет существенно разнятся, некоторые считают это неправдоподобным, некоторые полагают, что это может произойти, а некоторые

рассматривают это как умеренно вероятный риск, требующий внимания из-за его высокой потенциальной степени серьезности. Продолжающиеся эмпирические и математические исследования дают новый материал для этих дискуссий.

После публикации промежуточного доклада (май 2024 г.) новые исследования позволили получить некоторые новые сведения относительно рисков предвзятости и потери контроля. Число доказательств предвзятости в системах ИИ общего назначения становится все больше, и в недавних исследованиях были обнаружены дополнительные формы предвзятости ИИ. Исследователи отмечают небольшой дальнейший прогресс в развитии возможностей ИИ, которые, вероятно, необходимы для реализации широко обсуждаемых сценариев потери контроля. К ним относятся возможности автономного использования компьютеров, программирования, получения несанкционированного доступа к цифровым системам и выявления способов обхода человеческого надзора.

Системные риски: помимо рисков, напрямую связанных с возможностями отдельных моделей, широкое внедрение ИИ общего назначения сопряжено с рядом более широких системных рисков. Примеры системных рисков варьируются от потенциального воздействия на рынок труда до рисков для конфиденциальности и воздействия на окружающую среду:

- Риски для рынка труда. ИИ общего назначения, особенно если он продолжит быстро развиваться, позволит автоматизировать очень широкий спектр задач, что может оказать существенное влияние на рынок труда. Это означает, что многие люди могут потерять работу. Однако многие экономисты ожидают, что потенциальное сокращение числа рабочих мест может быть частично или даже полностью компенсировано за счет создания новых рабочих мест и повышения спроса в неавтоматизированных секторах.
- Глобальный разрыв в исследованиях и разработках в области ИИ. Исследования и разработки (НИОКР) в области ИИ общего назначения в настоящее время сосредоточены в нескольких западных странах и Китае. Этот «разрыв в области ИИ» может привести к усилению зависимости всего мира от этой небольшой группы стран. Некоторые эксперты также ожидают, что это будет способствовать росту глобального неравенства. У этого разрыва много причин, включая те, которые не являются характерными только для сферы ИИ. Однако в значительной степени это обусловлено разным уровнем доступа к весьма дорогостоящим вычислительным ресурсам, необходимым для разработки ИИ общего назначения. Большинство стран с низким и средним уровнем дохода (LMICs) имеют значительно меньший доступ к вычислительным ресурсам, чем страны с высоким уровнем дохода (HICs).
- Концентрация рынка и единые точки отказа. В настоящее время на рынке ИИ общего назначения доминирует небольшое количество компаний. Такая концентрация рынка может сделать общество более уязвимым к ряду системных рисков. Например, если организации в таких критически важных секторах, как финансы или здравоохранение, будут полагаться на небольшое количество систем ИИ общего назначения, то ошибка или уязвимость в одной из таких систем может привести к одновременным сбоям и нарушениям в работе в широком масштабе.
- Экологические риски. Рост использования вычислительных мощностей при разработке и развертывании систем ИИ общего назначения быстро привел к увеличению объемов энергии, воды и сырья, потребляемых при создании и эксплуатации необходимой

вычислительной инфраструктуры. Эта тенденция не демонстрирует явных признаков замедления, несмотря на прогресс в технологиях, позволяющих более эффективно использовать вычислительные ресурсы. ИИ общего назначения также имеет ряд применений, которые могут как принести пользу, так и нанести вред усилиям по обеспечению устойчивого развития.

- Риски нарушения конфиденциальности. ИИ общего назначения может стать причиной или способствовать нарушению конфиденциальности пользователей. Например, конфиденциальная информация, содержащаяся в обучающих данных, может непреднамеренно стать доступной другим лицам, когда пользователь взаимодействует с системой. Кроме того, когда пользователи делятся с системой конфиденциальной информацией, эта информация также может попасть к другим лицам. Однако ИИ общего назначения может также способствовать преднамеренному нарушению конфиденциальности, например, когда злоумышленники используют ИИ для получения конфиденциальной информации о конкретных лицах из больших объемов данных. Однако до сих пор исследователи не обнаружили доказательств широкомасштабных нарушений конфиденциальности, связанных с ИИ общего назначения.
- Нарушения авторских прав. ИИ общего назначения как учится, так и создает произведения творческого самовыражения, бросая вызов традиционным системам получения согласия на использование данных, компенсации и контроля. Сбор данных и создание контента могут быть связаны с различными законами о правах на данные, которые отличаются в разных юрисдикциях, и могут стать предметом активных судебных разбирательств. Учитывая правовую неопределенность в отношении методов сбора данных, компании, занимающиеся разработкой ИИ, предоставляют меньше информации об используемых ими данных. Эта непрозрачность затрудняет проведение сторонних исследований безопасности ИИ.

