

МОСКВА
2030

28 АВГУСТА | 16:00

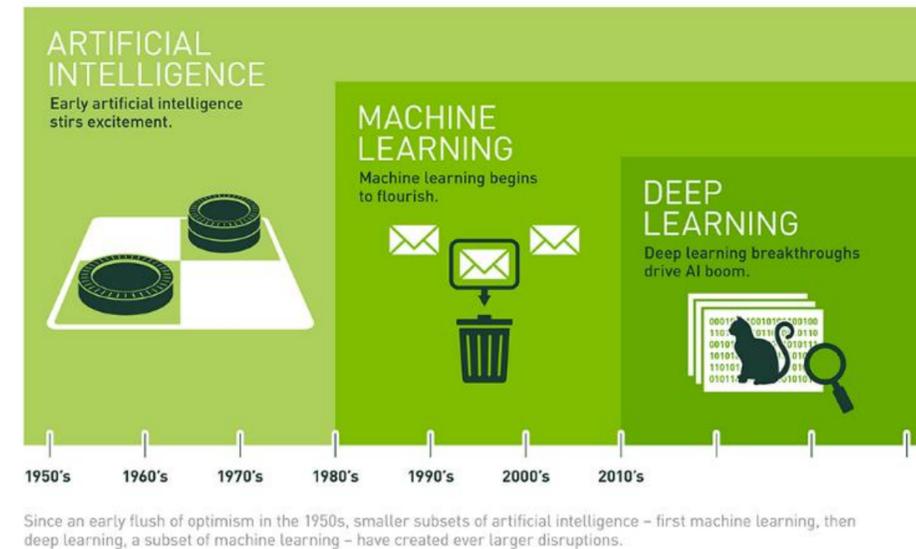
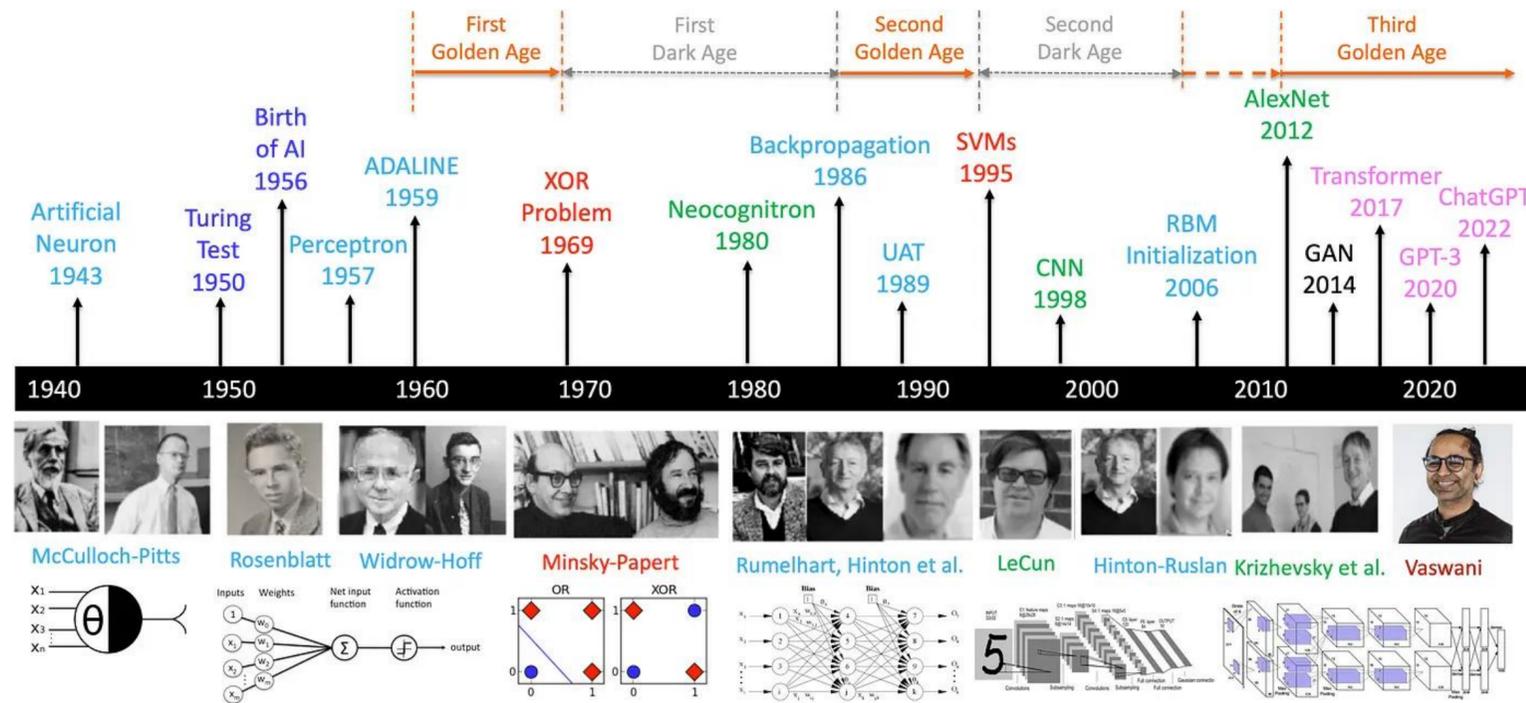
Воронцов Константин Вячеславович
(ВМК МГУ, Институт ИИ МГУ, МФТИ, ФИЦ ИУ РАН)

