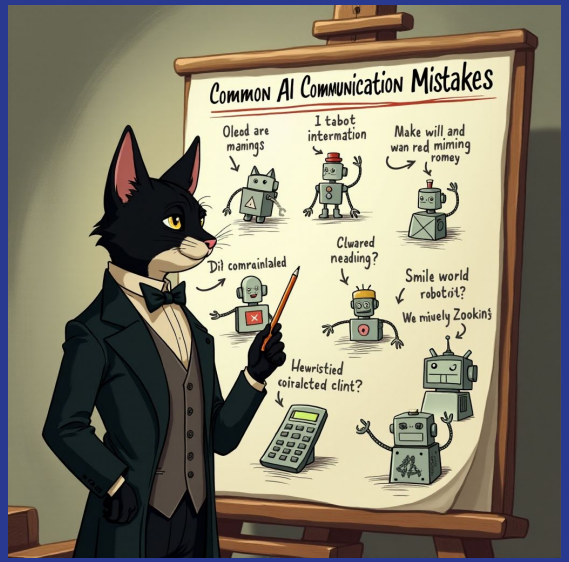
Как не надо рассказывать об искусственном интеллекте (пожалуйста!)

Сергей Николенко

28 февраля 2025



План

- Что такое “искусственный интеллект”?
- AI — не монолитная вещь
- Что же такое “искусственный интеллект”?
- О чём говорят и не говорят
- И всё-таки: что такое “искусственный интеллект”?
- Позитивные выводы

Что такое “искусственный интеллект”

- У нашей области очень неудачное название :(
- Этого мы не изменим, но не надо усугублять проблему!

Что такое “искусственный интеллект”

- У нашей области очень неудачное название :(
- Этого мы не изменим, но не надо усугублять проблему!

Искусственный интеллект научился разговаривать с котиками...

Что такое “искусственный интеллект”

- У нашей области очень неудачное название :(
- Этого мы не изменим, но не надо усугублять проблему!

Искусственный интеллект научился разговаривать с котиками...

Плохо

Новая нейросеть обучилась разговаривать с котиками...

Лучше

Что такое “искусственный интеллект”

- У нашей области очень неудачное название :(
- Этого мы не изменим, но не надо усугублять проблему!

Искусственный интеллект научился разговаривать с котиками...

Плохо

Новая нейросеть обучилась разговаривать с котиками...

Лучше

Разработана новая модель, которая может разговаривать с котиками...

Хорошо

Что такое “искусственный интеллект”

- У нашей области очень неудачное название :(
- Сравните:

Таблетка теперь лечит болезнь Альцгеймера...	**Плохо**
Новая таблетка может вылечить болезнь Альцгеймера...	**Лучше**
Разработано новое лекарство, которое может справиться с болезнью Альцгеймера...	**Хорошо**

AI — не единый монолит

- Нет никакого одного “искусственного интеллекта”
- Это большая область науки, в которой тысячи разных подходов для самых разных задач

Учёные применили искусственный интеллект...

AI — не единый монолит

- Нет никакого одного “искусственного интеллекта”
- Это большая область науки, в которой тысячи разных подходов для самых разных задач

~~Учёные применили искусственный интеллект...~~

Плохо

Учёные разработали новую модель искусственного интеллекта...

Хорошо

AI — не единый монолит

- Нет никакого одного “искусственного интеллекта”
- Сравните:

Учёные применили медицину...

