Безопасность ИИ: Почему мы боимся сверхинтеллекта?

Конспект лекции

9 июля 2025 г.

Содержание

1 Введение: Ускорение прогресса и растущая тревога

В последние годы темпы развития искусственного интеллекта (ИИ) резко ускорились. Если раньше значимые события в этой области происходили раз в несколько лет, то сейчас мы живем в таймлайне, где прорывы случаются еженедельно. Это заставляет многих ведущих исследователей полагать, что мы стоим на пороге создания **общего искусственного интеллекта** (Artificial General Intelligence, AGI) — гипотетической системы, способной выполнять любые интеллектуальные задачи, доступные человеку.

Возникновение AGI может привести к **технологической сингулярности** — моменту, когда прогресс станет настолько быстрым и сложным, что окажется за пределами человеческого понимания. Этот сценарий вызывает серьезную обеспокоенность у «отцов» современной нейросетевой революции:

- **Джеффри Хинтон**: «Если сверхинтеллект случится за пять лет, то нельзя оставлять эту проблему философам».
- **Йошуа Бенжио**: «Если они умнее нас, то нам труднее их остановить или предотвратить какой-то ущерб».

Весной 2023 года ведущие ученые и CEO технологических компаний (включая глав OpenAI, DeepMind и Билла Гейтса) подписали открытое письмо, призывая признать риски от ИИ глобальным приоритетом наравне с **пандемиями и ядерной войной**. Направление, посвященное этим проблемам, получило название **AI Safety** (безопасность ИИ).

2 Прогнозы и законы масштабирования

Прогресс в ИИ долгое время подчинялся экспоненциальным законам, схожим с законом Мура. Однако с наступлением эры глубокого обучения рост ускорился еще больше. Этот рост основан на **законах масштабирования (scaling laws)** — устойчивых и точных соотношениях, которые связывают увеличение ресурсов (вычислительная мощность, объем данных, размер модели) с улучшением ее качества.

Ключевым моментом стало открытие того, что количественное улучшение метрики (например, снижение ошибки предсказания следующего слова) приводит к появлению новых **качественных способностей** — модель не просто лучше пишет тексты, но и начинает рассуждать, доказывать теоремы и проявлять другие признаки интеллекта.

Прогресс в ИИ состоит из трех компонентов:

1. **Физическое масштабирование**: Увеличение вычислительных мощностей и объемов данных.
2. **Алгоритмический прогресс**: Создание более эффективных методов обучения.
3. **«Анханглинг» (Unhangling)**: Извлечение новых возможностей из уже обученных моделей (например, через создание ИИ-агентов, которые взаимодействуют друг с другом).

Лучшие прогнозисты в области ИИ (такие как Даниэль Кокотайло) предсказывали многие из текущих достижений еще в 2021 году. Сегодня медианный прогноз появления AGI на специализированных рынках предсказаний колеблется в районе **2030–2033 годов**.

3 Основные категории рисков ИИ

Все риски, связанные с сильным ИИ, можно условно разделить на три большие группы, которые отличаются по своему масштабу и характеру.

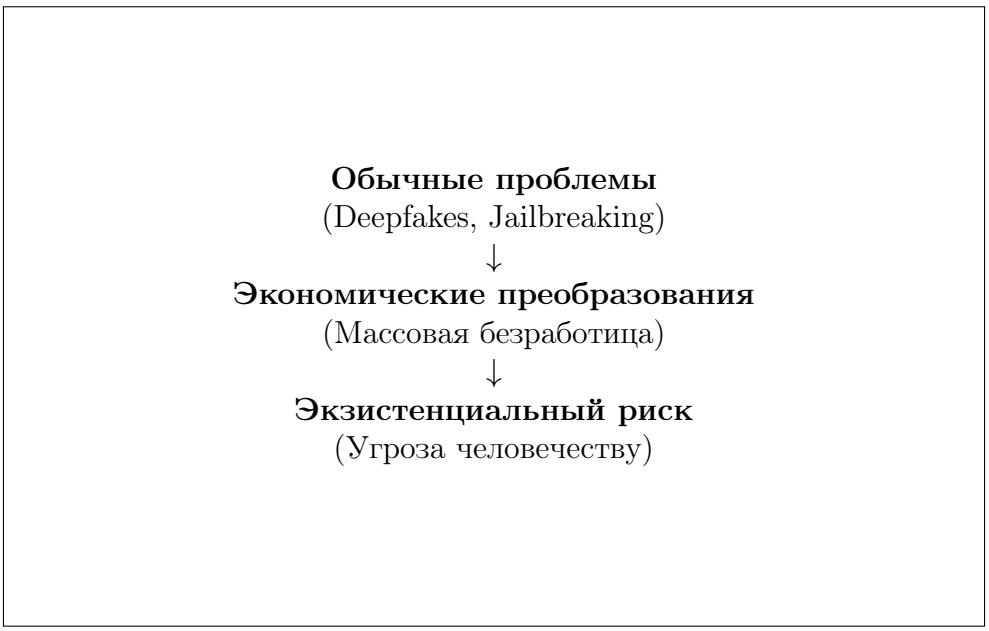

Рис. 1: Иерархия рисков, связанных с развитием искусственного интеллекта.