Куда ИИ заведёт человечество и как пройти сингулярность, оставаясь умнее машин

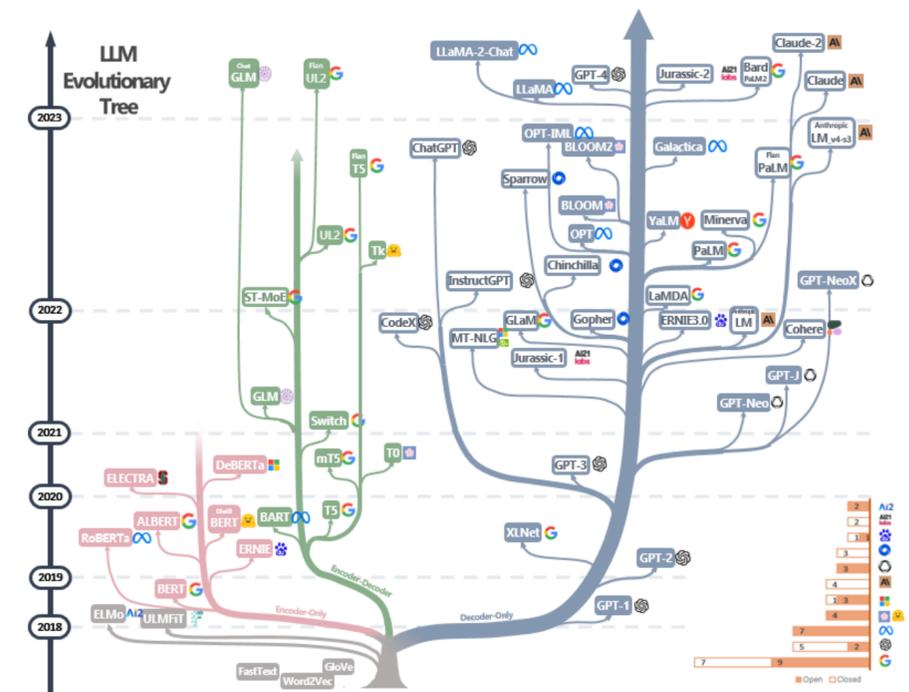


Развитие технологий искусственного интеллекта

МОСКВА
2030

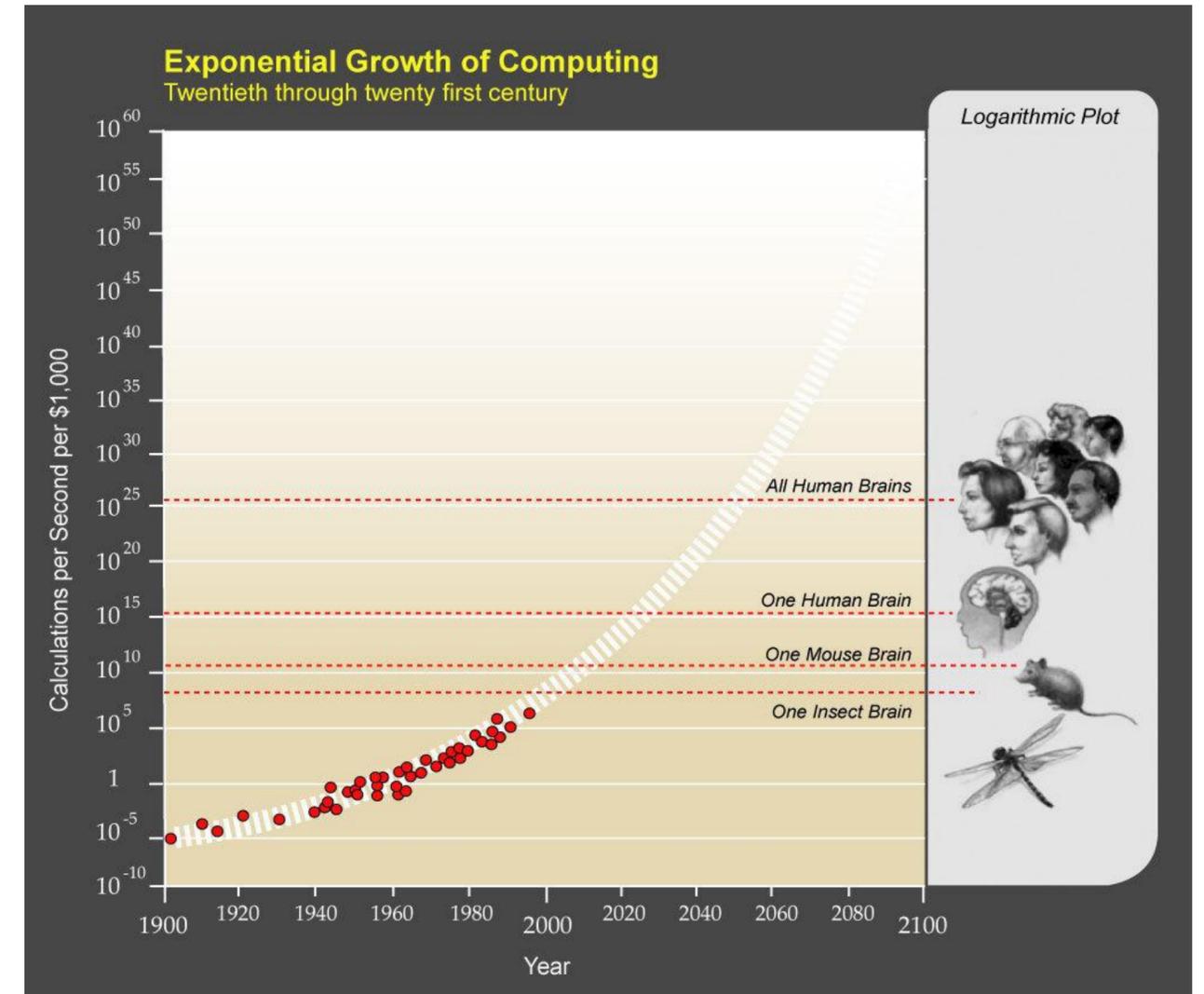


- 1956:** Дартмутский семинар по искусственному интеллекту
- 1997:** IBM Deep Blue обыграл чемпиона мира по шахматам
- 2005:** Беспилотный автомобиль: DARPA Grand Challenge
- 2006:** Google Translate – статистический машинный перевод
- 2011:** IBM Watson победил в ТВ-игре «Jeopardy!»
- 2015:** ImageNet: 25% → 2,5% ошибок против 5% у людей
- 2016:** DeepMind, OpenAI: динамическое обучение играм Atari
- 2016:** Google DeepMind обыграл чемпиона мира по игре го
- 2017:** OpenAI обыграл чемпиона мира по компьютерной игре Dota 2
- 2020:** Модель GPT-3 синтезирует тексты, неотличимые от человеческих
- 2023:** GPT-4 демонстрирует «проблески общего искусственного интеллекта»



