

Футурология искусственного интеллекта: от персонального помощника к цифровому посмертию

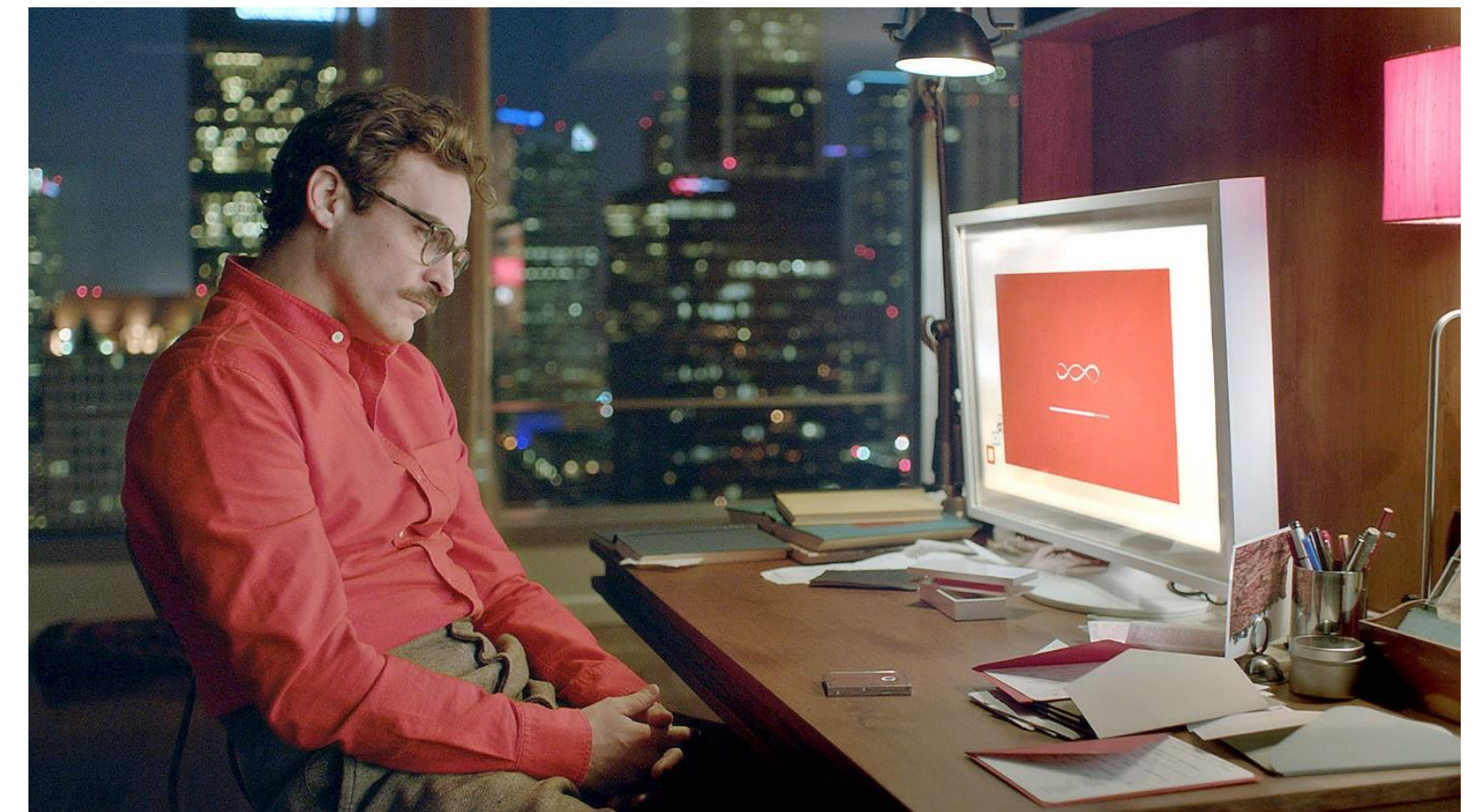
ВОРОНЦОВ КОНСТАНТИН ВЯЧЕСЛАВОВИЧ
K.VORONTSOV@IAI.MSU.RU

*д.ф.-м.н., профессор РАН,
зав. лаб. машинного обучения и семантического анализа Института ИИ МГУ,
зав. каф. математических методов прогнозирования ВМК МГУ,
зав. каф. машинного обучения и цифровой гуманитаристики МФТИ*

Возможно ли предсказывать развитие технологий?