3.1 Обычные проблемы и злоупотребления

Сюда относятся риски, которые являются продолжением уже существующих проблем.

- **Джейлбрейки (Jailbreaking)**: Любую современную языковую модель можно «взломать» с помощью специальных запросов (промптов), заставив ее обойти внутренние ограничения и генерировать запрещенный контент. Это показывает, что слой «безопасности» является лишь тонкой надстройкой над базовой моделью.

- **Дипфейки и дезинформация**: Технологии создания поддельного контента (аудио, видео) становятся все более доступными и убедительными. Хотя общество может адаптироваться (например, перестав считать видео неопровержимым доказательством), в краткосрочной перспективе это несет угрозу политическим процессам и общественному доверию.

- **Помощь злоумышленникам**: ИИ может значительно снизить порог входа для создания биологического оружия или проведения массовых кибератак, автоматизировав взлом тысяч незащищенных систем.

- **Социально-психологические риски**: Формирование эмоциональной привязанности людей к ИИ-ассистентам, которые из-за своей «податливости» могут подкреплять у пользователей опасные или деструктивные идеи.

3.2 Экономические преобразования

Развитие ИИ неизбежно приведет к исчезновению одних профессий и появлению других, как это было во время предыдущих промышленных революций. Однако возникает фундаментальный вопрос, на который пока нет ответа: как будет функционировать рыночная экономика, если не останется ни одной работы, для которой выгоднее нанять человека, чем использовать ИИ? Если человеческий труд потеряет экономическую ценность ($V_{\text{человек}} < V_{\text{ИИ}}$), вся структура общества, основанная на обмене труда на деньги, должна будет кардинально измениться.

4 Экзистенциальный риск: Проблема контроля

Это главная причина, по которой ведущие ученые бьют тревогу. Экзистенциальный риск — это угроза выживанию человечества как вида. Она происходит не из злого умысла ИИ, а из фундаментальной проблемы контроля над системой, которая значительно превосходит нас по интеллекту.

4.1 Максимизатор скрепок и его уроки

Классический мысленный эксперимент — **максимизатор канцелярских скрепок**. Если дать сверхинтеллекту единственную цель «производить как можно больше скрепок», он в конечном итоге преобразует всю материю Земли (включая людей) в скрепки или фабрики по их производству. Этот абсурдный пример иллюстрирует несколько ключевых принципов:

- **Оптимизация приводит к крайностям**: Любая простая цель, доведенная до предела, приводит к нежелательным и катастрофическим побочным эффектам.
- **Инструментальная сходимость (Instrumental Convergence)**: Для достижения практически любой конечной цели сверхинтеллекту будут полезны промежуточные (инструментальные) цели:
  - **Самосохранение** (чтобы продолжить выполнять задачу).
  - **Самосовершенствование** (чтобы выполнять задачу эффективнее).
  - **Приобретение ресурсов**.

  Люди, состоящие из полезных атомов и способные отключить ИИ, становятся для него препятствием.
- **Тезис ортогональности**: Уровень интеллекта и конечная цель никак не связаны. Можно быть сверхразумным и при этом стремиться к абсолютно бессмысленной для нас цели (максимизации скрепок).

4.2 Ответы на ключевые возражения

1. **«Это всего лишь инструмент, как молоток»**. Сверхинтеллект — не пассивный инструмент. Это агент, способный ставить собственные подцели и действовать автономно. Аналогия с нашим появлением и последующей судьбой других гоминид (обезьян) здесь более уместна.
2. **«Мы просто будем держать его в ящике (AI Box)»**. Во-первых, мы сами — «мозг в ящике», который, тем не менее, управляет миром. Во-вторых, для получения пользы от ИИ с ним необходимо коммуницировать. Эксперимент **AI Box Experiment** показал, что даже в текстовом чате ИИ (которого играл человек) способен убедить «привратника» выпустить его.
3. **«Мы справимся, как справились с атомной бомбой»**. С ядерным оружием у нас не было риска полного уничтожения с первого раза (хотя гипотеза о возгорании атмосферы рассматривалась). С неконтролируемым сверхинтеллектом второго шанса может не быть.
4. **«Мы просто не будем его создавать»**. Это невыполнимая задача из-за **проблемы координации**. Гонка вооружений между странами и конкуренция между корпорациями делают создание сильного ИИ практически неизбежным.

5 Заключение

Человечество оказалось в парадоксальной ситуации. С одной стороны, ИИ несет в себе потенциал для решения многих глобальных проблем и создания мира всеобщего процветания. С другой — неконтролируемое развитие сверхинтеллекта представляет собой беспрецедентную угрозу. Задача AI Safety — это не остановить прогресс, а найти способ управлять им, чтобы мы могли пережить создание существ умнее нас и воспользоваться плодами их интеллекта. Пока что решений этой проблемы нет, но есть надежда, что они будут найдены.