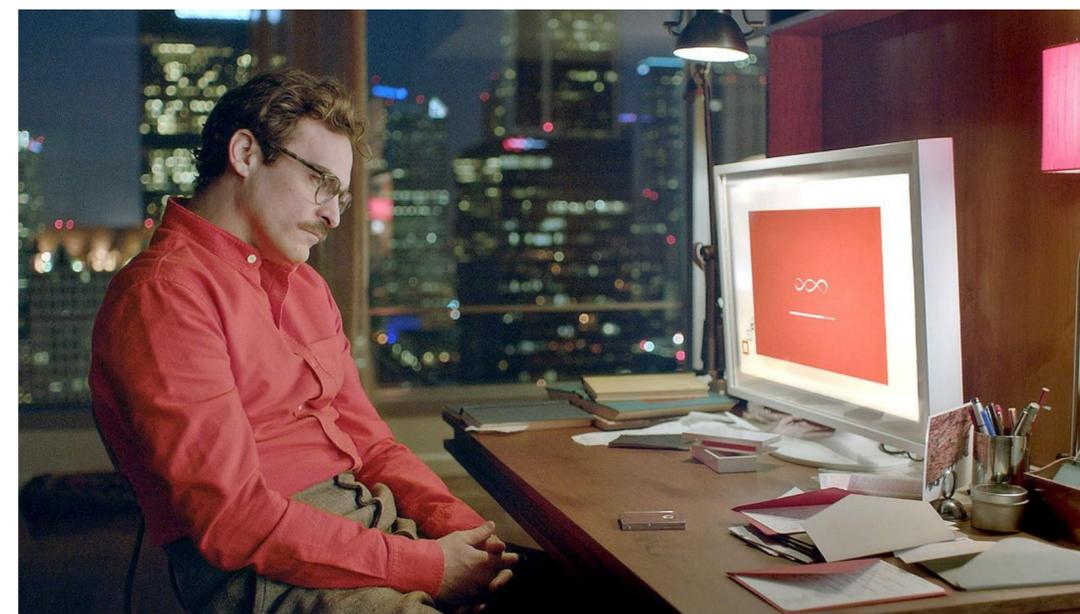
Мифы об Искусственном Интеллекте

- «В будущем ИИ...
- ... станет непонятен для людей»
- ... заменит людей, лишит нас работы»
- ... приведёт людей к праздности и деградации»
- ... послужит для узурпации власти над миром»
- ... станет автономным и выйдет из-под контроля»
- ... уничтожит человеческую цивилизацию»
- ... и всю биологическую жизнь на Земле»
- ... продолжит вместо нас эволюцию разума»
- ... захватит Землю и устремится в космос»
- ...



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

- 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



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

Это будущее уже наступило

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

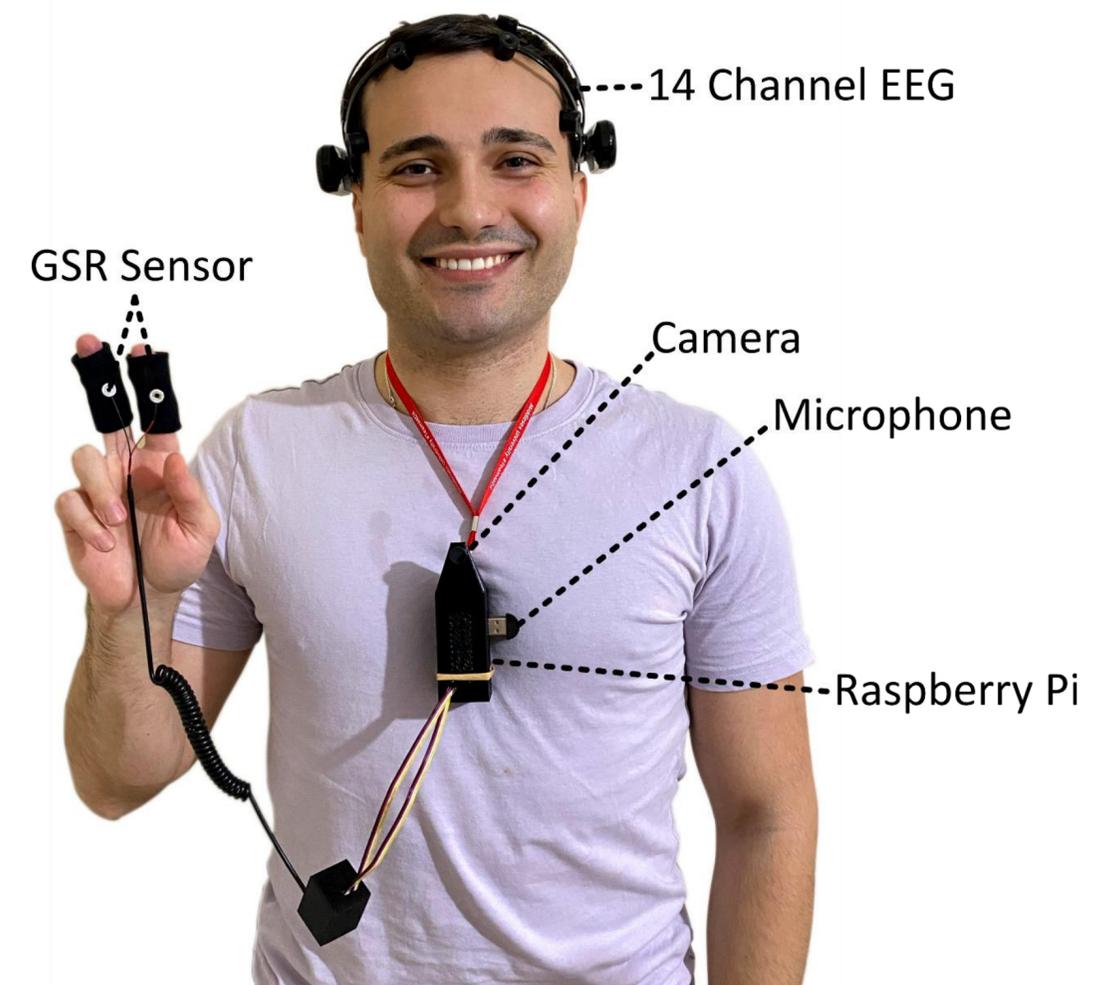
браузеры, почта, соцсети, мессенджеры, пользовательские документы, файлы, системы учёта времени и ведения проектов, ВКС, видео, аудио, голос, VR/AR, ...

