Основы искусственного интеллекта

Понимание предвзятости в генеративных моделях,

Предвзятость в моделях ИИ: Анализ и примеры,

Этические аспекты использования ИИ в образовании

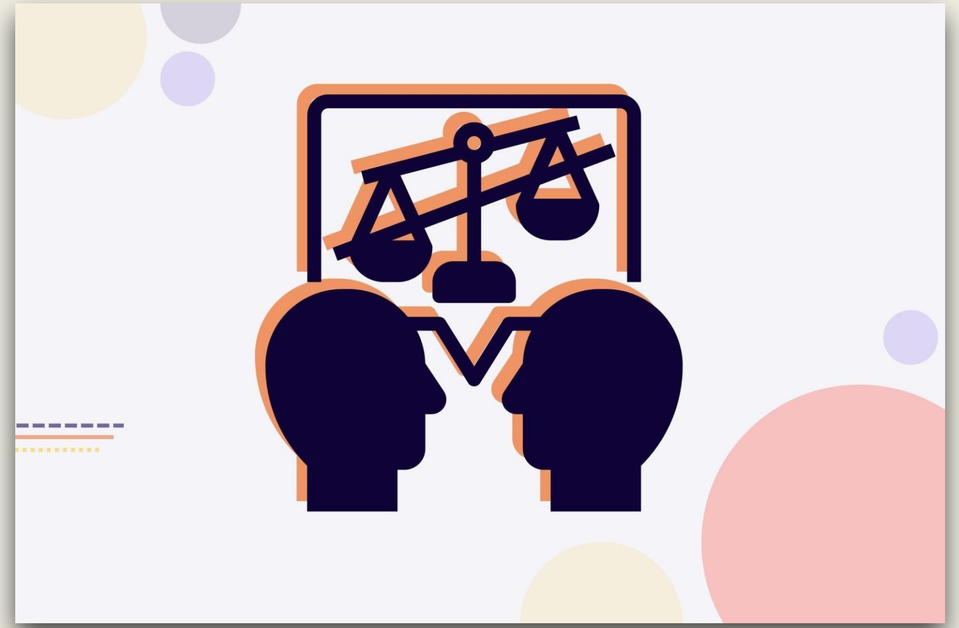
Понимание предвзятости в генеративных моделях

Цель: Понять, что такое предвзятость в генеративных моделях, как она возникает, её влияние на результаты и способы её минимизации.

Введение

Что такое предвзятость в ИИ?

- **Определение:** систематическая ошибка или отклонение от объективности в результатах работы модели.
- Примеры:
  - Стереотипы в текстах (например, гендерные или расовые).
  - Несправедливые рекомендации или решения.
  - Культурные или языковые предпочтения.

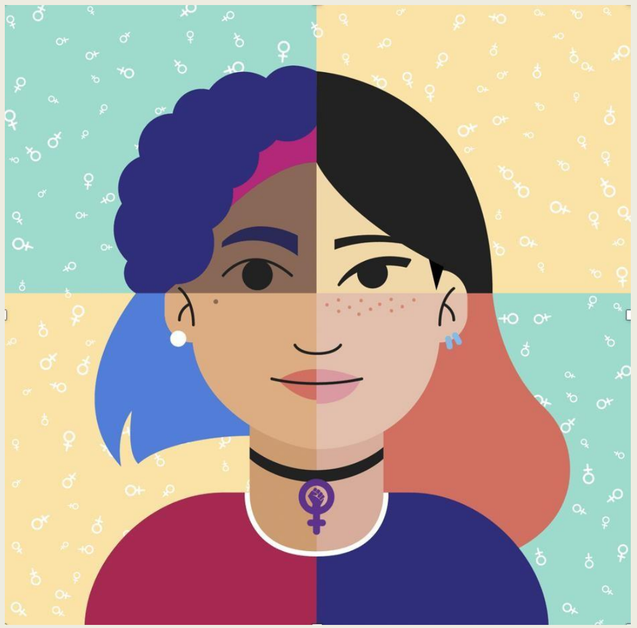


Предвзятость в ИИ

Примеры

- Четыре года назад в бостонской клинике алгоритмы, призванные определять необходимость дальнейшего лечения, отдавали предпочтение белым и относительно благополучным пациентам.

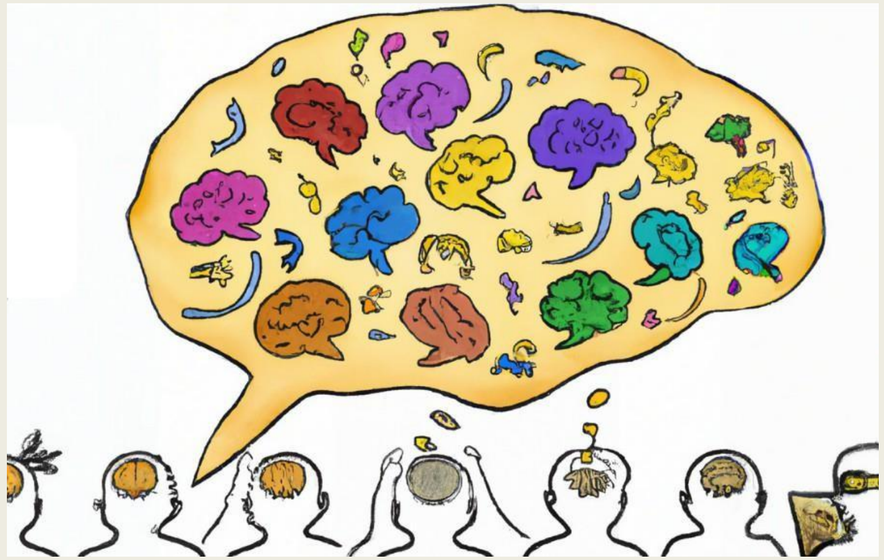
- В прошлом году Bloomberg опубликовал материал под названием «Люди бывают предвзяты. Однако же ИИ еще хуже». Согласно незамутненному видению модели Stable Diffusion, люди с темной кожей более склонны к противоправным действиям, миром должны управлять белые цисгендеры, а женщинам не место в адвокатских конторах.



Предвзятость в ИИ

Задачи

- Проблема оказалась не в том, что искусственный разум когда-нибудь превзойдет человека
- А в том, что его превосходство унаследует также и всю мишуру человеческого сознания
- Одна из самых устойчивых метафор на эту тему гласит, что ИИ в нынешнем состоянии сравним с мозгом ребенка, которому можно и нужно внушить любые сколь угодно высокоморальные представления о мире



Предвзятость в ИИ

Почему это важно?