После публикации промежуточного доклада (май 2024 г.) появились дополнительные данные о влиянии ИИ общего назначения на рынок труда, в то время как новые разработки усилили обеспокоенность по поводу конфиденциальности и авторских прав. Новый анализ данных о рынке труда показывает, что люди гораздо быстрее осваивают ИИ общего назначения по сравнению с предыдущими технологиями. Темпы внедрения новых технологий компаниями существенно различаются в зависимости от отрасли. Кроме того, недавнее расширение возможностей привело к тому, что ИИ общего назначения все чаще применяется в таких деликатных сферах, как здравоохранение и мониторинг на рабочем месте, что создает новые риски для конфиденциальности. Наконец, по мере того как споры об авторских правах обостряются, а технические меры по защите от нарушений авторских прав остаются ненадежными, правообладатели данных стремительно ограничивают доступ к своим данным.

Модели с открытыми весами. Важным фактором при оценке многих рисков, которые может представлять собой модель ИИ общего назначения, является то, как она будет представлена общественности. Так называемые «модели с открытыми весами» — это модели ИИ, центральные компоненты которых, называемые «весами», доступны для скачивания. Открытый доступ к весам способствует исследованиям и инновациям, в том числе в области безопасности ИИ, а также повышает прозрачность и позволяет научному сообществу проще обнаруживать дефекты в моделях. Однако модели с открытым весом также могут представлять риски, например, способствуя злонамеренному или ошибочному использованию, которое разработчику модели

сложно или невозможно отследить или смягчить. После того как веса модели становятся доступны для свободного скачивания, пропадает возможность массового отзыва (возврата) всех существующих копий модели и гарантия возможности получения обновлений безопасности всем пользователям модели. После выхода промежуточного доклада (май 2024 г.) на высоком уровне сложился консенсус относительно того, что риски, связанные с большей открытостью ИИ, следует оценивать с точки зрения «дополнительного» риска — степени, в которой выпуск модели с открытым весом увеличит или уменьшит определенный риск по сравнению с рисками, создаваемыми существующими альтернативами, такими как закрытые модели или другие технологии.

Раздел 3. Управление рисками: какие существуют методы управления рисками, связанными с ИИ общего назначения?

Несколько технических подходов способны улучшить управление рисками, но во многих случаях лучшие из имеющихся подходов по-прежнему имеют существенные ограничения и не обеспечивают количественной оценки рисков или гарантий, существующих в других критически важных для безопасности областях.

Управление рисками — выявление и оценка рисков, а затем их смягчение и мониторинг — является сложной задачей в контексте ИИ общего назначения. Хотя управление рисками является весьма сложной задачей и во многих других областях, есть некоторые особенности ИИ общего назначения, которые создают особые трудности.

Некоторые технические особенности ИИ общего назначения делают управление рисками в этой области особенно сложным. Среди них, в частности, следующие:

- **Спектр возможных применений и контекстов использования систем ИИ общего назначения необычайно широк.** Например, одну и ту же систему можно использовать для медицинских консультаций, анализа компьютерного кода на предмет уязвимостей и создания фотографий. Это усложняет всестороннее прогнозирование возможных вариантов использования, выявление рисков и тестирование работы систем в соответствующих реальных условиях.
- **Разработчики все еще мало понимают, как работают ими созданные модели ИИ общего назначения.** Такое непонимание затрудняет как прогнозирование поведенческих проблем, так и объяснение и решение известных проблем после их обнаружения. Понимание остается труднодостижимым главным образом потому, что модели ИИ общего назначения не программируются в традиционном смысле. Вместо этого их обучают: разработчики ИИ организуют процесс обучения на большом объеме данных, и в результате этого процесса обучения создаётся модель ИИ общего назначения. Внутреннее устройство этих моделей в значительной степени непостижимо, в том числе и для их разработчиков. Методы объяснения и «интерпретируемости» моделей способны улучшить понимание исследователями и разработчиками того, как работают модели ИИ общего назначения, но, несмотря на прогресс в последнее время, эти исследования все еще находятся на начальной стадии.

