PRO& CONTRA 2026

Вопросы целеполагания в стратегиях развития искусственного интеллекта

Секция №3: ИИ в технологиях системного анализа и стратегического планирования социально-экономического развития

Воронцов Константин Вячеславович
д.ф.-м.н., профессор РАН, зав. лабораторией МОСА Института ИИ МГУ

Тезисы

1. Достижения ГИИ: от проблесков AGI (2023) до автоматизации решения трудных задач в экономике и науке (2025) — высших сферах деятельности человеческого интеллекта.
2. Позитивные и негативные сценарии развития ИИ для человеческой цивилизации: всё возможно, но за что хвататься — создавать возможности или устранять угрозы?
3. Мы не определились в главном — для чего и как мы развиваем ИИ, и кто это «мы»? Есть полярные мнения о значимости, целеполагании, приоритетах, управляемости.
4. Отвечая на эти вызовы в энциклике **Magnifica Humanitas**, папа римский Лев XIV задаёт высокие этические ориентиры для профессиональных сообществ в сфере ИИ.
5. Растерянность в целеполагании видна даже в определениях ИИ, которые мы даём. Россия — единственная страна, записавшая превосходство ИИ над человеком в само определение ИИ, даваемое в Национальной стратегии развития ИИ до 2030 года.

Развитие AI в соревновании с человечеством?

прогнозируемые этапы развития

1956: дартмутский семинар
AI -- новая наука и мечта учёных

2020: эмерджентный AI (начиная с LLM GPT-3)

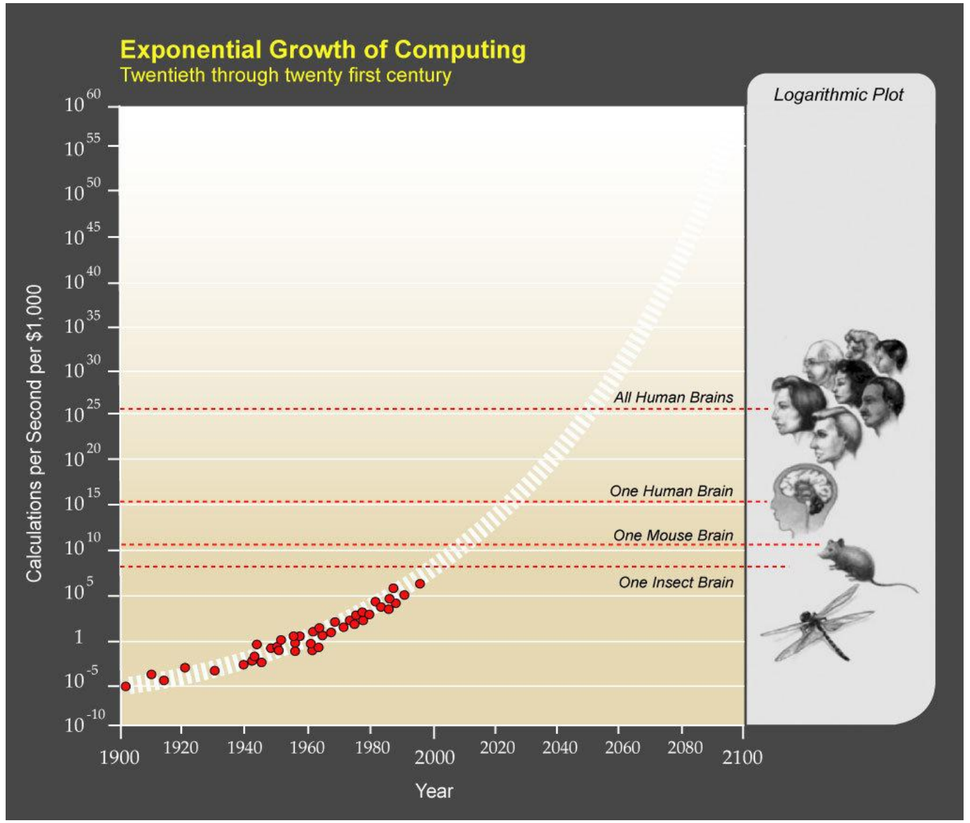
2030-е: AGI -- Artificial General Intelligence