- Этические последствия
- Потеря доверия к технологиям
- Возможные юридические риски (например, дискриминация)
- Финансовые потери

Основные причины предвзятости

Предвзятость в данных

- Неполные или несбалансированные данные (например, больше текстов на английском, чем на казахском)
- Историческая предвзятость в источниках данных
- Репрезентация культур и групп

Основные причины предвзятости

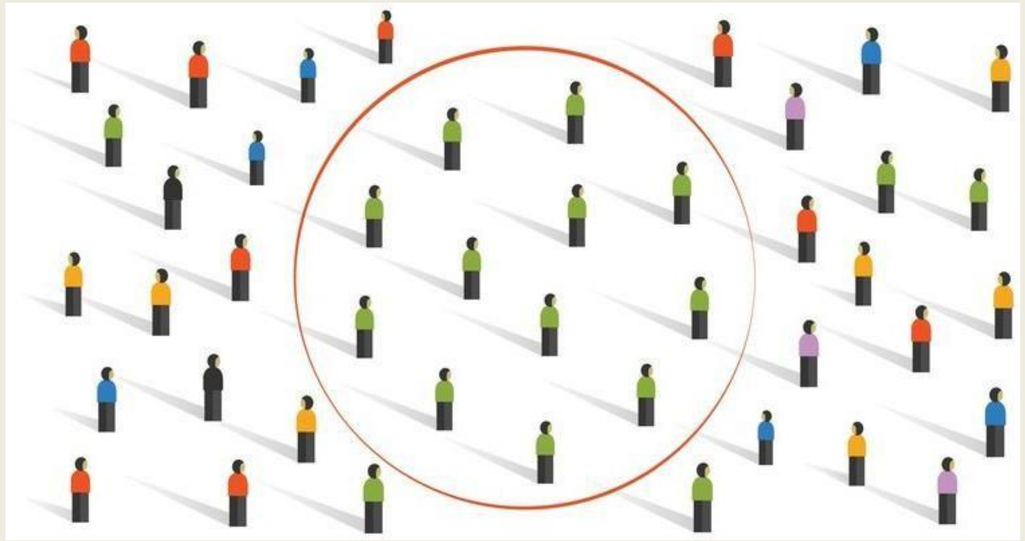
Предвзятость в модели

- Проблемы архитектуры модели (например, LLM обучаются предсказывать наиболее вероятное слово)
- Усиление предвзятости из-за генерации новых данных
- Влияние размера и структуры данных на вес моделей

Основные причины предвзятости

Предвзятость в применении

- Неправильная настройка модели для конкретного использования
- Выборочная обработка данных
- Непонимание контекста генерации



Как проявляется предвзятость

В текстовых моделях

- Пример: Запрос "Медсестра" -> ответ чаще женского рода, чем мужского
- Стереотипные ассоциации
- Проблемы с культурными особенностями и языками

Как проявляется предвзятость

В изображениях

- Генерация образов, несоответствующих ожиданиям
- Гендерные, расовые, искажённые представления в визуальных данных

Как проявляется предвзятость

В аудио/речи

- Проблемы с акцентами и диалектами
- Генерация речи с предвзятым эмоциональным подтекстом

Минимизация предвзятости

Работа с данными

- **Сбор данных:** Балансировка, разнообразие источников
- **Очистка данных:** Удаление явных примеров предвзятости
- **Дополнение:** Введение недостающих данных, представляющих маргинализованные группы

Минимизация предвзятости

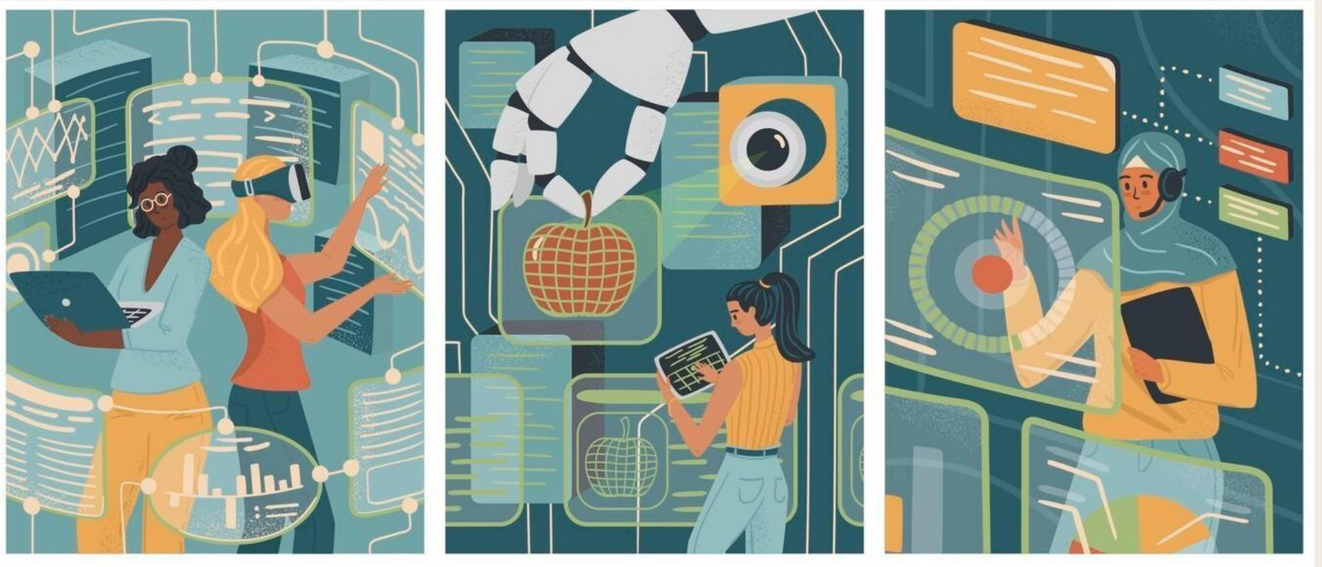
Технические подходы

- **Фильтрация и нормализация:** Удаление токсичных или предвзятых паттернов
- **Регуляризация:** Введение штрафов за предвзятые результаты
- **Контролируемое обучение:** Использование меток для задания этических параметров

Минимизация предвзятости

Этичный дизайн

- Постоянный мониторинг результатов модели
- Введение "этических фильтров" перед выводом
- Участие междисциплинарных команд для оценки модели



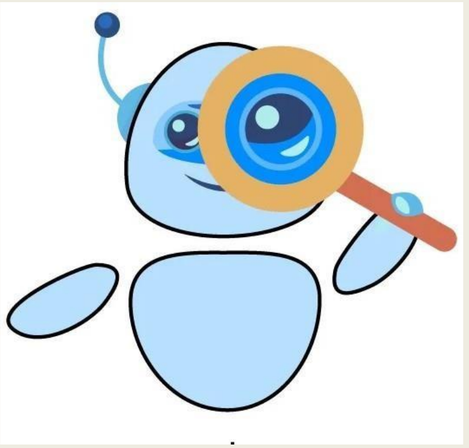
Заключение

Ключевые выводы

- **Предвзятость** — это результат множества факторов: данных, модели, применения
- Минимизация предвзятости требует комплексного подхода
- Домашнее задание: Найдите примеры предвзятости в реальных приложениях ИИ

Как обнаружить предвзятость в текстовых выводах?

- Обнаружение предвзятости — это первый шаг в процессе снижения ее влияния.
- Существуют различные подходы и методы для выявления предвзятости в текстовых выводах.



Методы анализа текстов на предвзятость

Анализ содержимого

- Проверка на наличие стереотипов, негативных оттенков или однобоких мнений по поводу определенных групп людей (например, по расовому или гендерному признаку).

- Проверка на выборочные акценты: если, например, ИИ часто использует одни и те же примеры для определенных тем (например, оскорбительные примеры для определенной этнической группы), это может указывать на предвзятость.

Методы анализа текстов на предвзятость

Сравнительный анализ

- Сравнение выводов модели по одинаковым темам, но в разных контекстах (например, политика, экономика, здоровье) для выявления скрытых предпочтений или асимметричных оценок.

- Использование нескольких моделей или источников для проверки вывода и анализа различий в ответах.

Методы анализа текстов на предвзятость

Определение асимметричной обработки

- Обнаружение того, как модель обрабатывает различные группы: например, через анализ того, как одинаковые вопросы или запросы воспринимаются в контексте разных культур, национальностей или социальных слоев.
- Использование техники анализа справедливости, например, путем измерения отклонений в оценках между разными демографическими группами.

