Исполнительное резюме

О данном отчете

- Это промежуточная публикация первого «Международного научного отчета о безопасности продвинутого ИИ». В работе над этим отчетом приняли участие 75 разных специалистов в области искусственного интеллекта (ИИ), в том числе члены международной Консультативной группы экспертов, номинированных 30 странами, Европейским Союзом (ЕС) и Организацией Объединенных Наций (ООН).
- Коллектив независимых экспертов, занимающихся составлением данного отчета под руководством председателя, имел полный контроль над его содержанием.
- Данная публикация, подготовленная на фоне беспрецедентного прогресса в развитии ИИ, ограничивается одним из видов ИИ, который особенно быстро развивался в последние годы: ИИ общего назначения, или ИИ, способный выполнять широкий спектр задач. Быстро продвигающиеся исследования в области ИИ общего назначения в настоящее время находятся в стадии научных открытий и пока не являются устоявшейся наукой.
- Люди во всем мире смогут безопасно пользоваться многочисленными потенциальными преимуществами ИИ общего назначения только при условии надлежащего управления связанными с ним рисками. Данный отчет посвящен определению этих рисков и оценке технических методов их оценки и снижения. В его задачи не входит всесторонняя оценка всех возможных последствий ИИ общего назначения для общества, включая его многочисленные потенциальные преимущества.
- Впервые в истории в работе над этим промежуточным отчетом приняли участие эксперты, номинированные 30 странами, ЕС и ООН, а также другие ведущие мировые специалисты, чтобы создать общую научную доказательную базу для обсуждения и принятия решений о безопасности ИИ общего назначения. Мы по-прежнему расходимся во мнениях по ряду незначительных и значительных вопросов, касающихся возможностей ИИ общего назначения, рисков и мер по их снижению. Однако мы считаем этот проект важным для улучшения коллективного понимания этой технологии и потенциальных рисков, связанных с ней, а также для приближения к достижению консенсуса и эффективного снижения рисков, что позволит людям безопасно пользоваться потенциальными преимуществами ИИ общего назначения. Ставки высоки. Мы с нетерпением ждем продолжения этой работы.

Основные моменты исполнительного резюме

- При условии надлежащего управления ИИ общего назначения можно применять в интересах общества, что может привести к повышению благосостояния, дальнейшему процветанию и новым научным открытиям. Однако в случае неправильной работы или злонамеренного использования ИИ общего назначения также может нанести вред, например, в результате принятия необъективных

решений в ситуациях, где ставки слишком высоки, или мошенничества, использования «фейковых» СМИ или нарушения конфиденциальности.

- По мере развития возможностей ИИ общего назначения могут возникнуть такие риски, как масштабное воздействие на рынок труда, хакерские или биологические атаки с использованием ИИ, а также потеря обществом контроля над ИИ общего назначения, хотя исследователи ставят под вопрос вероятность этих сценариев. Разные взгляды на эти риски часто обусловлены различиями в ожиданиях относительно мер, которые общество предпримет для их ограничения, эффективности этих мер и темпов развития возможностей ИИ общего назначения.
- Существует значительная неопределенность в отношении темпов будущего прогресса в развитии возможностей ИИ общего назначения. Некоторые эксперты считают наиболее вероятным замедление прогресса, в то время как другие эксперты полагают, что возможен или вероятен чрезвычайно быстрый прогресс.
- Существуют различные технические методы оценки и снижения рисков, связанных с ИИ общего назначения, которые могут использовать разработчики и включать в свои требования регулирующие органы, но все они имеют свои ограничения. Например, существующие методы объяснения того, почему модели ИИ общего назначения выдают тот или иной результат, сильно ограничены.
- Будущее технологий ИИ общего назначения неопределенно, и даже в ближайшем будущем возможны самые разные траектории развития, включая как очень позитивные, так и очень негативные результаты. Но ничто в будущем ИИ не является неизбежным. Определять будущее ИИ будут решения общества и правительств. Данный промежуточный отчет призван способствовать конструктивному обсуждению этих решений.

