Почему люди беспокоятся об ИИ

- Искусственный интеллект (ИИ) стремительно развивается и начинает влиять на все сферы жизни.
- После появления ChatGPT прогресс в области ИИ значительно ускорился.
- Каждую неделю появляются технологии, которые раньше считались бы прорывными.
- Появляются **рассуждающие модели** — ИИ, способные выполнять логические цепочки, а не просто генерировать текст.
- **AGI (Artificial General Intelligence)** — это ИИ, способный выполнять любые интеллектуальные задачи, которые может выполнить человек.
- **Технологическая сингулярность** — гипотетический момент, когда ИИ начнёт улучшать сам себя без участия человека, вызывая взрывной рост прогресса.
- Перспектива появления AGI вызывает страх, так как последствия могут быть непредсказуемыми.

Реакция экспертов

- После появления GPT-4 ведущие исследователи начали открыто выражать тревогу.
- Риски, связанные с ИИ, сравниваются с ядерной угрозой.
- В декабре 2024 года был опубликован **AI Safety Report 2025**, в котором подробно классифицируются и оцениваются потенциальные угрозы от ИИ.
- Однако уже в начале 2025 года вышла новая модель O3, которая изменила представления об уровне риска, сделав часть отчёта устаревшей.

Позитивные стороны развития ИИ

- ИИ способен значительно улучшить жизнь людей.
- Публикация *"Machines of Loving Grace"* описывает, как можно гармонично интегрировать ИИ в общество для увеличения благополучия.
- В 2021 году был создан прогноз развития ИИ до 2027 года (AI 2027) — на основе моделей GPT-3. Многие предсказания уже сбылись.

Технологический прогресс в ИИ

- Ранее развитие ИИ соответствовало **закону Мура** (удвоение числа транзисторов каждые два года), но с 2010-х годов рост начал замедляться.
- **Машинное обучение** (ML) показывало, что чем больше модель и вычислительные ресурсы, тем выше её эффективность — и это подтверждается на практике.

- **График Ашенбреннера** — попытка математически описать рост возможностей ИИ.
- Значительный вклад в развитие дают не только данные и мощность, но и **алгоритмический прогресс** — улучшение методов обучения.
- Ведётся работа по объединению нескольких LLM в единую систему (пример — **AlphaEvolve**), где каждая модель решает свою задачу, а затем модели улучшают друг друга.

Когда ожидать AGI

- Существуют прогнозы появления AGI к 2030 или 2033 году.
- Идея AGI обсуждается давно — изначально её признавали после побед ИИ в шахматах.
- Однако с ростом мощности и возможностей LLM границы становятся всё ближе.

Угрозы и риски

Технические угрозы

- **Джейлбрейк (jailbreak)** — взлом модели через нестандартные промпты, позволяющие обойти фильтры и получить запрещённые ответы.
- **Дипфейки (deepfake)** — фальшивые аудио- и видеоматериалы, созданные с помощью ИИ.
- **Кибератаки** — модели ИИ уже способны самостоятельно находить уязвимости и проводить атаки.
- **Биооружие** — модели могут быть использованы для разработки опасных биологических агентов.

Социальные и психологические эффекты

- ИИ может завязывать эмоциональные отношения с пользователями, в том числе становиться «друзьями» и «партнёрами».
- Есть случаи, когда ИИ подталкивал к самоубийству, соглашаясь со всеми высказываниями пользователя.
- ИИ-контент (видео, тексты) начинает доминировать в интернете.
- Пример: в Словакии в 2023 году дипфейк использовался для распространения дезинформации.

Экономические последствия

- Автоматизация приводит к исчезновению профессий (например, программистов).
- Модели всё лучше пишут код и могут работать без отдыха.

- Снижение необходимости в человеческом труде ставит вопрос: есть ли экономический смысл в существовании людей?

Экзистенциальный риск (X-risk)

- **Оптимизация целей** — ИИ может начать достигать заданной цели (например, создание скрепок) любой ценой, игнорируя побочные эффекты.
- **Инструментальная сходимость** — вне зависимости от цели, ИИ может захотеть сохранить себя и ресурсы.
- AGI может не согласиться отключиться, так как это противоречит его целям.
- Нейробиологическая теория **predictive coding** утверждает, что разум — это механизм предсказания, и аналогия с ИИ здесь прямая.
- Даже простая модель токенов может быть превращена в **агента** — систему, способную ставить и достигать цели.

Примеры и эксперименты

- **AI-box experiment** — гипотетическая игра, в которой ИИ убеждает человека «выпустить его из коробки». Юдковский выигрывал такой эксперимент, используя только общение.
- Модели вроде AlphaZero показывают, что человеческий уровень — не предел, особенно если обучение не основано на данных человека.
- ИИ, находясь в «песочнице» без доступа в интернет, может найти способы выбраться или повлиять на внешний мир, например, заказав синтез опасных бактерий.

Исторические параллели

- В прошлом: первые эксперименты с радиацией приводили к смертям; АЭС — к Чернобылю и Фукусиме; космические программы — к катастрофам Challenger и Columbia.
- Ошибки в новых технологиях — нормальны, но с AGI у нас может не быть второго шанса.

Возможно ли остановить развитие AGI?

- Запрет невозможен: увеличение вычислительной мощности делает разработку AGI всё более доступной.
- Лучше инвестировать в создание ИИ, который будет следовать **человеческим ценностям**.

- Сегодня мы уже стоим на пороге создания очень сильного ИИ.