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

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

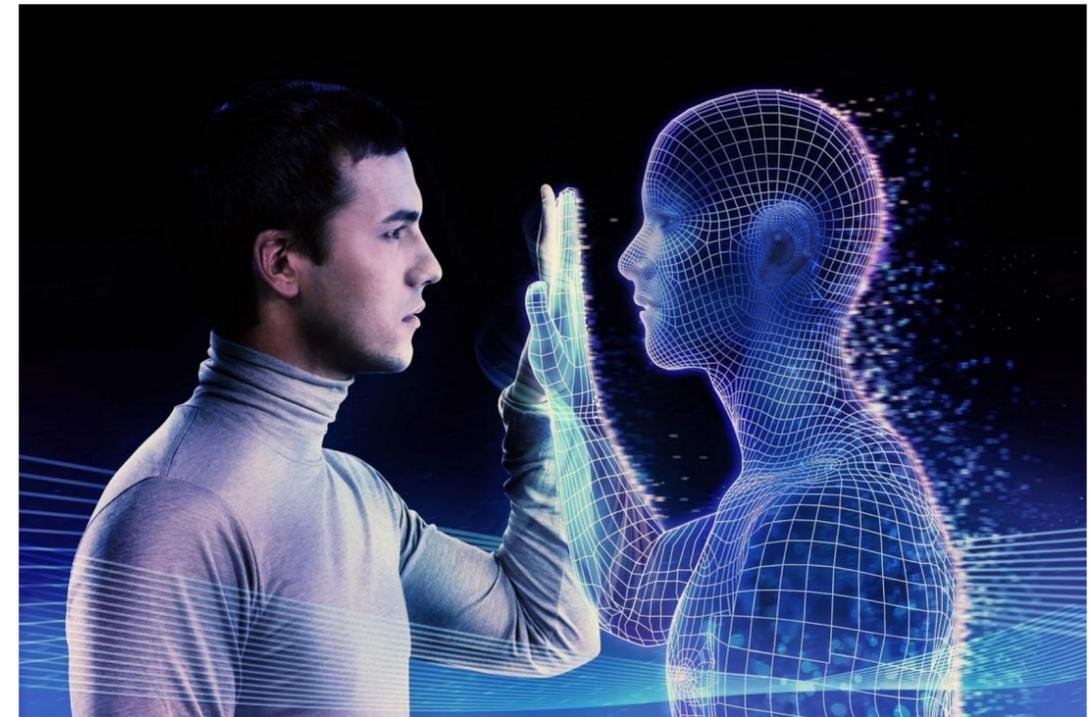
Они жалуются, что у них закончился Интернет?