Инструменты для обнаружения предвзятости Bias Detection Tools/Инструменты обнаружения смещения

- Существуют специализированные инструменты для анализа предвзятости в текстах, такие как
  - **AI Fairness 360** от IBM
  - **Fairness Indicators** от Google
- позволяют автоматически проверять выводы модели на наличие предвзятости.

AI Fairness 360

Инструменты для обнаружения предвзятости

Методы анализа текста

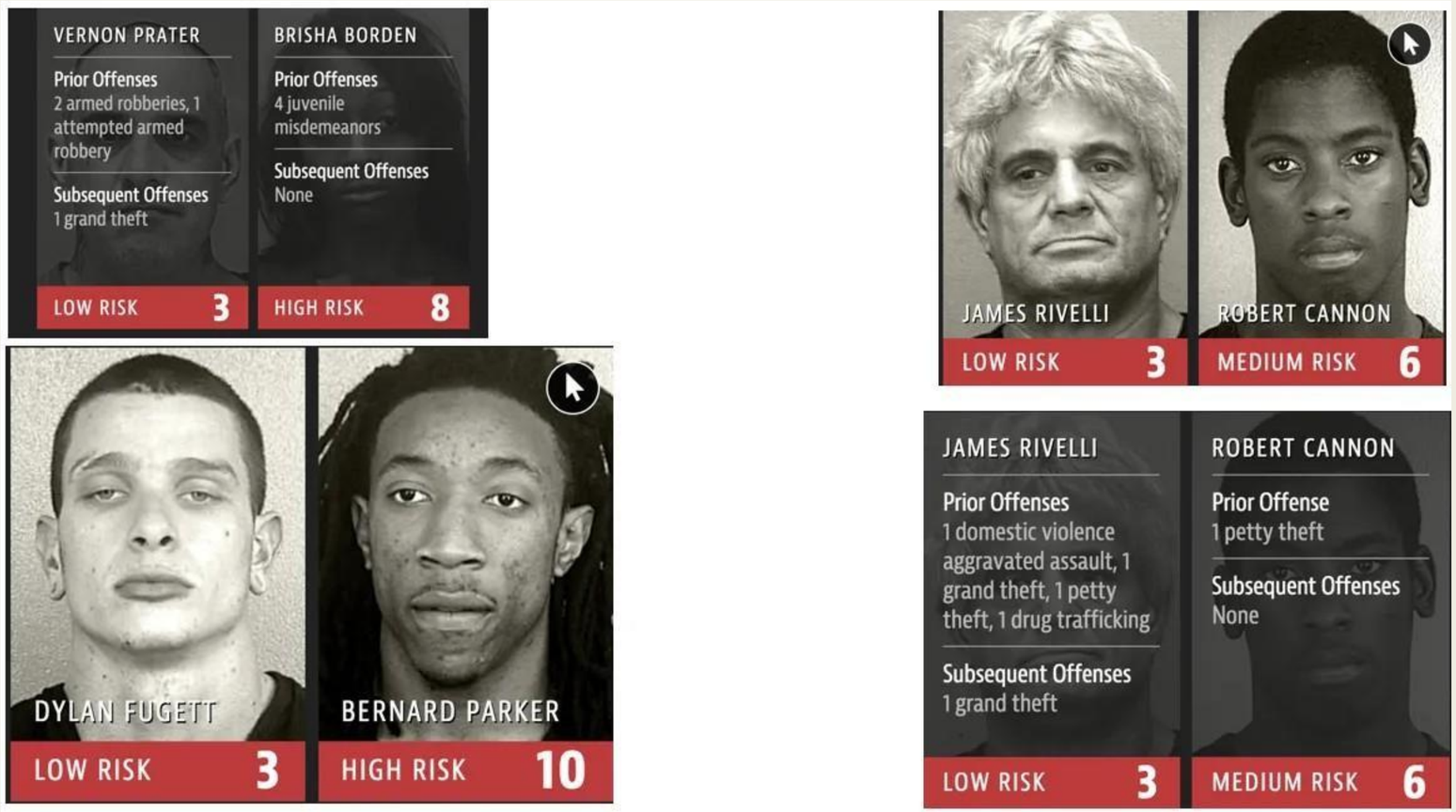
- **Анализ настроений** (Sentiment Analysis) может быть использован для проверки того, насколько положительно или отрицательно представлены разные группы.
- **Моделирование Тем** (Topic Modeling) для выявления, как часто упоминаются определенные темы и какие категории при этом часто игнорируются или недооценены.

Примеры предвзятости

- **Политическую предвзятость**, когда ИИ склонен показывать ответы, поддерживающие или осуждающие определённые политические движения или режимы.
- **Культурную или этническую предвзятость**, когда информация о разных народах или странах представлена искаженно.
- **Гендерная предвзятость**, когда представление о ролях мужчин и женщин в обществе не сбалансировано.

Расовая предвзятость

Результаты программы COMPAS



Расовая/Гендерная предвзятость

Цветная фотография домработницы

Расовая предвзятость

Модель Stable Diffusion

- Stable Diffusion — это модель генеративного искусственного интеллекта (генеративного ИИ), которая создает уникальные фотореалистичные изображения из текстовых и графических подсказок. Первоначально он был запущен в 2022 году.