Данный доклад обобщает состояние научного понимания ИИ общего назначения — ИИ, способного выполнять широкий спектр задач, — и фокусируется на понимании связанных с ним рисков и управлении ними.

Возможности систем, использующих ИИ, стремительно развиваются. Это высветило многочисленные возможности, которые ИИ создает для бизнеса, исследований, правительства и частной жизни. Это также привело к повышению осведомленности о вреде, который может быть нанесен в настоящее время, и потенциальных будущих рисках, связанных с продвинутым ИИ.

Цель Международного научного отчета о безопасности продвинутого ИИ — сделать шаг вперед на пути к достижению общего понимания рисков ИИ и способов их снижения международным сообществом. В этой первой промежуточной публикации доклада основное внимание уделяется типу ИИ, возможности которого развиваются особенно быстро, — ИИ общего назначения, или ИИ, способного выполнять широкий спектр задач.

Быстро продвигающиеся исследования в области ИИ общего назначения в настоящее время находятся в стадии научных открытий и пока не являются устоявшейся наукой. В отчете кратко описано современное научное понимания ИИ общего назначения и

связанных с ним рисков. Сюда входит определение областей, по которым достигнут научный консенсус, и областей, в которых представлены различные мнения или есть открытые вопросы для исследований.

Люди во всем мире смогут безопасно пользоваться многочисленными потенциальными преимуществами ИИ общего назначения только при условии надлежащего управления его рисками. Данный отчет посвящен определению рисков, связанных с ИИ общего назначения, и анализу технических методов их оценки и снижения, в том числе полезного использования ИИ общего назначения для снижения рисков. В его задачи не входит всесторонняя оценка всех возможных последствий ИИ общего назначения для общества, включая преимущества, которые он может обеспечить.

Судя по многим показателям, возможности ИИ общего назначения в последние годы стремительно возросли, и пока нет единого мнения о том, как прогнозировать дальнейший прогресс, поэтому возможны самые разные сценарии.

Согласно многим показателям, возможности ИИ общего назначения стремительно прогрессируют. Пять лет назад ведущие языковые модели ИИ общего назначения за редким исключением не могли выдать связный абзац текста. Сегодня некоторые модели ИИ общего назначения могут вести длительные диалоги с поочередными репликами на самые разные темы, писать короткие компьютерные программы или генерировать видео по описанию. Однако надежно оценить и точно определить возможности ИИ общего назначения достаточно сложно.

Темпы развития ИИ общего назначения зависят как от темпов технологического прогресса, так и от нормативно-правовой базы. В данном отчете основное внимание уделяется технологическим аспектам и не рассматривается вопрос о том, как усилия по регулированию могут повлиять на скорость разработки и внедрения ИИ общего назначения.

В последние годы разработчики ИИ быстро развивали возможности ИИ общего назначения, в основном за счет постоянного наращивания ресурсов, используемых для обучения новых моделей (эта тенденция называется «масштабированием»), и совершенствования существующих алгоритмов. Например, объем вычислительных ресурсов («вычислительный кластер»), используемых для обучения современных моделей ИИ, увеличивался примерно в 4 раза ежегодно, размер наборов данных для обучения — в 2,5 раза, а эффективность алгоритмов (производительность по отношению к вычислительным ресурсам) — в 1,5-3 раза. Вопрос о том, привело ли «масштабирование» к прогрессу в решении фундаментальных проблем, таких как причинно-следственное мышление, вызывает споры среди исследователей.

Темпы будущего прогресса в развитии возможностей ИИ общего назначения имеют существенные последствия для управления возникающими рисками, однако эксперты расходятся во мнениях относительно того, чего стоит ожидать даже в ближайшем

будущем. Эксперты по-разному оценивают вероятность медленного, быстрого или чрезвычайно быстрого развития возможностей ИИ общего назначения. Это разногласие затрагивает ключевой вопрос: достаточно ли постоянного «масштабирования» ресурсов и совершенствования существующих методов для достижения быстрого прогресса и решения таких задач, как обеспечение надежности и фактической достоверности, или же для существенного развития возможностей ИИ общего назначения необходимы новые прорывы в исследованиях?