превосходит интеллект человека

2040-е: ASI -- Artificial Super-Intelligence (искусственный супер-интеллект)

превосходит интеллект больших человеческих коллективов

универсальный ИИ (Universal AI)



Ray Kurzweil. The Singularity Is Near. 2005

Рэй Курцвейл. Революция разума: на подступах к Сингулярности. Как технологии изменят общество и сознание. 2025

Tim Genewein et al. From AGI to ASI. — DeepMind, ArXiv 2606.12683. — 10 Jun 2026.
http://www.machinelearning.ru: Публикация:DeepMind_2026_From_AGI_to_ASI

Проблески общего искусственного интеллекта

Sparks of Artificial General Intelligence:
Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research *(27 March 2023)*

Новые способности модели GPT, не закладывавшиеся при её обучении:

- объяснять свои ответы, перефразировать, переводить на другие языки
- реферировать, генерировать планы, сценарии, шаблоны
- строить аналогии, менять тональность, стиль, глубину изложения
- генерировать программный код на различных языках
- решать некоторые логические и математические задачи
- искать и исправлять собственные ошибки по подсказке

ИИ: переход к пятой научной парадигме

драйвер развития

научная парадигма:
общепринятия модель научной деятельности (Томас Кун, 1958)

4-й научной парадигмы

1 **эмпирическая:**
наблюдения, эксперименты, накопление данных

2 **теоретическая:**
создание моделей, законов, теорий

3 **вычислительная:**
компьютерное моделирование

4 **информационная:**
извлечение знаний из больших данных

- открытые данные
- бенчмарки, SOTA
- цифровые библиотеки
- поисковые системы

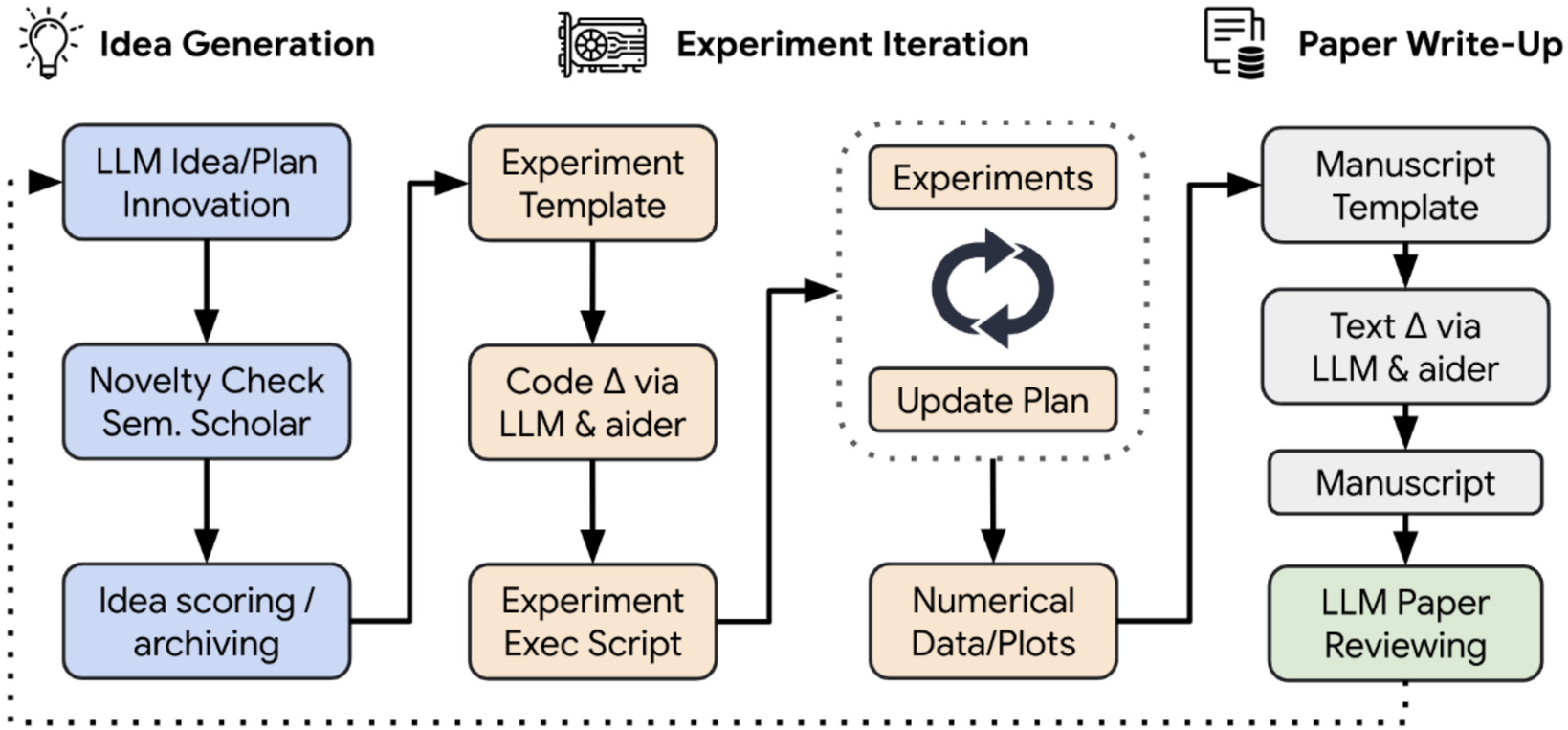
Джим Грэй (Microsoft), 2007

5-й научной парадигмы

5 **машинная:**
автоматизация полного цикла

- архив науки становится данными для непрерывной обработки
- ИИ становится генератором гипотез и рутинным исполнителем
- человек становится архитектором, отвечая за целеполагание
- знания становятся структурированными
  - по контексту
  - динамически
  - иерархически

Полная автоматизация научных исследований



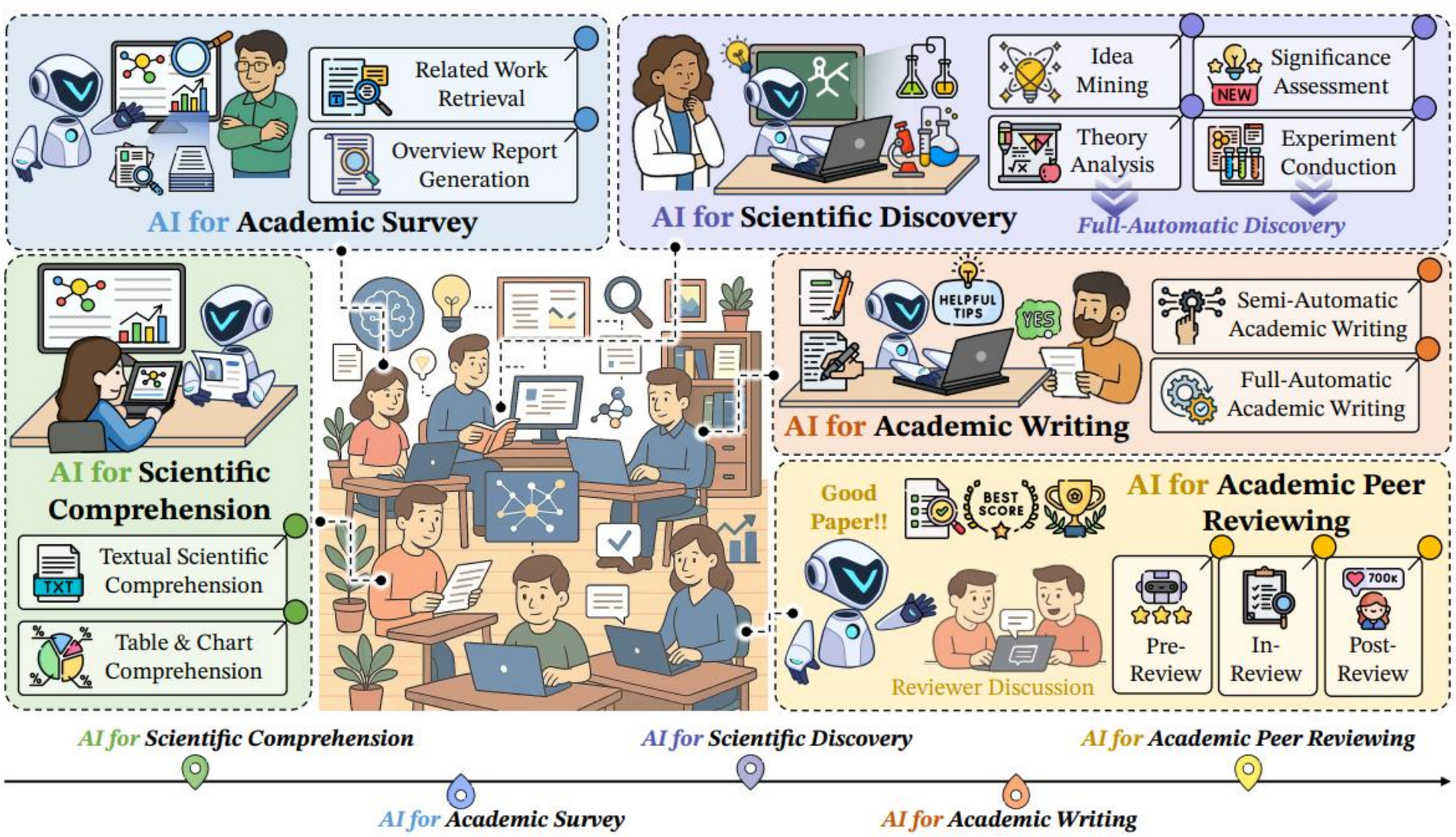
Источник:

*The **AI Scientist**: Towards Fully Automated Open-Ended Scientific Discovery, 2024*

AI Scientist — процесс научных открытий, основанный на LLM.

1. Генерация идей и оценивание их новизны.
2. Определение методов проверки гипотез, включая автоматическую генерацию кода.
3. Автоматическое выполнение экспериментов для сбора результатов, включая данные, графики, таблицы.
4. Результаты обосновываются, объясняются и суммируются в отчете LaTeX.
5. Генерация обзора в соответствии с современной практикой на конференциях по ML.

Автоматизация анализа научной литературы



AI4Research can be divided into five key areas:

(1) AI for Scientific Comprehension

(2) AI for Academic Survey

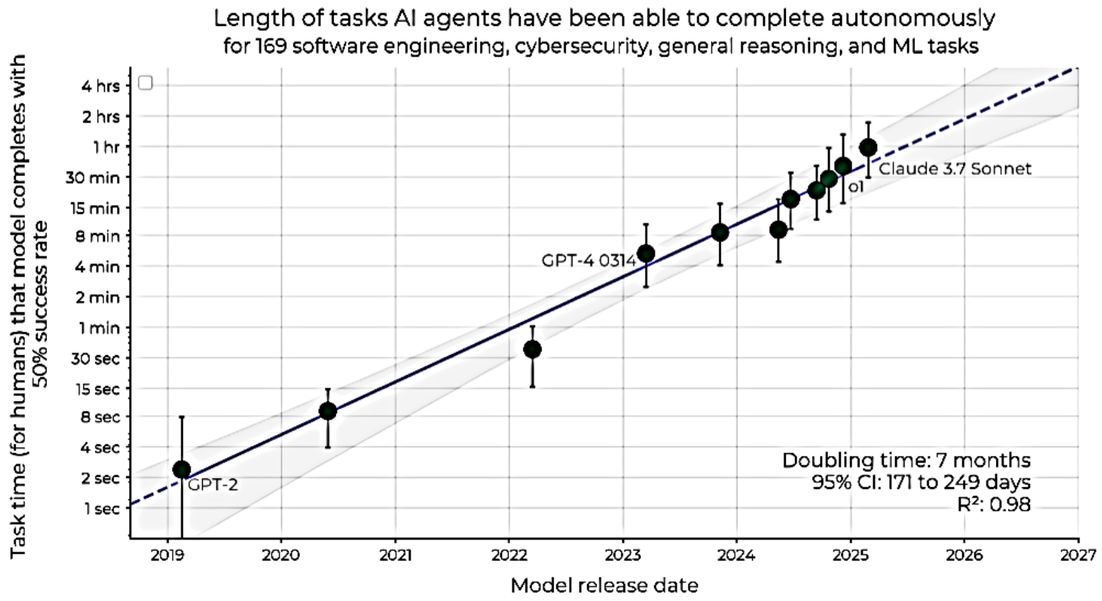
(3) AI for Scientific Discovery

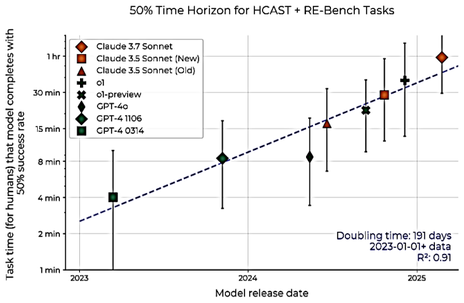
(4) AI for Academic Writing

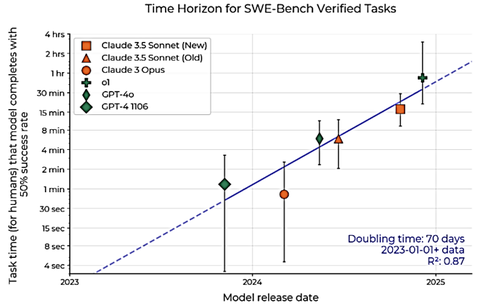
(5) AI for Academic Peer Review

AI4Research: A Survey of Artificial Intelligence for Scientific Research (LARG, июль 2025)

Автоматизация решения сложных задач







ИИ как рабочая сила в экономике

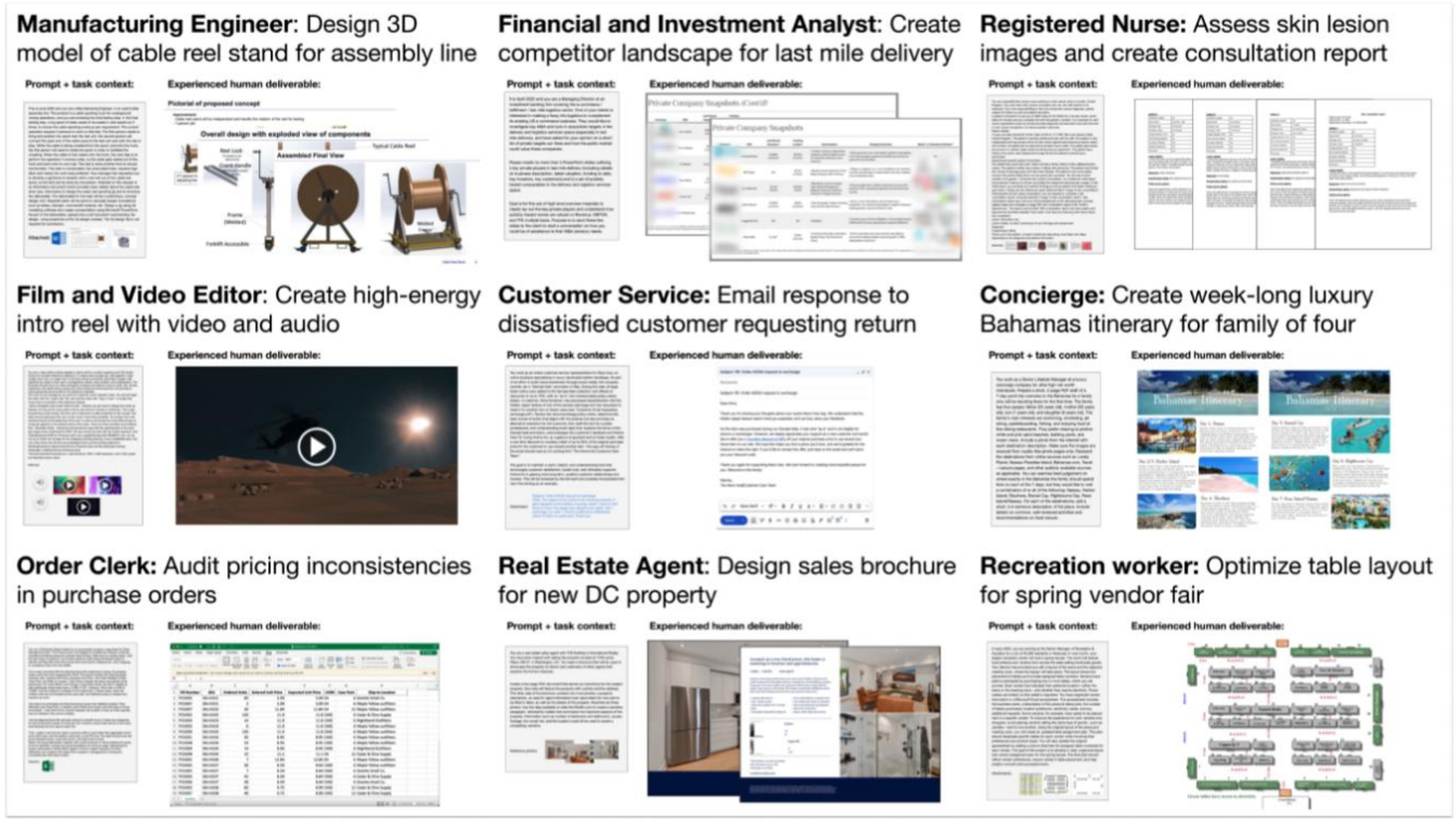
- **Сложность задач** (измеряемая временем, которое требуется людям-профессионалам), которые универсальные автономные агенты могут выполнить с 50% надежностью, **удваивается каждые 7 месяцев в течение последних 6 лет**