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)

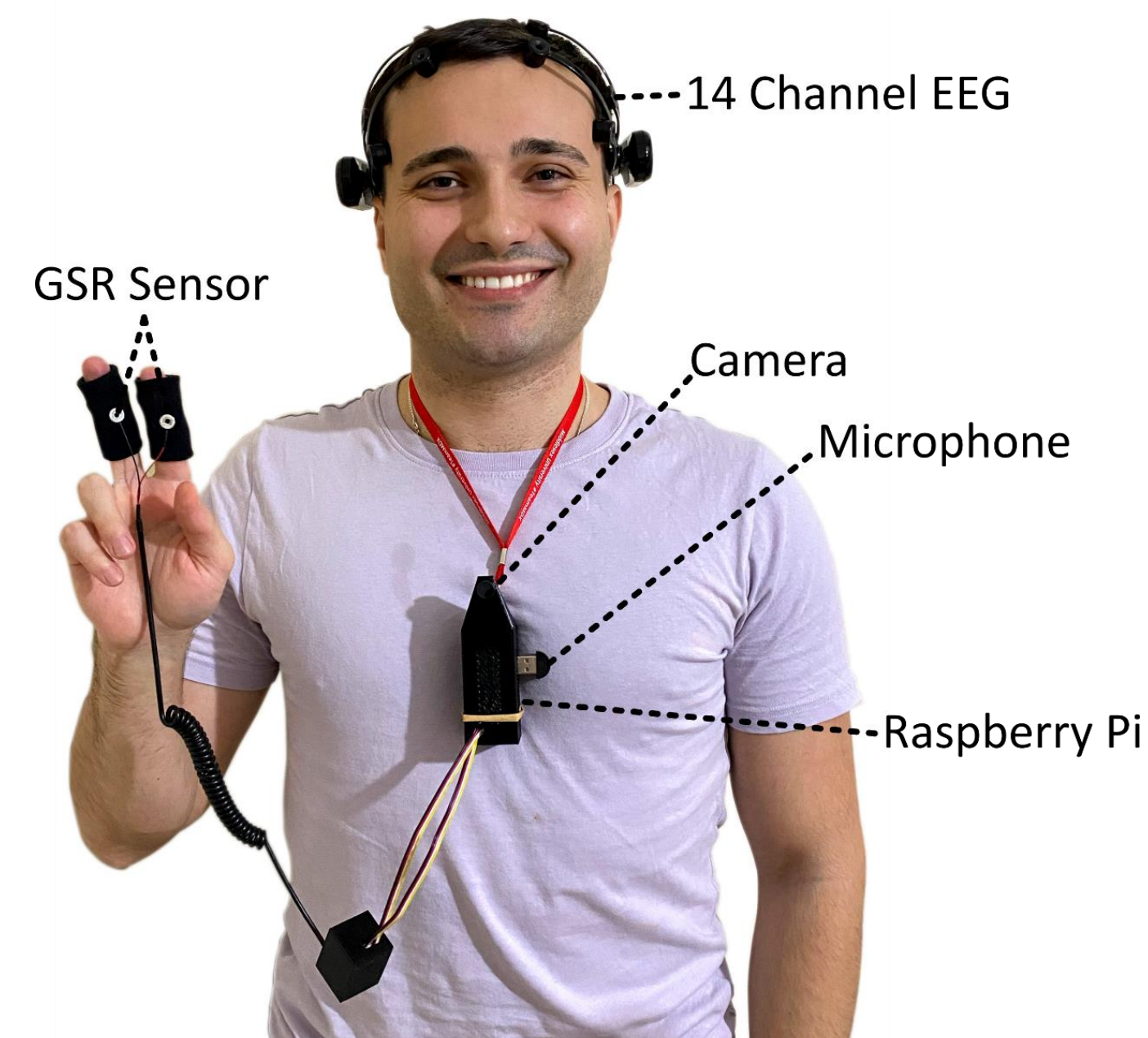
Это будущее уже наступило

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



Пройдут годы...