- **Все более эффективные агенты ИИ — системы ИИ общего назначения, способные действовать автономно, планировать действия и делегировать полномочия для достижения целей. Вероятно, они создадут новые серьезные проблемы в управлении рисками.** Агенты ИИ обычно работают над достижением целей автономно, используя распространенное программное обеспечение, такое как веб-браузеры и инструменты программирования. В настоящее время большинство из них еще недостаточно надежны для широкого использования, но компании прилагают большие усилия для создания более эффективных и надежных агентов ИИ и за последние месяцы добились определенных успехов. Агенты ИИ, вероятно, будут становиться все более полезными, но также могут обострить ряд рисков, обсуждаемых в этом докладе, и создать дополнительные трудности для управления рисками. Примеры таких потенциальных новых проблем включают возможность того, что пользователи не всегда будут знать, что делают их собственные агенты ИИ, возможность того, что агенты ИИ будут действовать без чьего-либо контроля, возможность того, что хакеры «взломают» агентов, а также возможность того, что взаимодействие между разными ИИ создаст новые сложные риски. Подходы к управлению рисками, связанными с агентами, только начинают разрабатываться.

Помимо технических факторов, управление рисками в области ИИ общего назначения затруднено рядом экономических, политических и других социальных факторов.

- **Темпы развития ИИ общего назначения создают «дилемму доказательств» для лиц, принимающих решения.**[†] Быстрое развитие возможностей приводит к тому, что некоторые риски возрастают скачкообразно; например, риск подделки аттестационных работ в сфере образования с использованием ИИ общего назначения за год превратился из незначительного в широко распространенный. Чем быстрее возникает риск, тем сложнее управлять им реактивно и тем ценнее становится подготовка. Однако до тех пор, пока доказательства риска остаются неполными, лица, принимающие решения, также не могут знать наверняка, возникнет ли риск или, возможно, уже возник. Это создает дилемму — принятие упреждающих или ранних мер по смягчению последствий может оказаться ненужным, однако ожидание убедительных доказательств может сделать общество уязвимым к быстро возникающим рискам. Компании и правительства разрабатывают системы раннего предупреждения и структуры управления рисками, которые возможно снизят остроту дилеммы. Некоторые из них инициируют конкретные меры по смягчению последствий при появлении новых доказательств рисков, другие же требуют от разработчиков предоставлять доказательства безопасности перед выпуском новой модели.
- **Существует информационный разрыв между тем, что компании, занимающиеся ИИ, знают о своих системах ИИ, и тем, что знают о них правительства и исследователи, не связанные с отраслью.** Компании часто делятся лишь ограниченной информацией о своих системах ИИ общего назначения, особенно в период до их широкого выпуска. В качестве причин ограничения разглашения информации компании приводят как коммерческие соображения, так и соображения безопасности. Однако этот информационный пробел также затрудняет эффективное участие других субъектов в управлении рисками, особенно в случае возникновения новых рисков.

[†] Ознакомьтесь с обновленной информацией от Председателя о последних достижениях в области ИИ после написания настоящего доклада.

- Как компании, работающие в сфере ИИ, так и правительства часто сталкиваются с сильным конкурентным давлением, которое может привести к тому, что они перестанут уделять должное внимание управлению рисками. При некоторых обстоятельствах конкурентное давление может побудить компании вкладывать в управление рисками меньше времени и других ресурсов, чем они могли бы в противном случае. Аналогичным образом, правительства могут вкладывать меньше средств в политику поддержки управления рисками в случаях, когда стоят перед выбором между успехом в международной конкуренции и снижением рисков.

Тем не менее, существуют различные методы и системы управления рисками, сопряженными с ИИ общего назначения, которые доступны компаниям и которые могут требовать регулирующие органы. К ним относятся методы выявления и оценки рисков, а также методы смягчения их последствий и мониторинга.