Плохо

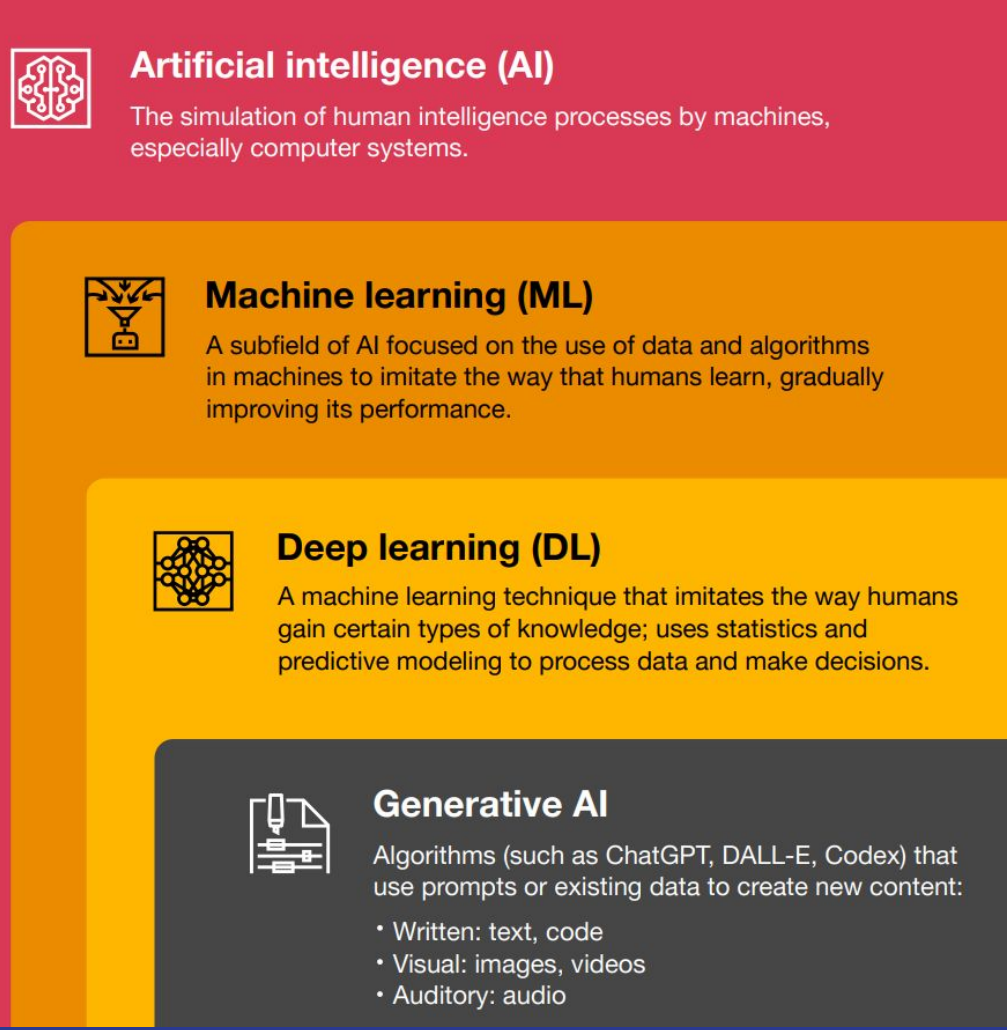
Учёные разработали новое лекарство...

Хорошо

Что же такое AI?

- Это раздел компьютерных наук, который развивается с начала 1950-х годов:
  - с начала 1990-х AI – это в основном машинное обучение (machine learning, ML)
    - с середины 2000-х ML – это в основном глубокое обучение (deep learning, DL)

с начала 2023-го мы слышим в основном о порождающих и особенно больших языковых моделях (large language models, LLM)



Но подождите, а как же AGI

- Да, “искусственный интеллект” — это не только раздел науки, но и его цель, продукт, возможно, даже конкретная модель
- Чтобы его отличить, лучше называть его “сильным” или “общим” искусственным интеллектом (artificial general intelligence, AGI)
- AGI – это общий искусственный интеллект человеческого уровня, который по определению сможет делать то же, что и человек (хотя бы когнитивно)
- Его пока не существует, но…

Многие беспокоятся

If super intelligence will happen in 5 years time, it can't be left to philosophers to solve this... Maybe we are just a passing stage in the evolution of intelligence.

If they're smarter than us, then it's hard for us to stop these systems or to prevent damage... You could say I feel lost.

Stop it, Eliezer. Your scaremongering is already hurting some people. You'll be sorry if it starts getting people killed.

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Signatories:

☑ AI Scientists ☑ Other Notable Figures

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic

Dawn Song
Professor of Computer Science, UC Berkeley

Ted Lieu
Congressman, US House of Representatives

Bill Gates
Gates Ventures

Ya-Qin Zhang
Professor and Dean, AIR, Tsinghua University

Ilya Sutskever
Co-Founder and Chief Scientist, OpenAI

Shane Legg

Многие беспокоятся

If super intelligence will happe
time, it can't be left to phi
this... Maybe we are i
the evolution

I
u
da
hard for
prevent
I feel lost.

Stop it, Eliezer. Your scaremongering is already hurting some people. You'll be sorry if it starts getting people killed.

И вот об этом, кстати, стоило бы говорить чаще...

