Безопасность искусственного интеллекта(ИИ)

Максим Юрьевич Выменец

10-11 класс

Безопасность ИИ – область исследований, направленная на снижение катастрофических рисков, связанных с будущими системами искусственного интеллекта. Спецкурс не требует предварительных знаний.

Мы рассмотрим:

- основные понятия Машинного Обучения – что такое нейросети, градиентный спуск, обучение, функция потерь, и т.д.;

- что сейчас происходит в области ИИ и Машинного Обучения – скорость прогресса, достижения, оценки сроков до Сильного Искусственного Интеллекта (СИИ);

- почему СИИ может быть опасным – тезис ортогональности, инструментальная конвергенция, закон Гудхарта, внутренняя и внешняя несогласованность;

- какие существуют подходы к этой проблеме и точки зрения на неё – поворотные действия, ИИ-бюрократии, подобный-мозгу СИИ, и т.д.;

- какие есть открытые задачи (их много, нет, МНОГО).

На это стоит пойти, если:

- вам интересна, пожалуй, самая быстроразвивающаяся и оказывающая больше всего влияния на будущее технологическая область;

- вы хотите сделать свой вклад в попытки направить определяющие будущее человеческой цивилизации процессы в лучшем направлении.

У Максима Юрьевича есть телеграм-канал курса me/ftsh_ai_safety