- Оценка рисков систем ИИ общего назначения является неотъемлемой частью управления рисками, но существующие оценки рисков сильно ограничены. Существующие оценки рисков ИИ общего назначения главным образом основаны на «выборочных проверках», т. е. тестировании поведения ИИ общего назначения в ряде конкретных ситуаций. Это помогает выявить потенциальные опасности до развертывания модели. Однако существующие тесты часто не учитывают опасности и переоценивают или недооценивают возможности и риски ИИ общего назначения, поскольку условия тестирования отличаются от реальных.
- Для эффективного выявления и оценки рисков нужны специалисты по оценке рисков со значительным опытом, ресурсы и достаточный доступ к необходимой информации. Строгая оценка рисков в контексте ИИ общего назначения требует сочетания нескольких подходов к оценке. Они варьируются от технического анализа самих моделей и систем до оценки возможных рисков, связанных с определенными моделями использования. Чтобы правильно провести такую оценку, специалистам по оценке требуется значительный опыт. Для всесторонней оценки рисков им часто требуется больше времени, прямой доступ к моделям и их обучающим данным, а также больше информации об используемых технических методах, чем обычно предоставляют компании, разрабатывающие ИИ общего назначения.
- Достигнут определенный прогресс в обучении моделей ИИ общего назначения более безопасному функционированию, однако ни один из существующих методов не позволяет надежно исключить даже очевидно небезопасные результаты. Например, метод, называемый «состязательным обучением», заключается в преднамеренном предъявлении моделям ИИ специально спроектированных примеров, которые заставляют их терпеть неудачу или вести себя ненадлежащим образом во время обучения, чтобы выработать устойчивость к таким случаям. Однако злоумышленники по-прежнему могут найти новые способы («атаки»), чтобы обойти эти средства защиты с минимальными или умеренными усилиями. Кроме того, последние данные свидетельствуют о том, что современные методы обучения, которые в значительной степени опираются на несовершенную обратную связь от человека, могут непреднамеренно побуждать модели вводить людей в заблуждение при ответах на сложные вопросы, затрудняя обнаружение ошибок. Одной из возможностей прогресса является увеличение объема и повышение качества такой обратной связи, хотя

перспективными являются и зарождающиеся методы обучения с использованием ИИ для выявления вводящего в заблуждение поведения.

- Мониторинг (выявление рисков и оценка эффективности после начала использования модели), а также различные вмешательства для предотвращения вредоносных действий могут повысить безопасность ИИ общего назначения после его развертывания среди пользователей. Современные инструменты способны обнаруживать контент, созданный ИИ, отслеживать производительность системы и выявлять потенциально опасные входные/выходные данные, хотя пользователи средней квалификации часто могут обойти эти меры безопасности. Несколько уровней защиты, сочетающих технические средства мониторинга и возможности человеческого надзора, повышают безопасность, но могут привести к дополнительным затратам и задержкам. В будущем механизмы с аппаратной реализацией могли бы помочь заказчикам и регулирующим органам более эффективно контролировать системы ИИ общего назначения в процессе развертывания и, возможно, также отслеживать соблюдение соглашений между странами, но надежных механизмов такого рода пока не существует.
- Существует множество методов защиты конфиденциальности, применимых на протяжении всего жизненного цикла ИИ. К ним относятся удаление конфиденциальной информации из обучающих данных, подходы к обучению моделей, контролирующие объем информации, извлекаемой из данных (например, подходы «дифференциальной приватности»), и методы использования ИИ с сенситивной информацией, которые затрудняют восстановление данных (например, «конфиденциальные вычисления» и другие технологии, повышающие конфиденциальность). Многие методы повышения конфиденциальности из других областей исследований пока неприменимы к системам ИИ общего назначения из-за их требований к вычислительным ресурсам. В последние месяцы методы защиты конфиденциальности были расширены с учетом растущего использования ИИ в конфиденциальных областях, включая помощников на смартфонах, агентов ИИ, голосовых помощников, работающих в режиме постоянного восприятия речи, а также использования в здравоохранении или юридической практике.

С момента публикации промежуточного доклада (май 2024 г.) исследователи добились определенного прогресса в попытках объяснить, почему модель ИИ общего назначения выдает тот или иной результат. Возможность объяснить решения ИИ может помочь управлять рисками, связанными с неисправностями, начиная от предвзятости и фактической неточности и заканчивая потерей контроля. Кроме того, во всем мире предпринимаются все более активные усилия по стандартизации подходов к оценке и смягчению последствий.

Заключение: в будущем возможны самые разные траектории развития ИИ общего назначения, и многое будет зависеть от действий общества и правительств.

Будущее ИИ общего назначения неопределенно, и даже в ближайшем будущем возможен широкий спектр траекторий, включающий как очень позитивные, так и очень негативные результаты. Но ничто в будущем ИИ общего назначения не является неизбежным. Как и кем разрабатывается ИИ общего назначения, для решения каких задач он предназначен, сможет ли общество реализовать полный экономический потенциал ИИ общего назначения, кому он выгоден, каким рискам мы себя подвергаем и сколько мы инвестируем в исследования по управлению рисками — ответы на эти и многие другие вопросы зависят от того, какой выбор делают общества и правительства сегодня и будут делать в будущем, определяя развитие ИИ общего назначения.

Чтобы способствовать конструктивному обсуждению этих решений, в данном докладе представлен обзор состояния научных исследований и дискуссий по управлению рисками на момент создания документа, связанными с ИИ общего назначения. Ставки очень высоки. Мы с нетерпением ожидаем продолжения нашей совместной работы.