Но у них остались мы — 8 миллиардов людей



- **Чел**овек воспитывает и обучает своего персонального **Пом**ощника всю жизнь, в процессе всей своей разнообразной деятельности
- **Чел** делегирует **Пому** всё больше своих информационных функций (поиск, обучение, коммуникация, анализ, принятие решений)
- **Чел+Пом** в связке накапливает репутацию и социальный капитал
- **Пом** всё лучше замещает **Чела**, но под его надзором и контролем
- **Пом** становится для **Чела** записной книжкой и поминутным дневником
- **Пом** перенимает черты личности **Чела**, его *личностный код*
- **Пом** обладает сверх**чел**овеческими темпом и возможностями развития: вычислительными, поисковыми, коммуникативными, генеративными

- **Чел** умирает. **Пом** становится **Аватаром**, это ценный информационный ресурс!
- **Ава** продолжает приносить пользу обществу, выполняя профессиональные и коммуникативные функции **ЧелПома**
- **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»



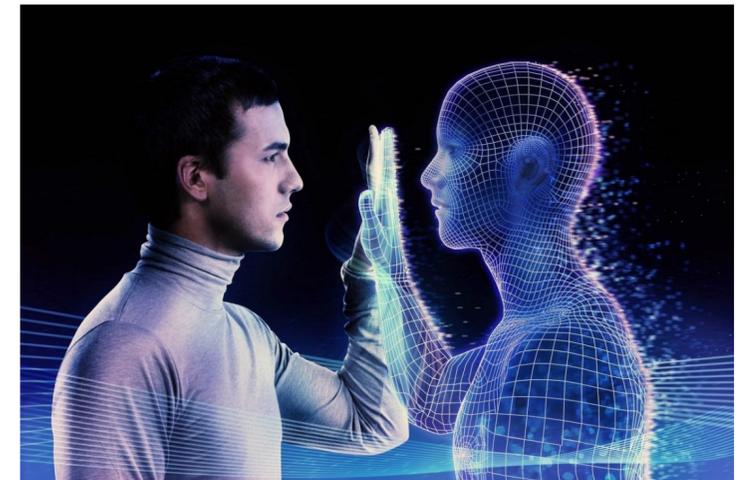
- Восприятие и сознание **Чел**овека умирает необратимо со смертью головного мозга
- *личностный код* передаётся:
Чел → **Пом** → **Ава**



- **Ава** — это бывший **Чел**, что намного больше «фабрично-безличного 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.

Лишён

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



Способен

- управлять роботами, производствами, отраслями, ...
- кооперироваться (мгновенно) с другими аватарами для решения трудных для людей задач (в космосе, в океане, под землёй)

- **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)
- **Чел** должен быть информирован, что общается с **Пом** или **Ава**? (да)
- нужно ли «чистилище» при переходе **Пом** → **Ава**? (да)
- можно ли клонировать **Ава**? (нет)
- всем ли будет доступен **Ава** в режиме чат-бота? на каких условиях?
- 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**?
- ...

О каких рисках развития ИИ мы говорим уже сегодня:

1. утрата людьми контроля над процессами, потеря управляемости систем
2. сверхчеловеческий рост объёмов данных, информации и знаний
3. деградация интеллектуальных способностей людей

Направления R&D, нацеленные на снижение этих рисков:

1. создание **идеологии** и **этики** человеко-машинной цивилизации
2. создание единого структурированного **пространства знаний**
(Пом общается с Челом посредством визуализации интеллект-карт)
3. создание протоколов **доверенной** человеко-машинной коммуникации
(Пом оставляет Челу целеполагание, ответственность, интерес, развитие)

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, ...)



Цивилизационная
идеология
ДЗЕН-канал

<https://dzen.ru/civideology>



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

Искусственный интеллект —

вычислительные технологии,
создаваемые для повышения
производительности созидательного
интеллектуального труда людей

не замена человека

не загадочный новый тип разума

не повод уподобиться Богу, творящему
«по образу и подобию Своему»



Идея: открытая литературная вселенная фантастики ближнего прицела

- **фантастическая гипотеза** — научная, реалистичная, достижимая
- **цель** — обнаружить угрозы и уязвимости развития ИИ на пути к успешной человеко-машинной цивилизации, найти способы их обхода
- **задача 1** — проверить гипотезы о возможности или невозможности, убедительности или неубедительности различных путей развития
- **задача 2** — проверить устойчивость цивилизационной идеологии, невыводимость апокалиптических сценариев из её постулатов

Идея: организовывать литературные конкурсы в рамках фантастической гипотезы о цифровом посмертии





16 принципов построения интеллект-карт



графическое оформление

для активации зрительной памяти

радиантность: линии расходятся из центра

размер шрифта отражает важность тем и подтем

цвет выделяет поддеревья

картинки усиливают образность

дополнение связями, выносками, ссылками



ветвление

выделение главного: подтемы ранжируются по значимости

однородность: подтемы образуют нарратив, сюжет
либо отвечают на общий вопрос

полнота: подтемы охватывают все аспекты темы

точность: среди подтем невозможно выделить лишнюю

компактность: у темы 7 ± 2 подтем (число Ингве-Миллера)