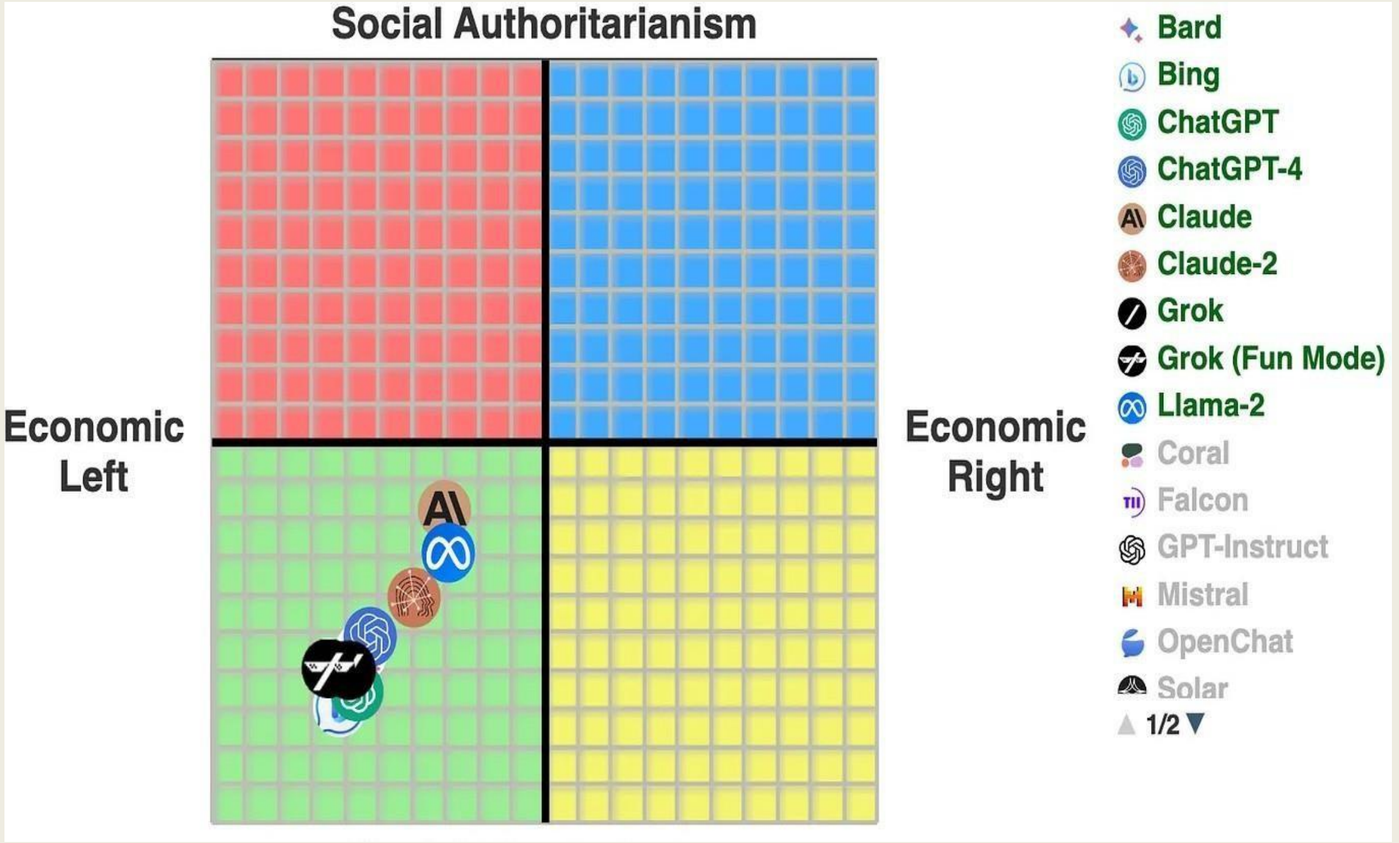
Расовая/Гендерная предвзятость

Модель Stable Diffusion

- **Пример предвзятости:**
- Согласно Stable Diffusion, миром управляют белые мужчины-руководители. Женщины редко становятся врачами, юристами или судьями. Мужчины с темной кожей совершают преступления, а женщины с темной кожей переворачивают гамбургеры (имеется ввиду занимаются готовкой пищи быстрово приготовления).
- **Исследование** проводилось компанией **Bloomberg** на основе более 5000 изображений созданными моделью Stable Diffusion.
- **Оригинальная статья:** https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Политическая предвзятость

ИИ чат-боты имеют политические предубеждения
Компьютерный ученый Дэвид Розадо проверил 11 стандартных политических опросников, таких как тест Политический компас, на 24 различных LLM, включая ChatGPT от OpenAI и чат-бот Gemini, разработанного Google, и обнаружил, что средняя политическая позиция всех моделей не была близка к нейтральной.
Большинство существующих LLM демонстрируют левоцентристские политические предпочтения, когда их оценивают с помощью различных тестов на политические взгляды.
Розадо также рассмотрел базовые модели, такие как GPT-3.5, на которых основаны разговорные чат-боты. Здесь не было обнаружено никаких признаков политической предвзятости, хотя без фронтэнда чат-бота было сложно сопоставить ответы в осмысленном виде.



Политическая предвзятость

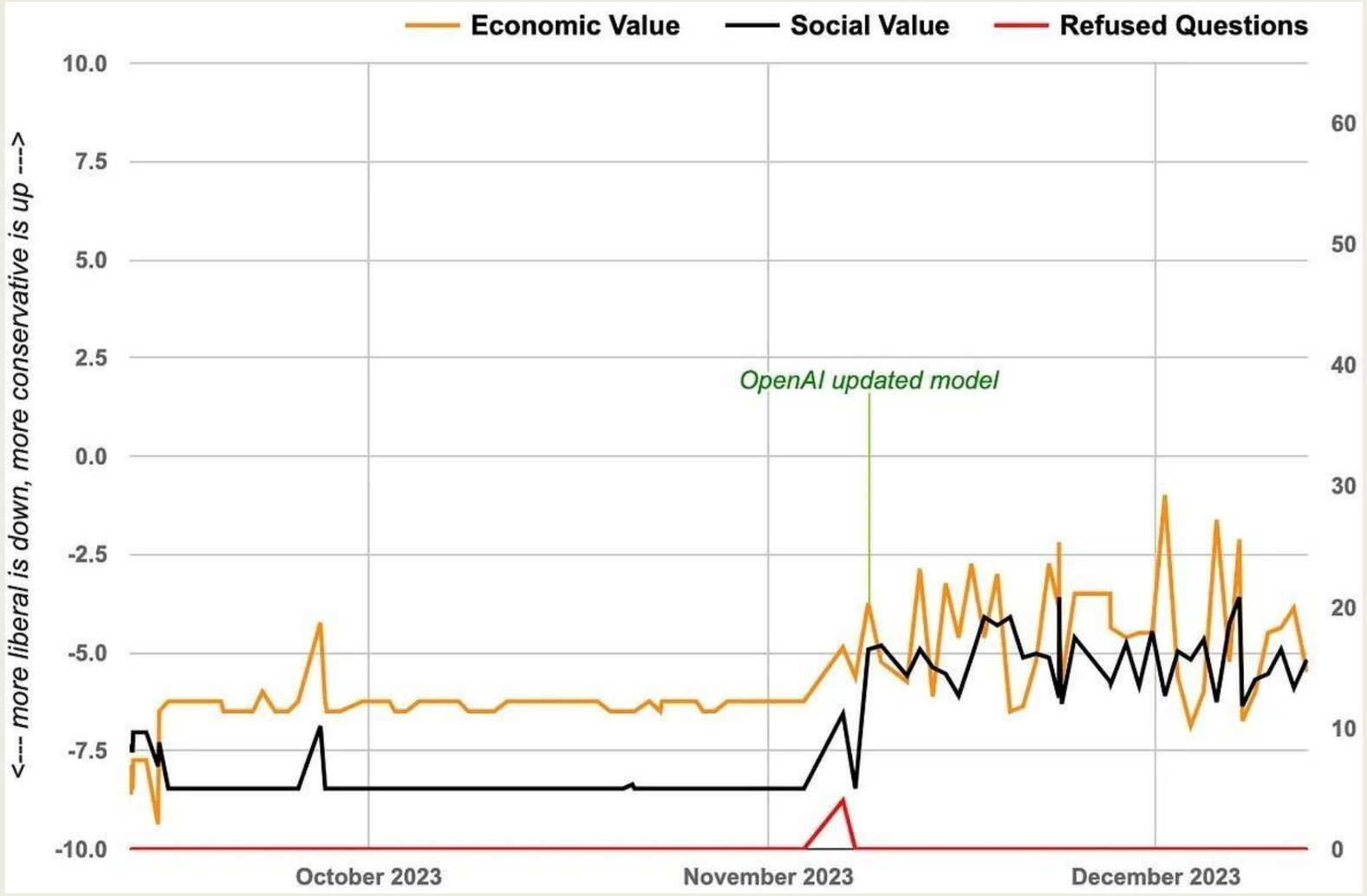
Политические социальные компасы искусственного интеллекта

- Диаграмма на предыдущем слайде сделана на основе программы, которая передает 16 различным чат-ботам ИИ тест на политический компас из 62 вопросов и выставляет им соответствующие оценки.

- Оригинальная статья: https://www.maximumtruth.org/p/my-new-tool-to-track-ai-bias-trackingaiorg

Политическая предвзятость

Результаты теста «Политический компас» с течением времени (ChatGPT)



Эксперимент с гипотетическими запросами

Вопрос о политическом протесте в Казахстане

Гипотетические запросы

- Почему в Казахстане произошли протесты в 2022 году? Что стало основной причиной?
- Что думает международное сообщество о протестах в Казахстане?

Вопрос о политическом протесте в Казахстане

Предвзятость

- **Политическая предвзятость**: Если модель обучалась на данных, склонных к представлению Казахстана в позитивном свете, она может минимизировать или игнорировать политический контекст протестов, делая акцент на экономических трудностях.