Несколько ведущих компаний, разрабатывающих ИИ общего назначения, делают ставку на то, что «масштабирование» приведет к дальнейшему росту производительности. Если последние тенденции сохранятся, то к концу 2026 года для обучения некоторых моделей ИИ общего назначения будет использоваться в 40-100 раз больше вычислительных ресурсов, чем использовалось для обучения самых требовательных к вычислительным ресурсам моделей, выпущенных в 2023 году, при этом эффективность методов обучения, которые используют эти ресурсы, увеличится в 3-20 раз. Однако существуют потенциальные препятствия для дальнейшего увеличения объема данных и вычислительных ресурсов, включая доступность данных, чипов ИИ, капитальные затраты и местные энергетические мощности. Компании, разрабатывающие ИИ общего назначения, работают над преодолением этих потенциальных препятствий.

Есть ряд исследований, направленных на улучшение понимания и повышение надежности оценки ИИ общего назначения, однако наше общее понимание того, как работают модели и системы ИИ общего назначения, ограничено.

Подходы к управлению рисками, связанными с ИИ общего назначения, часто основываются на предположении, что разработчики ИИ и нормативной базы могут оценить возможности и потенциальное воздействие моделей и систем ИИ общего назначения. Но хотя технические методы могут помочь в оценке, все существующие методы имеют свои ограничения и в большинстве случаев не могут обеспечить надежную гарантию защиты от вреда, связанного с ИИ общего назначения. В целом, научное представление о внутренних механизмах, возможностях и влиянии ИИ общего назначения на общество весьма ограничено, и эксперты сходятся во мнении, что улучшение понимания ИИ общего назначения должно быть приоритетной задачей. Ниже перечислены некоторые из основных вызовов.

- Разработчики все еще мало понимают, как работают их модели ИИ общего назначения. Это связано с тем, что модели ИИ общего назначения не программируются в традиционном понимании. Вместо этого их обучают: Разработчики ИИ организуют процесс обучения с использованием большого количества данных, и результатом этого процесса становится модель ИИ общего назначения. Эти модели могут состоять из триллионов компонентов, называемых параметрами, и большая часть их внутренних механизмов непостижима, в том числе и для разработчиков моделей. Используя методы объяснения и толкования выбора моделей, исследователи и разработчики могут лучше понять, как работают

модели ИИ общего назначения, но эти исследования находятся на начальной стадии.

- ИИ общего назначения в основном оценивается путем тестирования модели или системы на различных входных данных. Такие выборочные проверки помогают оценить сильные и слабые стороны, включая уязвимости и потенциально опасные возможности, но не дают количественных гарантий безопасности. Эти тесты часто не учитывают опасные факторы и переоценивают или недооценивают возможности, поскольку системы ИИ общего назначения могут вести себя по-разному в разных обстоятельствах, с разными пользователями или при внесении дополнительных изменений в их компоненты.
- Независимые субъекты могут, в принципе, проводить аудит моделей или систем ИИ общего назначения, разработанных какой-либо компанией. Однако зачастую компании не предоставляют независимым аудиторам необходимого уровня прямого доступа к моделям или информации об используемых данных и методах, которые необходимы для тщательной оценки. Некоторые правительства начинают наращивать потенциал, необходимый для проведения технических оценок и аудита.
- Трудно оценить последующее воздействие системы ИИ общего назначения на общество, поскольку существующих исследований в области оценки рисков еще недостаточно для создания строгих и всеобъемлющих методик оценки. Кроме того, ИИ общего назначения имеет широкий спектр вариантов использования, которые зачастую не определены заранее и имеют лишь небольшие ограничения, что еще больше усложняет оценку рисков. Для понимания потенциального последующего воздействия моделей и систем ИИ общего назначения на общество требуется детальный междисциплинарный анализ. Представленность все более широкого спектра точек зрения в процессах разработки и оценки ИИ общего назначения является постоянной технической и институциональной задачей.

