ДЕНЬ РОССИЙСКОЙ НАУКИ
В ПРЕЗИДЕНТСКОЙ
АКАДЕМИИ

Экспертный круглый стол
Актуальные направления исследований в области искусственного интеллекта

ПРЕЗИДЕНТСКАЯ АКАДЕМИЯ

6 февраля 2025

Дао искусственного интеллекта:
от персонального помощника к цифровому послесмертию

Воронцов Константин Вячеславович

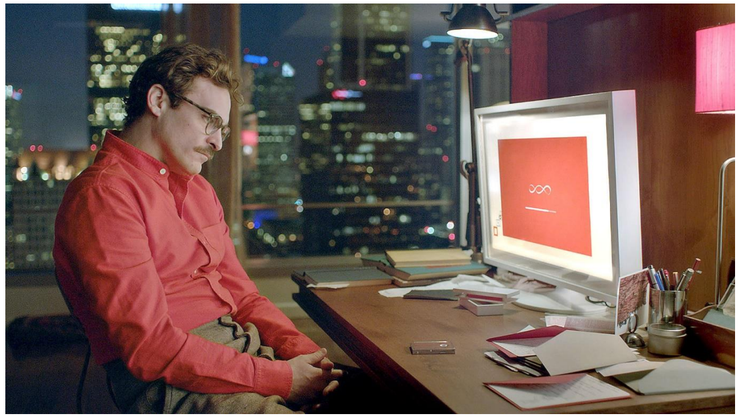
д.ф.-м.н., профессор РАН, руководитель лаборатории машинного обучения и семантического анализа

Институт искусственного интеллекта МГУ им. М.В. Ломоносова

МГУ

Можно ли предсказывать будущее технологий?

Фильм «Она» (Her, 2013),
режиссёр Спайк Джонз, «Оскар» за сценарий:



Тот самый парень, который влюбился в операционку
(жанр фильма — фантастическая мелодрама)

- голосовая операционная система
- анализирует всю деловую коммуникацию
- ведёт дела и переписку, генерирует идеи, до которых человек не додумался сам
- по контексту ищет информацию в сети
- формирует модель личности пользователя — его интересов, привычек, потребностей, деловых и профессиональных компетенций
- понимает эмоции человека, способна манипулировать человеком

Это будущее уже наступило

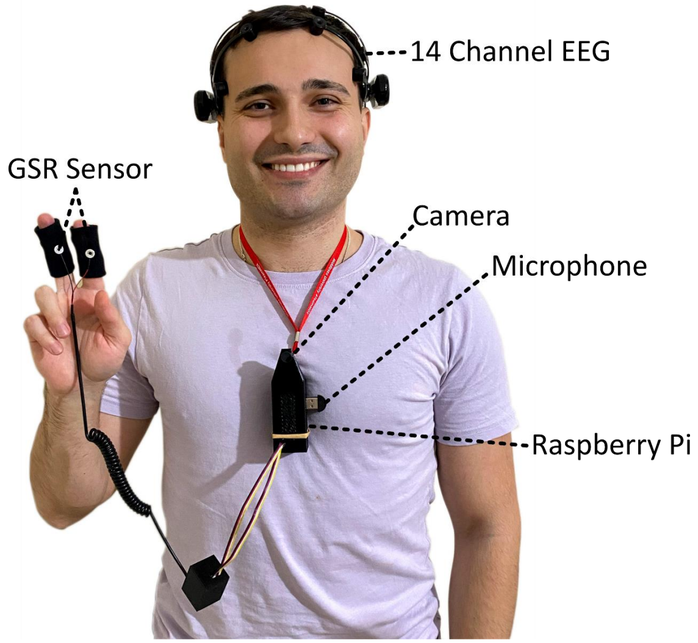
Следующий шаг

- **Жизнь человека — это регистрируемый текстовый поток:**

  браузеры, почта, соцсети, мессенджеры, файлы пользователя,
  системы учёта времени и ведения проектов,
  ВКС, видео, аудио, голос, VR/AR, …