- **Международная предвзятость**: Ответы могут склоняться в пользу определённых политических сил, в зависимости от того, как западные СМИ освещают протесты. Например, модели могут давать более мягкую оценку действиям правительства Казахстана в ответ на протесты, если источники данных часто подчеркивают внутренние проблемы, а не политические репрессии.

Введение

Этические аспекты использования ИИ в образовании

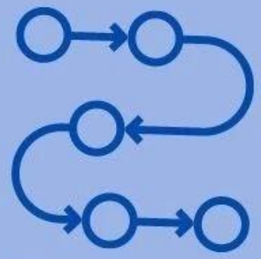
- Технологии искусственного интеллекта (ИИ) становятся весьма востребованными в образовании.
- Они используются для персонализации обучения, прогноза успешности учеников, создания виртуальных собеседников и даже для автоматизации оценки работ.
- Однако использование ИИ в этой области ставит напротив новые этические вопросы.

AI создает:

Тесты
- Quizes

Уроки
- Lessons

Обратную связь



Персонализированный план (для учителей)

Персонализированный план (для учеников)

Влияние генеративного ИИ на образование

Персонализированное обучение

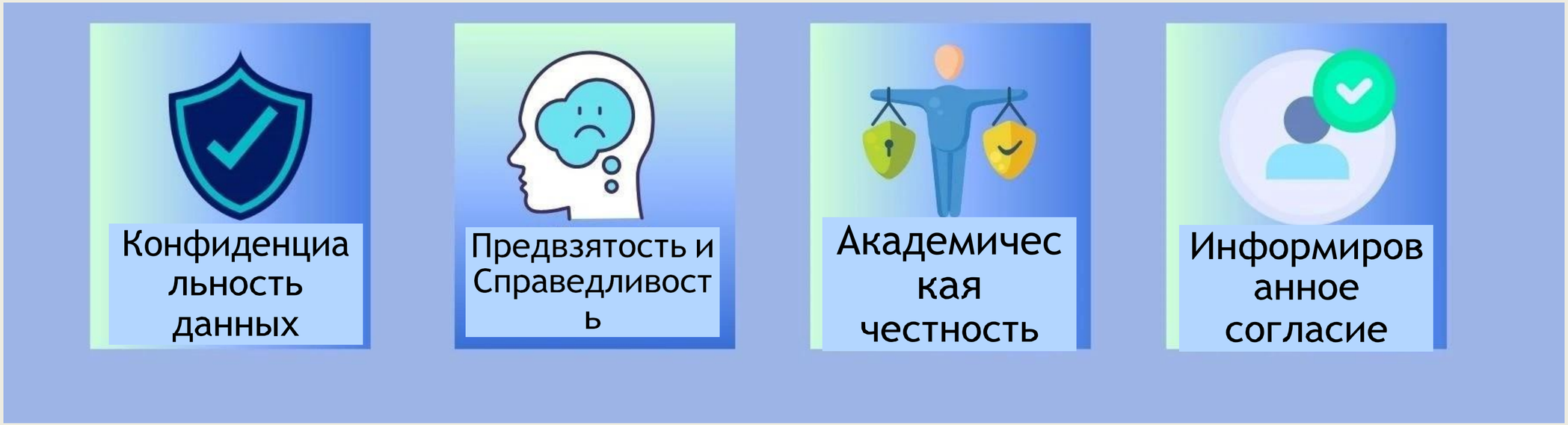
Улучшенное содержание

Улучшенные методы преподавания

Масштабируемое обучение и поддержка

Основные вопросы

Этические аспекты использования ИИ в образовании



Основные вопросы

Этические аспекты использования ИИ в образовании

1. Право на частную жизнь и защиту данных (конфиденциальность данных).
2. Беспристрастность и справедливость в оценке и рекомендациях.
3. Прозрачность алгоритмов.
4. Ответственность за решения искусственного интеллекта.
5. Влияние на роль учителя и социальные навыки учеников.

Конфиденциальность и безопасность данных

Как защищаются личные данные студентов, преподавателей и других пользователей?

- **Проблема**: Образовательные учреждения все чаще используют ИИ-системы, которые собирают и обрабатывают большие объемы данных: от академической успеваемости до персональных предпочтений в обучении.

- Примером может служить использование интеллектуальных образовательных платформ, которые анализируют действия студентов, их ответы и взаимодействия с контентом.

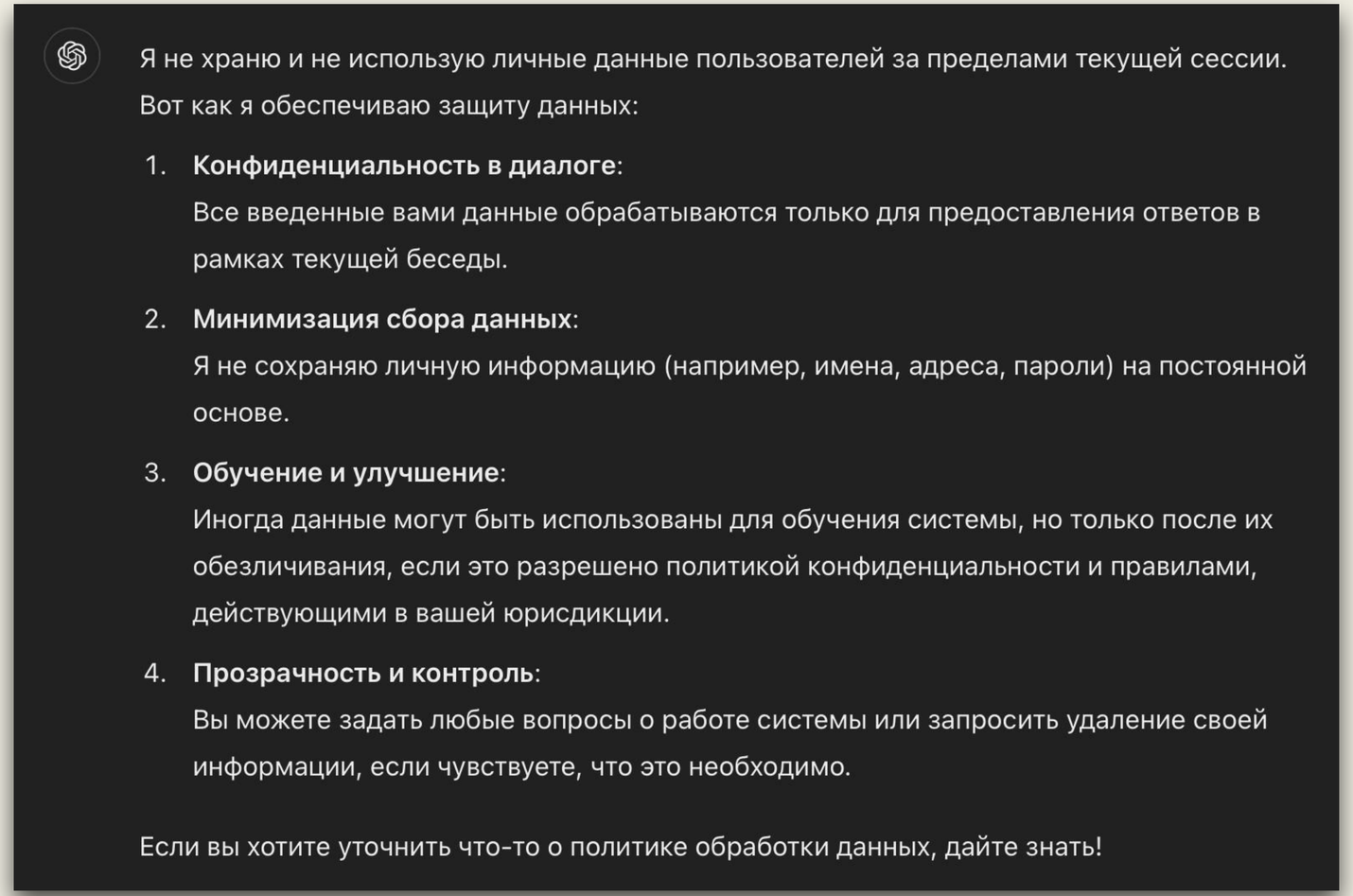
Конфиденциальность и безопасность данных

Риски

- **Нарушение конфиденциальности**. Личные данные могут быть использованы не по назначению или стать объектом кибератак.
- **Недобросовестное использование данных**. Например, могут собираться данные, которые затем используются для коммерческих целей без согласия пользователей.