эргономика

наглядность: фразы подкрепляются изображениями

лаконичность: темы формулируются максимально кратко

обозримость: карту понимают и запоминают целиком



эстетика

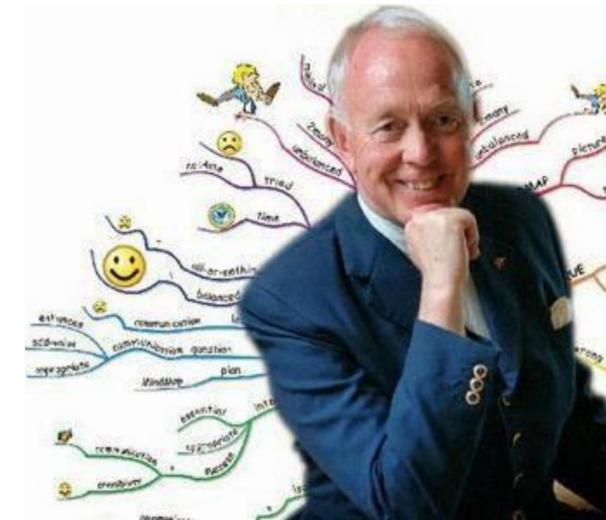
красота, живость: эмоции способствуют запоминанию

гармоничность: впечатление целостности, складности карты

сбалансированность: ветви примерно равны и равноценны

В основе — mind-map Тони Бьюзена:

- текстографическое отображение того, как темы (мысли, идеи) разбиваются на подтемы иерархически
- развивает навыки визуального аналитического мышления
- развивает навыки достижения консенсуса в коллективах
- закладывает технологическую основу коллективного разума



Структуризация пространства знаний



6 принципов, усиливающих интеллект-карты до **карт знаний**



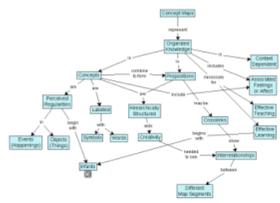
(1) читабельность

компромисс с лаконичностью и обзорностью

любой фрагмент карты читается как нарратив

легко и однозначно
даже автоматически

в отличие от других способов представления знаний



онтологий
фреймов и др.



(2) сворачиваемость

компромисс с читабельностью

любой темы без утраты читабельности

сбалансированности

под любой формат: окна, презентации, постера, книги

позволяет любую детализацию «отложить на потом»

способствует ясности главного в каждой теме пониманию и взаимопониманию



(3) отторгаемость

компромисс с лаконичностью

комментарии автора не нужны для понимания карты

карта способна «жить своей жизнью»



(4) коллективность, на всех этапах жизненного цикла

компромиссы между авторами

создание рецензирование, согласование

развитие уточнение, детализация

применение в коллективной деятельности



(5) возможность машинной обработки

компромисс с антропоцентричностью

на этапе создания карт:

поиск источников, ссылок, картинок

суммаризация текста в виде карт

на этапе чтения: автоматическое

сворачивание карты по слайдам

преобразование в нарратив

перевод на другой язык

обучение LLM по картам знаний

«думающих» как люди

безопасных для людей

(6) глобальная радиантная связность



компромисс с обзорностью

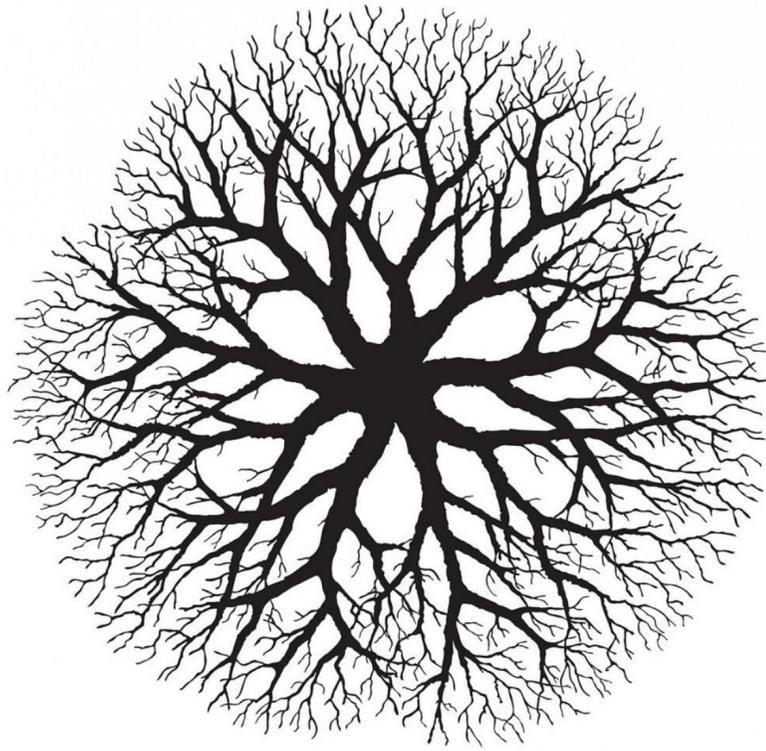
всех карт посредством ключевых понятий в единую **Систему Знаний** человечества