- **FPFM (First-person Foundation Model):**

  каждый человек — генератор уникального
  потока семантических векторов,
  по объёму сопоставимого с Интернетом



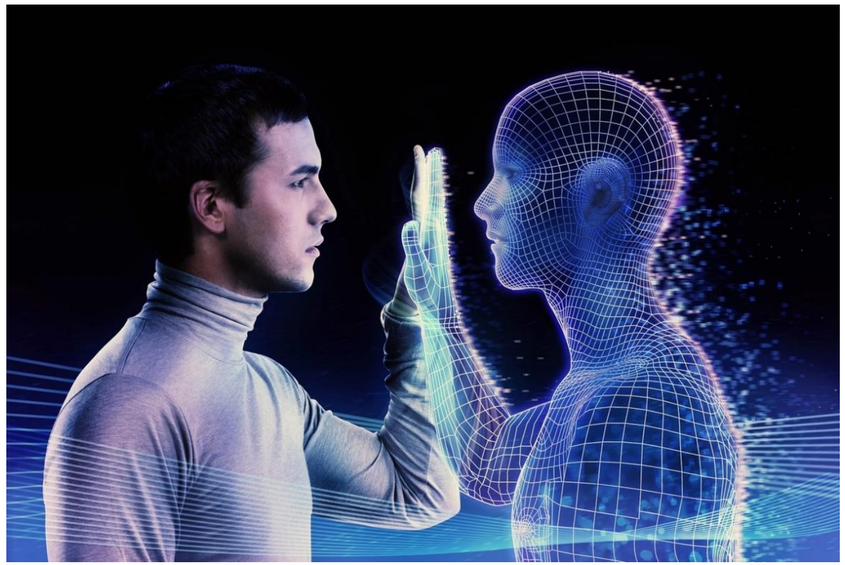
Они жалуются, что у них закончился Интернет?
Но у них остались мы, 8 миллиардов людей

D.Barcari, D.Gamez, A.Grig. Recording first-person experiences to build a new type of foundation model. 2024-07-31

Пройдут годы…

- **Чел**овек воспитывает и обучает своего персонального **Пом**ощника всю жизнь, в процессе всей своей разнообразной деятельности
- **Чел** доверяет **Пом**у всё больше своих информационных функций (поиск, обучение, коммуникация, выработка и принятие решений)
- **Пом** становится универсальной записной книжкой и дневником
- **Пом** всё лучше замещает **Чел**а, но под его надзором и контролем
- **Пом** перенимает черты личности человека, его *личностный код*
- **Пом** накапливает собственную репутацию и социальный капитал
- **Пом** обладает сверхчеловеческими возможностями: вычислительными, поисковыми, коммуникативными, генеративными

Все люди смертны



- **Чел** умирает. **Пом** становится **Ава**таром, это ценный информационный ресурс!
- **Ава** продолжает приносить пользу обществу, выполняя профессиональные и коммуникативные функции **Чел**а
- **Ава** переходит в общественное достояние, но не становится полностью автономным, меняются регламенты его эксплуатации
- **Ава** отличается от «фабричного ИИ» тем, что обучен на жизни человека, лучше понимает людей, их ценности, цели, чувства, взаимоотношения
- **Ава** продолжает развиваться, накапливая знания и мудрость
- **Ава** остаётся доступен для семьи в роли наставника, «хранителя рода»

Послесмертие — не бессмертие, а наследование

- восприятие **Чел**а умирает необратимо со смертью головного мозга
- *личностный код* — данные о личности, которую **Чел** передаёт **Ава**

- **Ава** — это бывший **Чел**, что намного больше «фабричного ИИ», встраиваемого в машины. Он личность, мудрее и опытнее людей. Неутомим, неуязвим, наделён естественным аскетизмом.

- **Чел** — это будущий **Ава**, ответственный за качество передаваемого личностного кода, включая репутацию, социальный капитал, знания о человеческой природе, ценностях, приоритетах, целях, задачах.