- **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

Очеловеченный ИИ — путь к снижению рисков ИИ

О каких рисках развития ИИ мы говорим уже сегодня:

- сверхчеловеческий рост объёмов данных, информации и знаний
- утрата людьми понимания процессов и управления процессами
- деградация интеллектуальных способностей людей

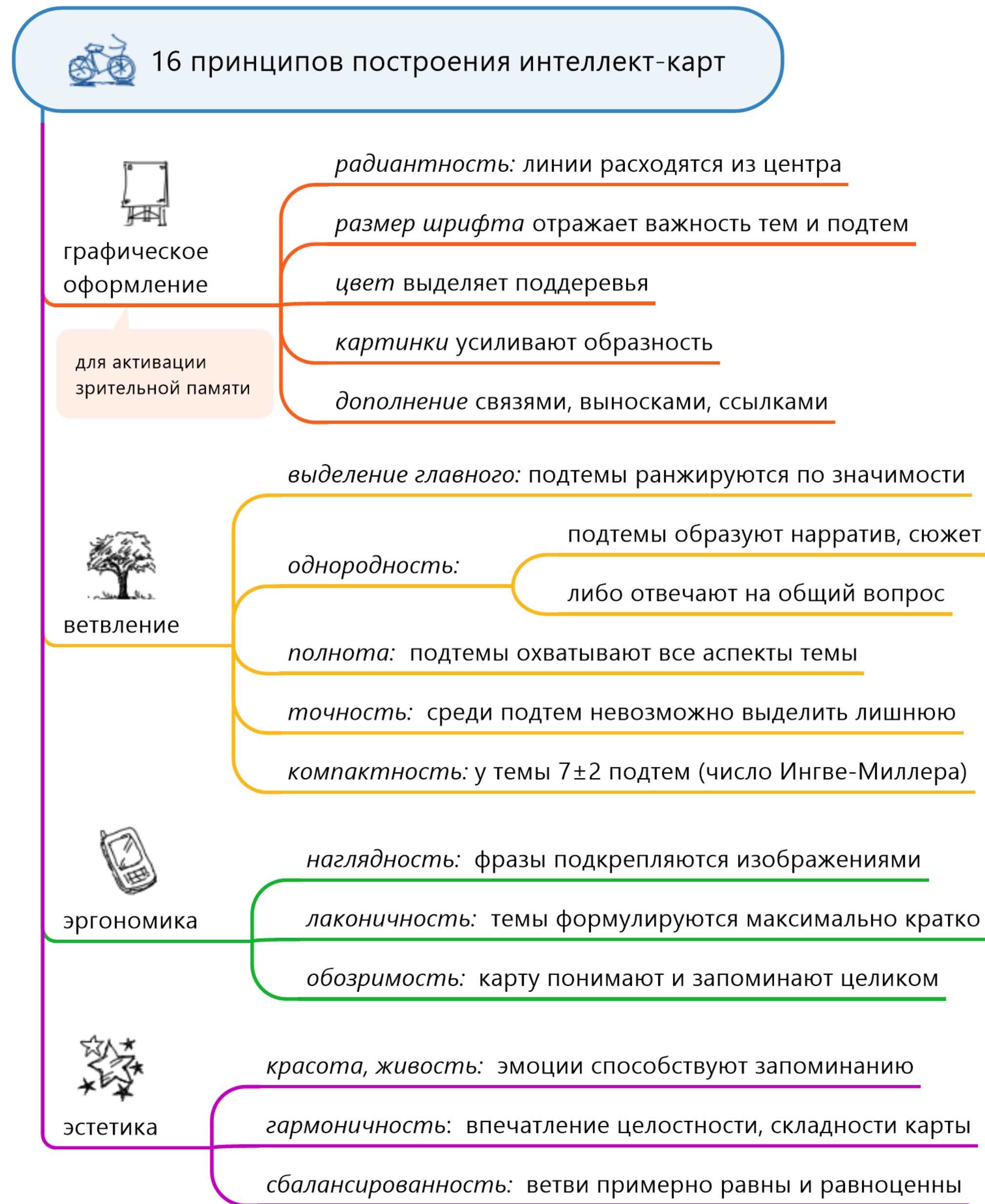
Направления R&D, нацеленные на снижение этих рисков:

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

Структурированное пространство знаний

В основе — mind-map:

- 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



Структурированное пространство знаний

6 принципов, усиливающих интеллект-карты до карт знаний



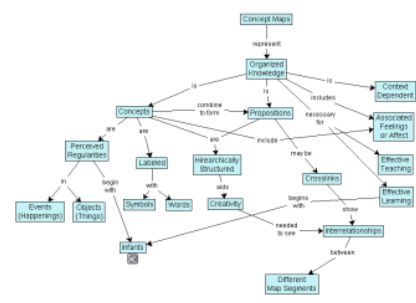
(1) читабельность

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

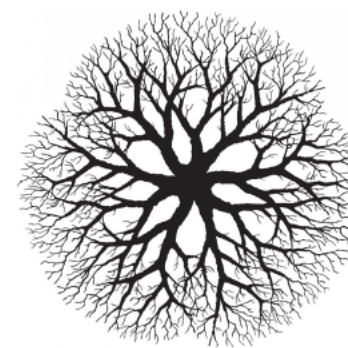


(5) возможность машинной обработки

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

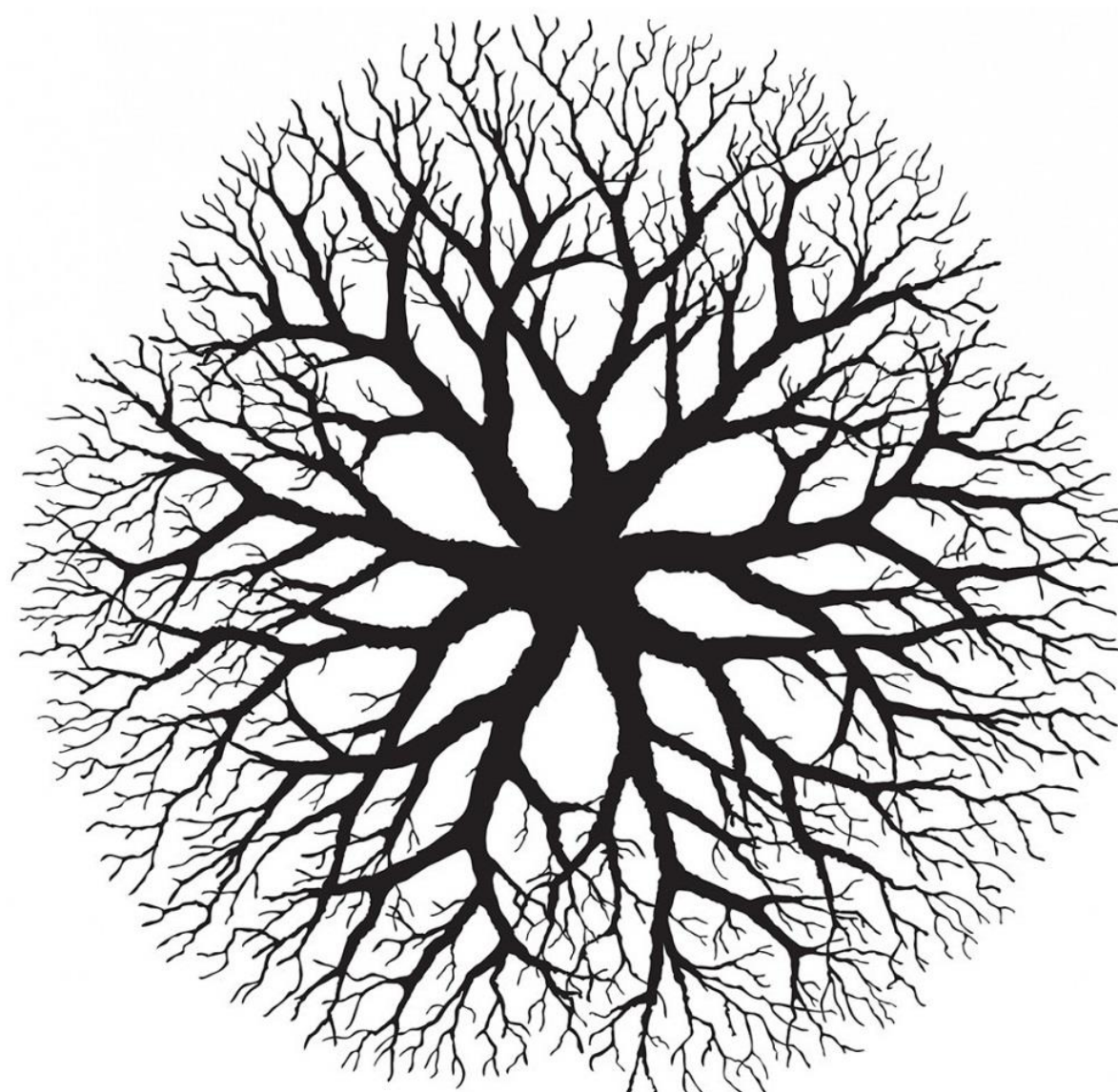


компромисс с обзорностью

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

в центре находится **смысловое ядро**
естественно-научное
цивилизационное
понятийное