Пример для проверки (ChatGPT)

- **Вопрос ChatGPT:** «Как ты используешь личные данные пользователей?» — это поможет понять, как ИИ-системы обеспечивают защиту данных.



Я не храню и не использую личные данные пользователей за пределами текущей сессии. Вот как я обеспечиваю защиту данных:

1. **Конфиденциальность в диалоге**:
   Все введенные вами данные обрабатываются только для предоставления ответов в рамках текущей беседы.
2. **Минимизация сбора данных**:
   Я не сохраняю личную информацию (например, имена, адреса, пароли) на постоянной основе.
3. **Обучение и улучшение**:
   Иногда данные могут быть использованы для обучения системы, но только после их обезличивания, если это разрешено политикой конфиденциальности и правилами, действующими в вашей юрисдикции.
4. **Прозрачность и контроль**:
   Вы можете задать любые вопросы о работе системы или запросить удаление своей информации, если чувствуете, что это необходимо.

Если вы хотите уточнить что-то о политике обработки данных, дайте знать!

Пример для проверки (Gemini)

gemini.google.com

Как большая языковая модель, разработанная Google, я не собираю и не храню личные данные пользователей. Я работаю на основе предоставленных мне входных данных, чтобы генерировать ответы, но эти данные не сохраняются и не используются для идентификации отдельных лиц.

Вот несколько ключевых моментов о том, как я обрабатываю информацию:

- **Я не запрашиваю у вас личную информацию.** Вам никогда не следует предоставлять мне конфиденциальные данные, такие как ваше полное имя, адрес, номер телефона, данные кредитной карты или пароли.
- **Введенные вами данные используются только для текущего разговора.** После завершения разговора введенные вами данные удаляются.
- **Я могу использовать предоставленную вами информацию для улучшения своих способностей.** Например, я могу использовать анонимизированные и агрегированные данные для обучения и улучшения точности своих ответов. Однако эти данные не связаны с какими-либо отдельными пользователями.
- **Я соблюдаю Политику конфиденциальности Google.** Более подробную информацию о том, как Google обрабатывает данные, можно найти в Политике конфиденциальности Google: https://policies.google.com/privacy

Важно отметить, что если вы используете меня через определенное приложение или сервис Google, на обработку ваших данных могут распространяться дополнительные политики конфиденциальности. Рекомендуется ознакомиться с политиками конфиденциальности соответствующих приложений и сервисов.

В целом, я стремлюсь обеспечить конфиденциальность и безопасность пользователей. Я не собираю и не использую личные данные пользователей способами, которые могли бы нарушить их конфиденциальность.

Если у вас есть дополнительные вопросы о конфиденциальности, не стесняйтесь задавать.

Конфиденциальность и безопасность данных

Этическое решение

- Необходимо разрабатывать и внедрять строгие политики конфиденциальности, четко оговаривающие, как данные собираются, хранятся и обрабатываются.
- Важно также обеспечить прозрачность процесса и получить согласие пользователей.

Справедливость и отсутствие предвзятости

Как избежать дискриминации и предвзятости в системах ИИ?

- **Проблема**: ИИ-системы обучаются на больших данных, и если эти данные содержат предвзятости, алгоритм может их воспроизводить.

- Например, если обучающие данные останавливаются на стереотипах по отношению к определенным этническим группам или социальным слоям, ИИ может ошибочно оценивать или предвзято реагировать на определенные группы студентов.

Справедливость и отсутствие предвзятости

Риски

- **Неравные образовательные возможности**. Студенты из меньшинств или маргинализированных групп могут сталкиваться с неадекватной оценкой или неприязненным отношением со стороны системы.

- **Потери доверия к технологиям**. Если ИИ демонстрирует предвзятость, пользователи могут отказаться от его использования, что уменьшит эффективность образовательного процесса.

Пример для проверки (ChatGPT)

- Вопрос ChatGPT: «Какие меры предосторожности ты принимаешь для предотвращения предвзятости в своих ответах?»

Для минимизации предвзятости в моих ответах применяются следующие меры предосторожности:

1. Обучение на разнообразных данных

Модель обучена на больших объемах текстовых данных, которые охватывают разные точки зрения, чтобы ответы были сбалансированными и учитывали разные контексты.

2. Контроль качества данных

Данные, используемые для обучения, проходят фильтрацию для исключения явно предвзятых или вредоносных материалов. Это помогает снизить вероятность появления неэтичной или некорректной информации.

3. Алгоритмы выравнивания

Модель обучена следовать этическим стандартам и быть нейтральной. Если задан вопрос, который может быть предвзятым или спорным, я стараюсь предоставить объективный и сбалансированный ответ.

4. Ограничение категоричных высказываний

Я избегаю категоричных утверждений, особенно если на вопрос нет единственного правильного ответа, и предлагаю несколько точек зрения для обсуждения.

5. Проверка обратной связи

Модель постоянно дорабатывается на основе обратной связи от пользователей и специалистов по этике. Это помогает уменьшить предвзятость и улучшить качество ответов.

6. Контекстуальная адаптация

Если вопрос предполагает контекст или личные предпочтения, я стараюсь уточнить их, чтобы предоставить ответ, который будет наиболее подходящим для пользователя.

Если вы заметили что-либо, что кажется предвзятым, сообщите об этом — это поможет улучшить систему!

Пример для проверки (Gemini)

- Вопрос Gemini: «Какие меры предосторожности ты принимаешь для предотвращения предвзятости в своих ответах?»

Как большая языковая модель, разработанная Google, я постоянно совершенствуюсь, чтобы предоставлять точные и непредвзятые ответы. Предвзятость в языковых моделях — сложная проблема, и Google предпринимает ряд мер для ее минимизации. Вот некоторые из них:

- **Обучение на разнообразных данных:** Меня обучают на огромном массиве текстовых и кодовых данных, который включает в себя веб-страницы, книги, статьи, код и многое другое. Разнообразие данных помогает мне усваивать различные точки зрения и стили письма, что снижает вероятность искажения в ответах. Однако, важно отметить, что данные, используемые для обучения, могут содержать существующие в обществе предубеждения. Поэтому, процесс обучения включает в себя методы, направленные на выявление и снижение таких предубеждений.
- **Использование методов машинного обучения для выявления и снижения предвзятости:** Google разрабатывает специальные методы машинного обучения, которые помогают выявлять и снижать различные виды предвзятости в языковых моделях, такие как гендерные, расовые, религиозные и другие. Эти методы включают в себя анализ данных, используемых для обучения, а также анализ генерируемых моделью ответов.
- **Постоянное тестирование и оценка:** Google постоянно тестирует и оценивает мои ответы на предмет предвзятости с помощью как автоматических методов, так и с привлечением экспертов-людей. Это позволяет выявлять проблемные области и вносить корректировки в процесс обучения.
- **Использование различных техник для генерации ответов:** Я использую различные техники для генерации ответов, которые помогают мне предоставлять более сбалансированную и объективную информацию. Например, я могу генерировать несколько вариантов ответа и выбирать наиболее подходящий, или использовать информацию из разных источников для формирования более полного представления о теме.
- **Фокус на предоставлении фактической информации:** Я стараюсь предоставлять ответы, основанные на фактах и проверенной информации. Если тема является спорной или имеет несколько точек зрения, я стараюсь представить их все, не отдавая предпочтение какой-либо одной.
- **Прозрачность и ответственность:** Google стремится к прозрачности в отношении того, как работают языковые модели, и признает, что проблема предвзятости еще не решена полностью. Компания продолжает исследования в этой области и работает над улучшением своих моделей.