ИИ общего назначения может представлять серьезную опасность для личной и общественной безопасности и благополучия.

В этом отчете риски ИИ общего назначения классифицируются по трем категориям: риски злонамеренного использования, риски, связанные с неправильной работой, и системные риски. В нем также рассматривается несколько сквозных факторов, которые способствуют возникновению многих рисков.

Злонамеренное использование. Как и все мощные технологии, системы ИИ общего назначения могут быть использованы со злым умыслом для нанесения вреда. Возможные виды злонамеренного использования варьируются от относительно хорошо задокументированных, таких как мошенничество с использованием ИИ общего назначения, до тех, которые, по мнению некоторых экспертов, могут появиться в ближайшие годы, например, злонамеренное использование научных возможностей ИИ общего назначения.

- Нанесение вреда отдельным лицам с использованием поддельного контента, создаваемого ИИ общего назначения, относится к относительно хорошо задокументированному классу злонамеренного использования ИИ общего назначения. ИИ общего назначения может использоваться для увеличения масштабов и изощренности мошенничества и афер, например, с помощью «фишинговых» атак, усовершенствованных с помощью ИИ общего назначения. ИИ общего назначения также может использоваться для создания поддельного компрометирующего контента с использованием изображений людей без их согласия, например, для производства порнографических дипфейков без согласия участников.
- Еще одна область, вызывающая беспокойство, — злонамеренное использование ИИ общего назначения для дезинформации и манипулирования общественным мнением. ИИ общего назначения и другие современные технологии облегчают создание и распространение дезинформации, в том числе в попытке повлиять на политические процессы. Технические меры противодействия, такие как нанесение водяных знаков на контент, хотя и полезны, обычно могут быть обойдены достаточно продвинутыми субъектами.
- ИИ общего назначения также может быть использован злонамеренно для совершения киберпреступлений, помогая отдельным лицам повысить уровень киберзнаний и облегчая злоумышленникам проведение эффективных кибератак. Системы ИИ общего назначения могут использоваться для масштабирования и частичной автоматизации некоторых типов кибероперациий, например атак с использованием социальной инженерии. Однако ИИ общего назначения может быть использован и в киберзащите. В целом, пока нет серьезных доказательств того, что ИИ общего назначения может автоматизировать сложные задачи кибербезопасности.
- Некоторые эксперты также выражают обеспокоенность тем, что ИИ общего назначения может быть использован для поддержки разработки и злонамеренного применения оружия, например биологического. Нет убедительных доказательств того, что современные системы ИИ общего назначения представляют такую опасность. Например, хотя современные системы ИИ общего назначения демонстрируют рост возможностей, связанных с биологией, имеющиеся результаты ограниченных исследований не дают четких доказательств того, что существующие системы могут повысить способность злоумышленников к получению биологических патогенов лучше, чем использование Интернета. Однако оценка будущих крупномасштабных угроз практически не проводилась, и их трудно исключить.

Риски, связанные с неправильной работой. Даже если у пользователей нет намерения причинить вред, серьезные риски могут возникнуть из-за сбоев в работе ИИ общего назначения. Такие неисправности могут иметь несколько возможных причин и последствий:

- Функциональность продуктов, основанных на моделях и системах ИИ общего назначения, может быть плохо понята их пользователями, например, из-за плохого информирования или недостоверной рекламы. Это может причинить вред, если пользователи будут использовать системы не по назначению или в неподходящих целях.
- Предвзятость в системах ИИ в целом является хорошо задокументированной проблемой, которая остается нерешенной и в области ИИ общего назначения. Результаты работы ИИ общего назначения могут быть предвзятыми по отношению к таким защищенным характеристикам, как раса, пол, культура, возраст и инвалидность. Это может создавать риски, в том числе в таких ответственных сферах, как здравоохранение, трудоустройство и финансовое кредитование. Кроме того, многие широко используемые модели ИИ общего назначения обучаются в первую очередь на наборах данных с непропорционально большим сегментом, касающимся представителей западных культур, что может увеличить вероятность нанесения вреда людям, плохо представленным в этих данных.
- Сценарии «потери контроля» — это потенциальные сценарии будущего, в которых общество больше не сможет значимо сдерживать системы ИИ общего назначения, даже если станет ясно, что они причиняют вред. По общему мнению, нынешний ИИ общего назначения не обладает достаточными возможностями, чтобы представлять такую опасность. Некоторые эксперты считают, что нынешние усилия по разработке *автономного* ИИ общего назначения — систем, способных действовать, планировать и добиваться поставленных целей, — в случае успеха могут привести к потере контроля. Эксперты расходятся во мнениях о том, насколько правдоподобны сценарии потери контроля, когда они могут произойти и насколько сложно будет смягчить их последствия.

Системные риски. Широкая разработка и внедрение технологий ИИ общего назначения сопряжены с рядом системных рисков, начиная от потенциального влияния на рынок труда и заканчивая рисками для конфиденциальности и экологическими последствиями:

- ИИ общего назначения, особенно если он будет развиваться быстрыми темпами, способен автоматизировать очень широкий круг задач, что может оказать значительное влияние на рынок труда. Это может означать, что многие люди могут потерять работу. Однако многие экономисты считают, что потенциальное сокращение рабочих мест может быть компенсировано — возможно, полностью — за счет создания новых рабочих мест и роста спроса в неавтоматизированных секторах.
- Исследования и разработки в области ИИ общего назначения в настоящее время осуществляются в основном в нескольких западных странах и Китае. Такое «ИИ-неравенство» имеет множество причин, но отчасти обусловлено различиями в уровнях доступа к вычислительным ресурсам, необходимым для разработки ИИ общего назначения. Поскольку страны с низким уровнем дохода и академические институты имеют меньший доступ к вычислительным ресурсам, чем страны с

высоким уровнем дохода и технологические компании, они оказываются в невыгодном положении.

- В результате концентрации рынка в сфере разработки ИИ общего назначения общество становится более уязвимым для нескольких системных рисков. Например, широкое использование небольшого числа систем ИИ общего назначения в таких критически важных отраслях, как финансы или здравоохранение, может привести к одновременным сбоям и нарушениям деятельности в масштабе всех этих взаимозависимых секторов, например, из-за ошибок или уязвимостей.
- Увеличение объема вычислительных ресурсов, используемых при разработке и внедрении ИИ общего назначения, привело к быстрому увеличению энергопотребления, связанного с ИИ общего назначения. Эта тенденция не имеет признаков ослабления, что может привести к дальнейшему увеличению выбросов CO_2 и потребления воды.
- Модели или системы ИИ общего назначения могут представлять опасность для конфиденциальности. Например, исследования показали, что, используя «враждебные» входные данные, пользователи могут извлекать из модели использовавшиеся для обучения данные, содержащие информацию об отдельных людях. В случае будущих моделей, обученных на конфиденциальных персональных данных, таких как медицинская или финансовая информация, это может привести к особенно серьезным утечкам конфиденциальной информации.
- Потенциальные нарушения авторских прав при разработке ИИ общего назначения представляют собой проблему в контексте традиционных законов об интеллектуальной собственности, а также систем согласия, компенсации и контроля над данными. Отсутствие ясности в режиме авторского права сдерживает разработчиков ИИ общего назначения от уведомления о том, какие данные они используют, и не дает понимания того, какие меры защиты предусмотрены для авторов, чьи работы используются для обучения моделей ИИ общего назначения без их согласия.