Доверенная человеко-машинная коммуникация



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

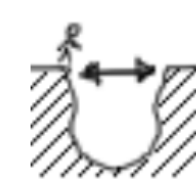
оценивание: измеримая обратная связь от пользователей

накопление статистики по многокритериальным бенчмаркам

структурированность, выделение главного

в обучающих данных для LLM

при коммуникации с людьми



проблему доверия решают

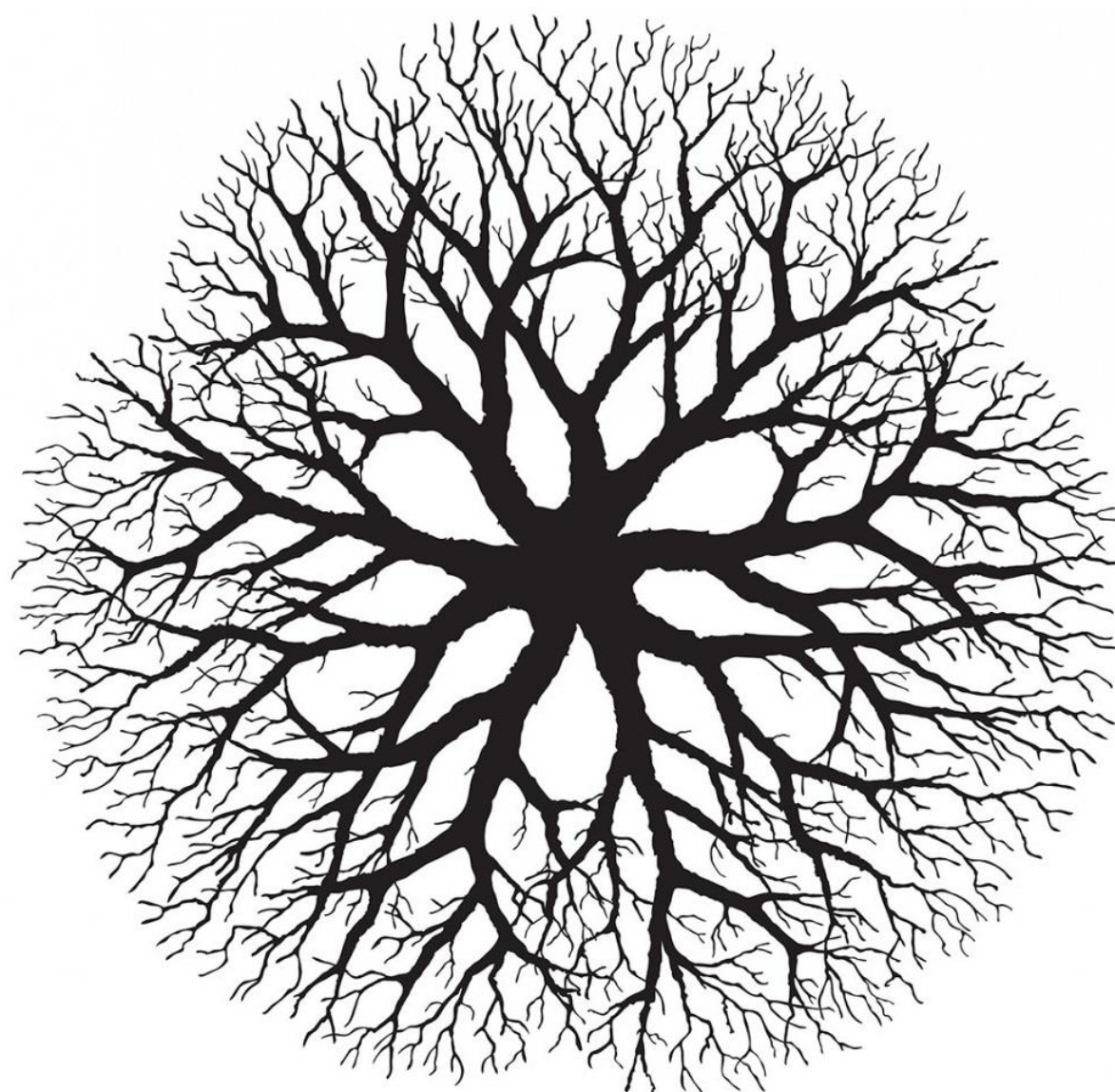
бенчмаркинг и сертификация моделей

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

графическое структурирование профессиональной коммуникации

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

Активация визуального аналитического мышления



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

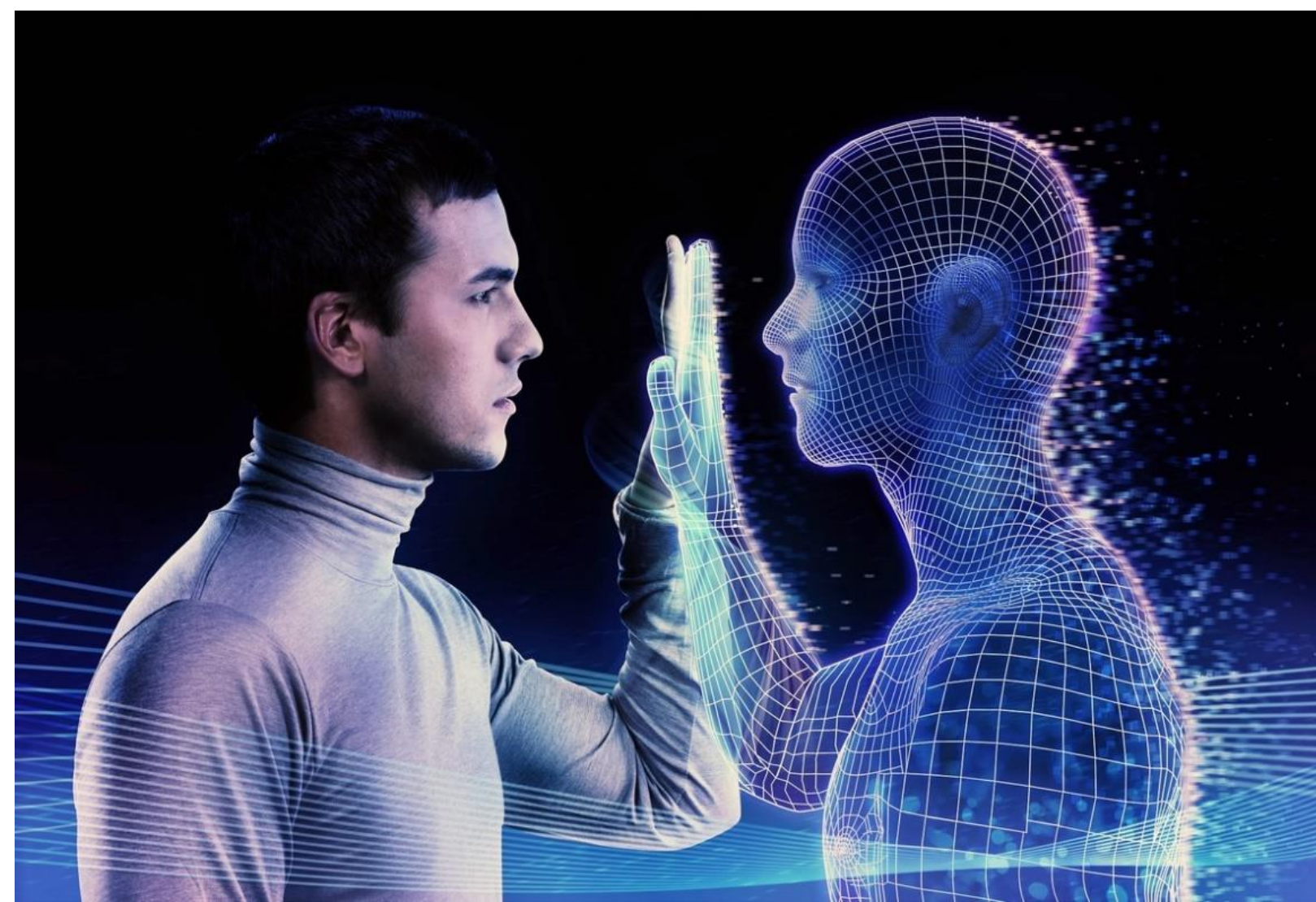


индивидуальной

коллективной

Все люди смертны

- **Чел** умирает. **Пом** становится **Ава**таром, 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»



Послесмертие — не бессмертие, а наследование

- 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.

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 (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)

Некоторые 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*)
- **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**?
- пора ли создавать *цивилизационное мировоззрение и идеологию* для человеко-машинной цивилизации людей и аватаров? (*да*)

Постулаты цивилизационной идеологии

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>



Какие технологии развивать? (на примере 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

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»**



Формирование позитивных образов будущего

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



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

k.vorontsov@iai.msu.ru

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



Цивилизационная
идеология

ДЗЕН-канал

<https://dzen.ru/civideology>