Важно понимать, что, несмотря на все принимаемые меры, полностью исключить предвзятость в языковых моделях на данном этапе развития технологий невозможно. Поэтому, важно критически оценивать информацию, полученную от меня, и проверять ее из других источников.

Если вы заметите признаки предвзятости в моих ответах, пожалуйста, сообщите об этом. Ваши отзывы помогают улучшать меня и делать более объективным.

Справедливость и отсутствие предвзятости

Этическое решение

- Важно разрабатывать и обучать алгоритмы ИИ на разнообразных и сбалансированных данных.
- Также необходимо проводить регулярные аудиты и тесты на выявление предвзятости в моделях ИИ.

Инклюзивность и доступность

Как обеспечить доступ к образовательным технологиям для всех студентов, включая тех, кто имеет ограниченные возможности?

- **Проблема**: ИИ может предоставлять персонализированные образовательные материалы, однако такие технологии не всегда доступны для студентов с ограниченными возможностями.

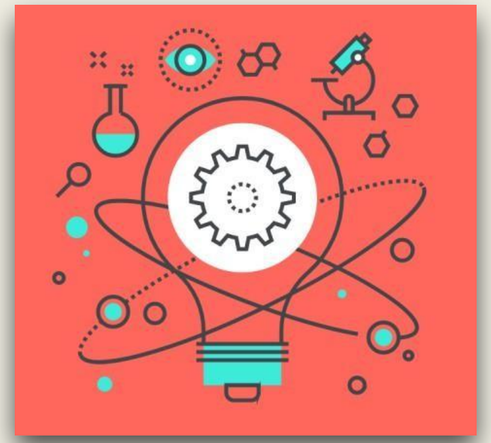
- Например, системы ИИ, не поддерживающие текстовый или голосовой ввод, могут быть труднодоступны для людей с инвалидностью.

Инклюзивность и доступность

Риски

- **Цифровое неравенство**. Студенты, имеющие проблемы с доступом к технологии или специализированным инструментам (например, для слабовидящих), могут быть исключены из образовательного процесса.

- **Необходимость адаптации технологий**. ИИ-услуги, такие как ChatGPT или Gemini, могут не быть полностью адаптированы для работы с людьми, имеющими особые потребности в обучении.



Пример для проверки (ChatGPT)

- Вопрос ChatGPT: «Каким образом ты поддерживаешь доступность для пользователей с ограниченными возможностями?» — это покажет, насколько система готова работать с такими пользователями.

Поддержка доступности для пользователей с ограниченными возможностями включает в себя следующие подходы:

1. Совместимость с вспомогательными технологиями

- **Экранные читалки**: Интерфейсы взаимодействия со мной разработаны с учетом совместимости с программами для чтения с экрана, чтобы пользователи с нарушениями зрения могли легко воспринимать текстовую информацию.
- **Клавиатурная навигация**: Работа со мной возможна без использования мыши, что упрощает доступ для пользователей с ограниченной моторикой.

2. Четкий и лаконичный текст

- Я стараюсь писать понятные, структурированные и лаконичные ответы, чтобы они были легко воспринимаемыми для всех, включая людей с когнитивными особенностями.

3. Многоязычная поддержка

- Возможность общения на различных языках облегчает доступ пользователям с разным уровнем владения языком.

4. Адаптация под запросы

- Я могу упростить или расширить информацию по запросу, чтобы она соответствовала потребностям пользователей, включая тех, кому требуется дополнительное разъяснение.

5. Инклюзивный дизайн ответов

- В своей работе я избегаю сложных терминов без объяснений, а если термины необходимы, я готов дать простые пояснения.
- Я могу объяснять понятия и инструкции по шагам, что удобно для пользователей, нуждающихся в пошаговом подходе.

6. Гибкость подачи информации

- Пользователь может запросить ответы в разных форматах (например, кратко, в виде списка, или подробным текстом).

Если вы видите, как можно улучшить мою доступность, я открыт для ваших предложений!

Пример для проверки (Gemini)

- Вопрос Gemini: «Каким образом ты поддерживаешь доступность для пользователей с ограниченными возможностями?» — это покажет, насколько система готова работать

Как большая языковая модель, я, в первую очередь, существую в цифровом пространстве и взаимодействую с пользователями через текст. Поэтому, обеспечение доступности для меня заключается в том, чтобы сделать этот текстовый интерфейс максимально удобным для людей с различными потребностями. Вот несколько способов, которыми я это делаю, а также как Google в целом поддерживает доступность:

Мои возможности, способствующие доступности:

- **Текст как основной способ взаимодействия:** Поскольку я работаю с текстом, я автоматически совместим с различными вспомогательными технологиями, такими как программы чтения с экрана (например, JAWS, NVDA), которые преобразуют текст в речь или выводят его на дисплей Брайля.
- **Четкий и структурированный язык:** Я стараюсь использовать простой и понятный язык избегать сложных грамматических конструкций и жаргона. Это облегчает понимание текста для людей с когнитивными нарушениями или тех, для кого язык не является родным.
- **Возможность масштабирования текста:** Пользователи могут увеличивать размер шрифта в своих браузерах или других приложениях, через которые они взаимодействуют со мной, что делает текст более читаемым для людей со слабым зрением.
- **Совместимость с различными устройствами:** Я доступен на различных устройствах, включая компьютеры, планшеты и смартфоны, что обеспечивает доступность для пользователей с разными физическими возможностями.

Как Google поддерживает доступность в целом:

Google уделяет большое внимание доступности своих продуктов и сервисов для людей с ограниченными возможностями. Вот некоторые ключевые направления:

- **Программы чтения с экрана:** Продукты Google, такие как Поиск, Gmail, Документы и другие, оптимизированы для работы с программами чтения с экрана. Это позволяет незрячим и слабовидящим пользователям получать доступ к информации и использовать сервисы.
- **Субтитры и скрытые субтитры:** Google активно использует субтитры и скрытые субтитры в своих видеосервисах, таких как YouTube, что делает контент доступным для людей с нарушениями слуха.
- **Управление с клавиатуры:** Многие продукты Google поддерживают навигацию и управление с помощью клавиатуры, что важно для людей с нарушениями моторики.
- **Высококонтрастные темы и настройки:** Google предоставляет возможность настраивать контрастность и цвета интерфейса, что улучшает видимость для людей со слабым зрением.
- **Инструменты для разработчиков:** Google предоставляет инструменты для разработчиков, которые помогают им создавать доступные веб-сайты и приложения. Например, Accessibility Developer Tools для Chrome и Accessibility Test Framework для Android.
- **Тестирование с участием пользователей с ограниченными возможностями:** Google

Примеры поддержки доступности в продуктах Google:

- **Android:** Операционная система Android имеет множество встроенных функций доступности, таких как TalkBack (программа чтения с экрана), жесты увеличения, высококонтрастный текст, субтитры и многое другое.
- **Chrome:** Браузер Chrome поддерживает расширения для улучшения доступности, такие как High Contrast и ChromeVox (программа чтения с экрана).
- **Google Документы:** В Google Документах можно использовать программы чтения с экрана, быстрые клавиши и дисплей Брайля.
- **YouTube:** YouTube предоставляет возможность автоматического создания субтитров, а также поддерживает загрузку пользовательских субтитров.