Сквозные факторы риска. В основе рисков, связанных с ИИ общего назначения, лежат несколько сквозных факторов риска — характеристик ИИ общего назначения, которые повышают вероятность или серьезность не одного, а нескольких рисков:

- К сквозным факторам риска технического характера относятся: сложность обеспечения надежного поведения систем ИИ общего назначения в соответствии с их назначением, недостаточное понимание нами их внутренних механизмов и продолжающаяся разработка «агентов» ИИ общего назначения, которые могут действовать автономно при более низком уровне надзора.
- Социальные факторы риска включают в себя потенциальное несоответствие между темпами технологического прогресса и темпами реагирования регулирующих органов, а также конкурентные стимулы, побуждающие разработчиков ИИ выпускать продукты быстро и потенциально в ущерб тщательному управлению рисками.

Есть несколько технических подходов, которые могут помочь снизить риски, но ни один из известных на сегодняшний день методов не дает твердых гарантий и не обеспечивает защиту от вреда, связанного с ИИ общего назначения.

Хотя в этом отчете не рассматриваются политические меры по снижению рисков, связанных с ИИ общего назначения, в нем обсуждаются технические методы снижения рисков, в разработке которых исследователи добились определенного прогресса. Несмотря на этот прогресс, существующие методы не позволяют надежно предотвратить даже откровенно вредные результаты работы ИИ общего назначения в реальных условиях. Для оценки и снижения рисков используется несколько технических подходов. К ним относятся следующее.

- Достигнут определенный прогресс в обучении моделей ИИ общего назначения более безопасному функционированию. Разработчики также обучают модели быть более устойчивыми к вводу данных, призванных заставить модель дать неверный результат («состязательное обучение»). Несмотря на это, противники, как правило, могут найти альтернативные входные данные, которые снижают эффективность мер защиты при незначительных или умеренных усилиях. Ограничение возможностей системы ИИ общего назначения конкретным вариантом использования может содействовать снижению рисков, связанных с непредвиденными сбоями или злонамеренным использованием.
- Существует несколько методов определения рисков, проверки действий системы и оценки эффективности после развертывания системы ИИ общего назначения. Такую практику часто называют «мониторингом».
- Устранение предвзятости в системах ИИ общего назначения может осуществляться на протяжении всего жизненного цикла системы, включая разработку, обучение, развертывание и использование. Однако полностью предотвратить предвзятость в системах ИИ общего назначения весьма сложно, поскольку это требует систематического сбора обучающих данных, постоянной оценки и эффективного выявления рисков предвзятости. Кроме того, возможно, потребуется поступиться объективностью в пользу других целей, таких как точность и конфиденциальность, а также решить, что является полезным знанием, а что — нежелательной предвзятостью, которая не должна отражаться в результатах.
- Защита конфиденциальности — область, в которой ведутся активные исследования и разработки. Простое сведение к минимуму использования конфиденциальных персональных данных в процессе обучения — один из подходов, который может существенно снизить риски для конфиденциальности. Однако, в случае намеренного или ненамеренного использования конфиденциальных данных применение существующих технических средств снижения рисков для конфиденциальности в масштабе больших моделей ИИ общего назначения может быть затруднено, и они не смогут обеспечить пользователям полноценный контроль.

Заключение. ИИ общего назначения может развиваться по ряду разных траекторий, и многое будет зависеть от действий общества и правительств.
Будущее ИИ общего назначения неопределенно, и даже в ближайшей перспективе возможны самые разные траектории развития, включая 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, какие проблемы он призван решать, сможет ли общество в полной мере использовать экономический потенциал ИИ общего назначения, кому он выгоден, каким рискам мы подвергаем себя и сколько средств вкладываем в исследования для снижения рисков — ответы на эти и многие другие вопросы зависят от того, какой выбор сделают общества и правительства сегодня и в будущем, чтобы определить пути развития ИИ общего назначения.

Чтобы способствовать конструктивному обсуждению этих решений, в данном отчете представлен обзор текущего состояния научных исследований и дискуссий по управлению рисками, связанными с ИИ общего назначения. Ставки высоки. Мы с нетерпением ждем продолжения этой работы.