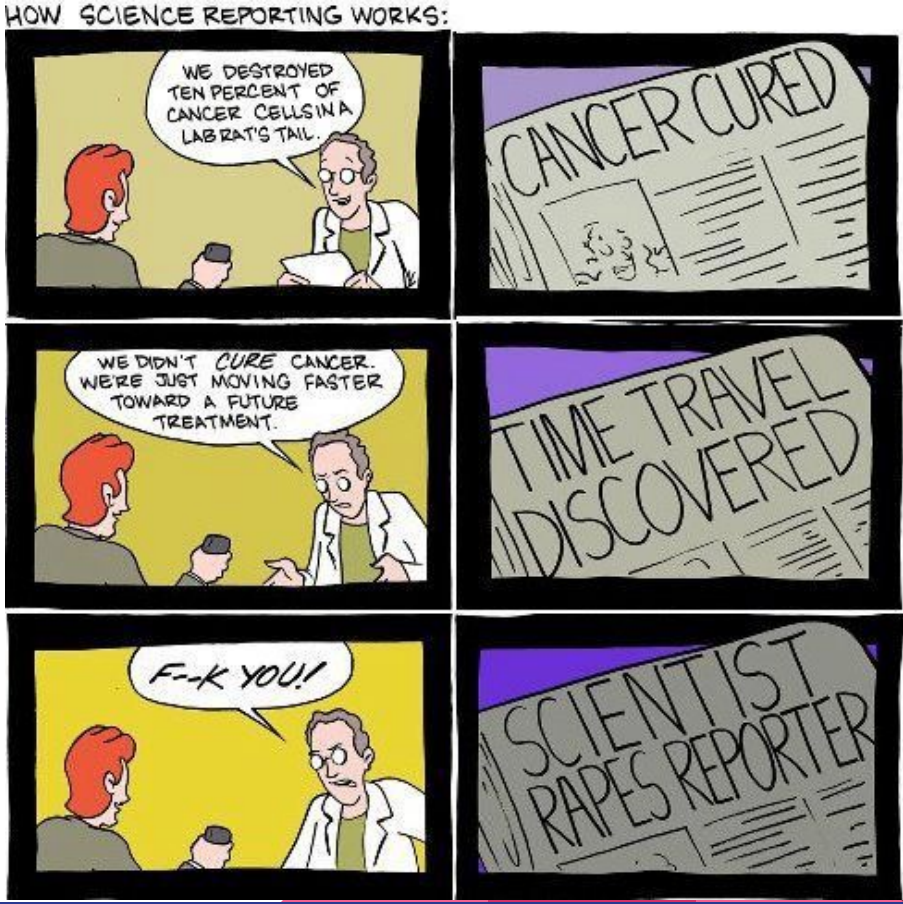
risk of extinction from
global priority alongside
ale risks such as
uclear war.

Другие частые ошибки

- Антропоморфизация AI
  - пока AGI нет, антропоморфизировать некого, да и когда будет, мы ничего не будем знать о его самосознании
  - “ИИ полагает”, “ИИ утверждает”, “ИИ исследует” – это тоже антропоморфизация; это может быть разумным сокращением для конкретной модели: “o1-pro исследовал”, “Claude Sonnet 3.7 полагает”
- AI – это просто калькулятор, “stochastic parrot”
  - ну да, а человек – это набор нейрончиков, которые обмениваются электрическими сигналами

Другие частые ошибки

- AI всех заменит, а потом убьёт всех человеков
  - это важные проблемы! человечество их ещё не решило
  - хорошо бы привлекать людей ими заниматься
- Чрезмерное упрощение
  - как на классической иллюстрации



Ещё примеры

Плохо

Искусственный интеллект научился диагностировать рак лучше врачей

Хорошо

Новая модель компьютерного зрения помогает радиологам выявлять признаки опухолей

Ещё примеры

Плохо

ИИ решил математическую задачу, с которой не справлялись учёные

Хорошо

С помощью машинного обучения исследователи решили математическую задачу

Ещё примеры

Плохо

Искусственный интеллект теперь понимает эмоции людей

Хорошо

Новая модель компьютерного зрения распознаёт эмоции людей с точностью до 85%

Итого — как рассказывать про AI лучше

- Используйте точные термины
  - говорите о конкретных технологиях, моделях или системах вместо абстрактного "искусственного интеллекта"
- Избегайте антропоморфизации
  - описывайте возможности систем через их функции, а не человеческие качества
- Подчёркивайте роль человека
  - акцентируйте внимание на том, что модели AI создаются, обучаются и управляются людьми
- Объясняйте ограничения
  - честно говорите о том, чего система не может делать и почему, приводите конкретные примеры применения
- Лучше всего, конечно, рассказывать об идеях
  - но я понимаю, что это обычно невозможно

Спасибо за внимание!