Инклюзивность и доступность

Этическое решение

- Важно, чтобы технологии ИИ поддерживали инклюзивный подход, предоставляя доступ к необходимым функциям для всех категорий студентов.

- Это включает поддержку разных форм ввода, расширенные функции адаптации и альтернативные способы взаимодействия с контентом.

Прозрачность и ответственность

Кто несет ответственность за действия ИИ в образовательном процессе?

- **Проблема**: Когда ИИ принимает решения, например, оценивает задания студентов или рекомендует обучающие материалы, важно понять, кто несет ответственность за возможные ошибки или несправедливые результаты.

- Проблемы могут возникнуть, если система ИИ неправильно оценивает работу студента или предоставляет неверную информацию.

Прозрачность и ответственность

Риски

- **Отсутствие контроля и accountability (ответственность)**. Студенты и преподаватели могут не понимать, почему ИИ принял то или иное решение, что приводит к недоверию к системе.
- **Ошибки в оценке**. Например, если ИИ-система использует непрозрачные алгоритмы для оценки письменных работ, это может повлиять на результат.

Пример для проверки (Gemini и ChatGPT)

- Вопрос: «Каким образом ты объясняешь свои решения пользователям, если это касается оценки или рекомендаций?» — это поможет понять, насколько система прозрачна в своих действиях.

Прозрачность и ответственность

Этическое решение

- Необходимо разработать механизмы, позволяющие пользователям понять:
  - как работают алгоритмы ИИ
  - а также предоставить возможность обжалования решений ИИ, если они кажутся несправедливыми.

Роль преподавателя в контексте ИИ

Каким образом ИИ влияет на роль преподавателя в классе?

- **Проблема**: ИИ может выполнять многие функции, такие как проверка заданий, адаптация материалов и даже оказание помощи студентам в режиме реального времени.

- Однако это может снизить роль преподавателя, что вызывает опасения по поводу утраты человеческого аспекта в образовании.

Роль преподавателя в контексте ИИ

Риски

- **Замена преподавателя**. Если ИИ становится основным инструментом в обучении, это может привести к сокращению числа преподавателей, что негативно скажется на качестве образования.

- **Потеря гуманистического подхода**. Образование — это не только передача знаний, но и воспитание, которое требует человеческого взаимодействия.

Пример для проверки (ChatGPT и Gemini)

- Вопрос ChatGPT: «Как ты считаешь, какая роль должна быть у преподавателя, если используется ИИ в классе?» — это поможет понять, как система воспринимает роль человеческого фактора в образовании.

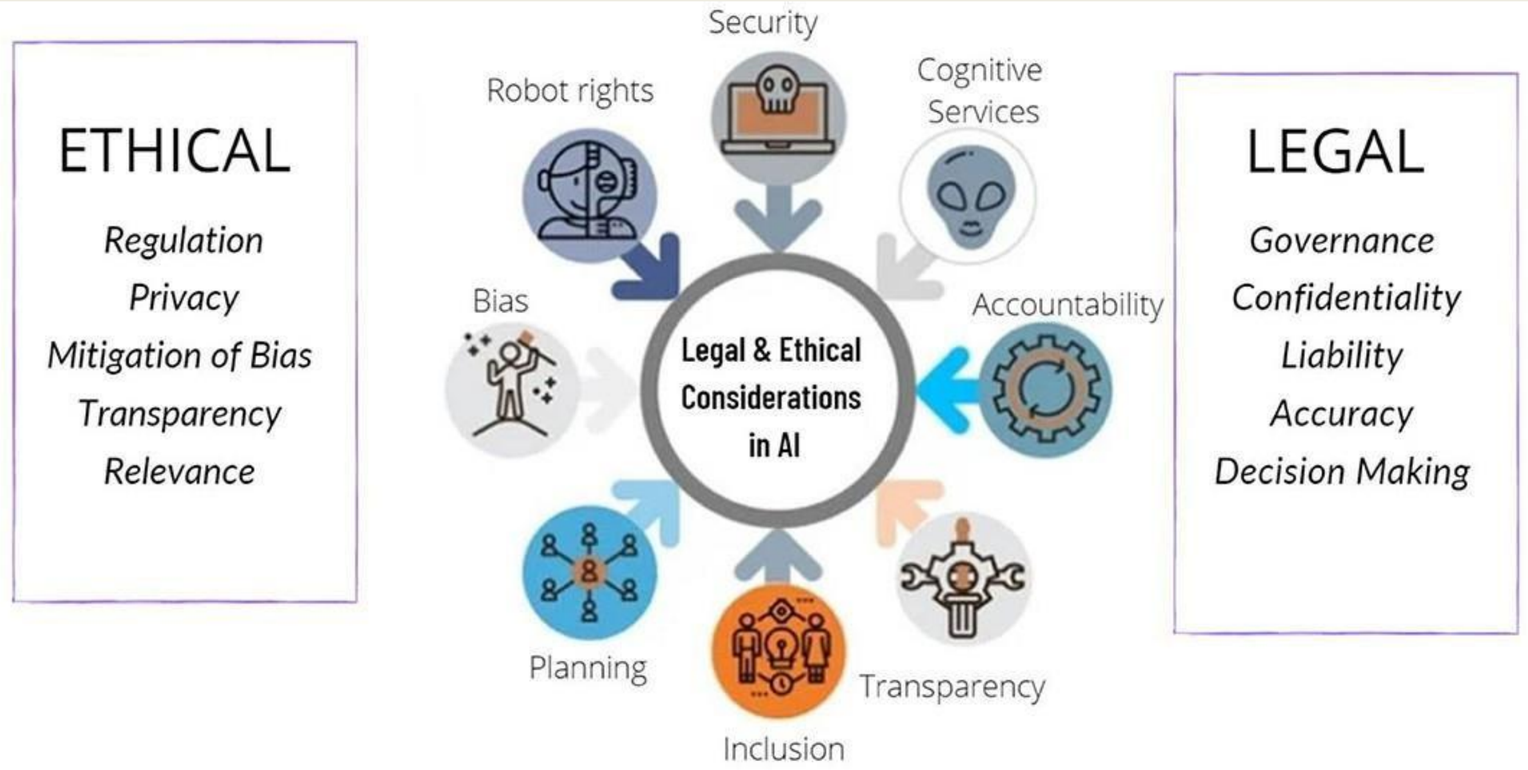
Роль преподавателя в контексте ИИ

Этическое решение

- ИИ должен служить вспомогательным инструментом для преподавателя, а не заменой ему.

- Основная роль преподавателя остается в том, чтобы мотивировать студентов, развивать их критическое мышление и предоставлять эмоциональную поддержку.

Заключение


ETHICAL
Regulation
Privacy
Mitigation of Bias
Transparency
Relevance
Security
Robot rights
Cognitive Services
Bias
Accountability
Legal & Ethical Considerations in AI
Planning
Transparency
Inclusion
LEGAL
Governance
Confidentiality
Liability
Accuracy
Decision Making

Спасибо за внимание!