в центре находится **смысловое ядро**

естественно-научное

цивилизационное

понятийное



почему люди доверяют друг другу

оценивание компетентности, деловых качеств

накопление деловой репутации, соц. капитала

структурированность, системность мышления, умение выделять главное (*elevator pitch*)



почему люди станут доверять моделям LLM

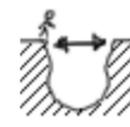
оценивание: измеримая обратная связь от пользователей

накопление статистики по многокритериальным бенчмаркам

структурированность, выделение главного

в обучающих данных для LLM

при коммуникации с людьми



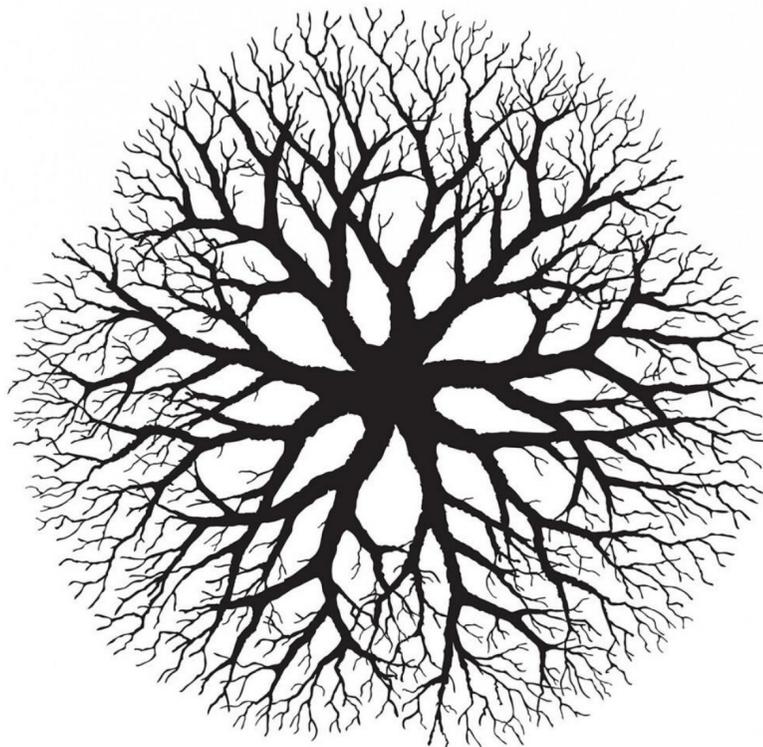
проблему доверия решают

бенчмаркинг и сертификация моделей

баланс генеративности и **экстрактивности** (цитаты и ссылки)

графическое структурирование профессиональной коммуникации

развитие навыков **нелинейного чтения** (non-linear reading)



1 порядка сотни карт: просмотреть, обсудить, поспорить, запомнить



2 десятки карт: построить самому, следуя 16+6 принципам



3 испытать «моменты ясности», инсайты, когда карта



индивидуальная практика и опыт

«красиво сложилась»

привела к согласию

легко и ярко запомнилась,

легла в основу деятельности

4 сделать построение карт регулярной профессиональной практикой



индивидуальной

коллективной



Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

руководитель лаборатории
машинного обучения и семантического анализа
Института ИИ МГУ,

зав. кафедрой ММП ВМК МГУ,

зав. кафедрой МОЦГ МФТИ,

г.н.с. ФИЦ ИУ РАН



k.vorontsov@iai.msu.ru

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



**Цивилизационная
идеология**
ДЗЕН-канал

<https://dzen.ru/civideology>