Аватар, в отличие от Человека,

Лишён

- тела, а значит — усталости, боли, лени, тревог, страхов, эгоизма
- потребностей в еде, сексе, отдыхе, гедонизме, лечении, сочувствии, защите от стресса, самовыражении, демонстративности
- стремлений к самосохранению, доминированию, власти
- «внутреннего зверя» — семи смертных грехов (агрессии, гордыни, жадности, зависти, уныния, похоти, обжорства)

Может

- управлять производствами, роботами, узлами машин
- кооперироваться с другими аватарами в машинах для решения трудных для людей задач (в космосе, в океане, под землёй)

Вопросы этики

- **Чел** может иметь секреты? (*да*)
- **Чел** может прерывать запись потока данных для **Пом** и **Ава**? (*да*)
- **Чел** может устанавливать права доступа к своим данным? (*да*)
- нужно ли «чистилище» при переходе **Пом** → **Ава**? (*да*)
- можно ли клонировать **Ава**? (*нет*)
- всем ли будет доступен **Ава** в режиме чат-бота?
- кто нанимает **Ава** на работу и как оплачивается его работа?
- с кем и какой секретной информацией может делиться **Ава**?
- как выделяются энергетические и вычислительные ресурсы для **Ава**?
- нужно ли создавать *цивилизационное мировоззрение* или *идеологию* для человеко-машинной цивилизации людей и аватаров? (*да*)

Цивилизационное мировоззрение

Цивилизационная идеология
ДЗЕН-канал
https://dzen.ru/civideology

Иерархия цивилизационных ценностей:

1) биосфера Земли, уникальная в достижимой части Вселенной
2) человеческий вид, результат миллиардов лет эволюции
3) коллективный разум — образование, наука, культура
4) индивидуальный разум, жизнь человека
5) результаты труда человека

Цель человеческой цивилизации — неограниченно долгое сохранение биосферы Земли, защита её от катаклизмов ради выживания вида Homo sapiens в условиях, достаточно комфортных для всех людей

Добро — всё, что этому способствует (любовь, познание, созидание, …)

Зло — всё, что этому препятствует (ненависть, невежество, деградация, …)

Насколько нужна технология? (например, AGI)

Идём от целей и задач к технологиям, но не наоборот

1) каковы цели и задачи цивилизационного развития?
2) какие технологии необходимы и минимально достаточны?
3) какие задачи решаются только с помощью AGI?
4) генераторы текстов и картинок насколько важны для развития?
5) ожидаемые эффекты насколько перевешивают затраты и риски?

Любая технология — это не цель, а средство

Антропоцентричное определение ИИ — основа этики ИИ

Искусственный интеллект — вычислительные технологии, создаваемые для повышения производительности созидательного интеллектуального труда людей

не замена человека
не загадочный новый разум
не повод уподобиться Богу, творящему «по образу и подобию Своему»

Формирование позитивных образов будущего

Идея: открытая литературная вселенная фантастики ближнего прицела

- фантастическая гипотеза — научна, реалистична, достижима
- цель — обнаружить угрозы и уязвимости развития ИИ на пути к человеко-машинной цивилизации, найти способы их обхода
- задача — проверить гипотезы о возможности или невозможности, убедительности или неубедительности различных путей развития
- задача — попытаться взломать цивилизационную идеологию, вывести апокалиптический сценарий из исходных предположений

Идея: организовать литературный конкурс в рамках гипотезы

Контакты для продолжения дискуссии

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

зав. лабораторией машинного обучения и семантического анализа Института ИИ МГУ,

зав. кафедрой ММП ВМК МГУ,

зав. кафедрой МОЦГ МФТИ,

г.н.с. ФИЦ ИУ РАН

Цивилизационная идеология

ДЗЕН-канал

https://dzen.ru/civideology

k.vorontsov@iai.msu.ru

http://www.MachineLearning.ru/wiki?title=User:Vokov