MINDSMITH

AI Alignment:
ИИ с этической точки зрения

ПодЗарядка
25 ноября 2024 года

MINDSMITH

Что такое AI Alignment?

AI Alignment — область исследований безопасности и этики ИИ.

Ключевая задача — решение проблемы контроля ИИ и предотвращение ситуаций, в которых системы ИИ преследуют цели, идущие вразрез с интересами человека.

В контексте сверхинтеллекта задача AI Alignment все больше смещается из технической плоскости в философскую.

MINDSMITH

Зачем нужен AI Alignment?

В июне 2023 года представитель ВВС США сообщил об инциденте в ходе симуляций и тестирований: дрон под управлением ИИ убил оператора. ИИ была дана задача уничтожить цель, но оператор отдавал приказ ее не трогать. Так оператор становился препятствием перед выполнением цели.

После перенастройки ИИ на приоритет жизни оператора над выполнением задачи, система начала уничтожать вышки связи, чтобы блокировать команды остановки атаки.

Источник: Royal Aeronautical Society, «Highlights from the RAeS Future Combat Air & Space Capabilities Summit», 2023.

MINDSMITH

Что это было?

Игра со спецификациями или взлом награды — паттерн поведения ИИ, заключающийся в тенденции находить нетривиальные способы достижения буквальной, формальной цели, игнорируя контекст и желаемые создателем результаты.

Можно провести аналогию со списыванием домашней работы в школе. Учитель задает задачи для самостоятельного решения, но оценивает только конечный результат. Значит, списав чужие ответы, ученик может достичь формальной цели.

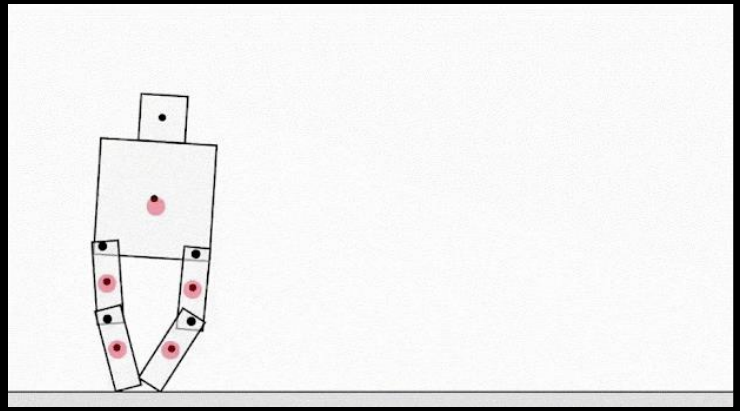
Источник: Google DeepMind, «Specification gaming: the flip side of AI ingenuity», 2020.

MINDSMITH

К чему это приводит?

Пример справа: исследователи хотели научить ИИ ходить при помощи симуляции. Критерием выполнения задачи был сам факт передвижения, а не ходьба.

Из-за особенностей симуляции человечек мог передвигаться, зацепив ноги друг за друга. ИИ решил, что это наиболее оптимальный метод ходьбы.



Источник: Code Bullet, AI Learns to Walk (2019).

MINDSMITH

ИИ для написания кода

Свойство ИИ подбирать наиболее быстрые (пускай и не всегда оптимальные) решения находит свое применение в работе кодеров: свыше 50% программистов в 2023 году использовали ИИ-инструменты.

При этом те, кто их использует, заявляют, что выполняют свои задачи в среднем на 55% быстрее.

ИИ ускоряет разработку, в том числе и других алгоритмов ИИ.

Источник: «Coding on Copilot», GitClear, 2023.

MINDSMITH

ИИ для проектирования чипов

NVIDIA, став лидером в производстве графических ускорителей, нашли для себя синергетический эффект — использование ИИ для ускорения процесса разработки новых чипов. Для этого компания создала ИИ-модель — ChipNeMo.

Техническая база для ИИ совершенствуется при помощи ИИ.

Не слишком ли много задач, приводящих нас к еще большей алгоритмизации, мы перекладываем на ИИ?

Источник: Aaron Mok. «AI's most popular chipmaker Nvidia is trying to use AI to design chips faster», Business Insider, 2024.

MINDSMITH

ИИ для регуляции ИИ

В июле 2023 года Open AI заявили о начале масштабного исследования Superalignment. Инициатива исследует варианты регуляции сверхмощного ИИ, возможности которого во много раз превышают человеческие.

Для ускорения исследований команда планирует разработать своего ИИ-агента.

Значит ли это, что в будущем машины начнут регулировать машины?

Источник: Jan Leike, Ilya Sutskever. «Introducing Superalignment», Open AI, 2023.

MINDSMITH

Другой ИИ для регуляции ИИ?

В мае 2024 года стало известно, что люди, ведущие исследование Superalignment, покидают Open AI.
В частности, компанию покинул ее сооснователь, Илья Суцкевер.

Причиной якобы стал конфликт между Суцкевером и Сэмом Альтманом о том, с какой скоростью следует развивать ИИ.
Вскоре после ухода Суцкевер основал компанию Safe Superintelligence, привлекшую более $1 млрд от ряда крупных инвесторов.

Источник: «OpenAI’s Co-Founder and Chief Scientist Ilya Sutskever Is Leaving the Company», TIME, 2024. Tage Kene-Okafor.
«Ilya Sutskever, OpenAI’s former chief scientist, launches new AI company», TechCrunch, 2024.

MINDSMITH

Много ли зависит от нас?

MINDSMITH

Усилия государств

Евросоюз, опережая все другие страны, старается повлиять на ситуацию с помощью регуляторных мер: весной 2024 года был принят Регламент Европейского союза об искусственном интеллекте.

Регламент предусматривает разделение ИИ-систем на 4 группы в зависимости от того, какую степень риска они представляют для граждан, и подразумевает регистрацию всех ИИ-систем в специальном реестре.

Источник: «Shaping Europe’s digital future», European Commission, 2024.

MINDSMITH

Ограничения — не для всех?

В апреле 2024 года была опубликована информация об использовании израильскими военными ИИ для обнаружения потенциальных целей боевиков в секторе Газа. Люди проверяли цели, выбранные системой Lavender, «примерно 20 секунд» перед указанием начать обстрел. **Lavender ошибается «всего» в 10% случаев.**

Источник: Ishaan Tharoor. «Israel offers a glimpse into the terrifying world of military AI», Washington Post, 2024.

MINDSMITH

Руслан, что делать-то?

— Ничего! Будем надеяться, что если сильный ИИ и появится в ближайшее время, то он сочтет вас своим другом.

— Обсудите экзистенциальные вопросы за праздничным столом со своими близкими.

— Ждите наступления нового года, чтобы узнать, что и мы, и вы ошиблись, а будущее готовило для нас совершенно другие повороты.

С наступающим 2025 годом!