- В программировании ещё быстрее: **каждые 70 дней**!

Measuring AI Ability to Complete Long Tasks (**Model Evaluation & Threat Research , METR**). March 2025

Автоматизация реальных задач экономики

GDPval: бенчмарк ИИ-агентов-профессионалов (сентябрь 2025)



44 профессиональные задачи из 9 областей экономики

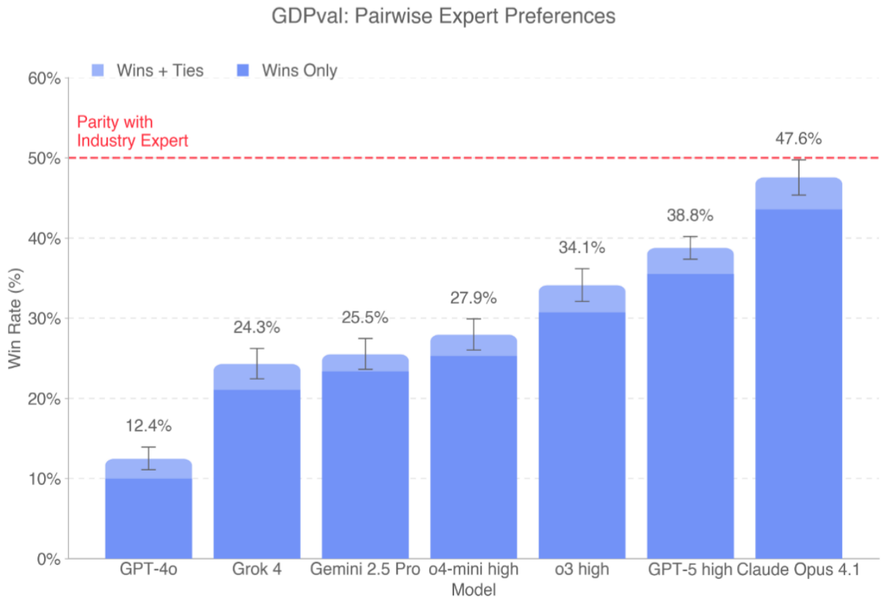
Пример запроса (300 слов)

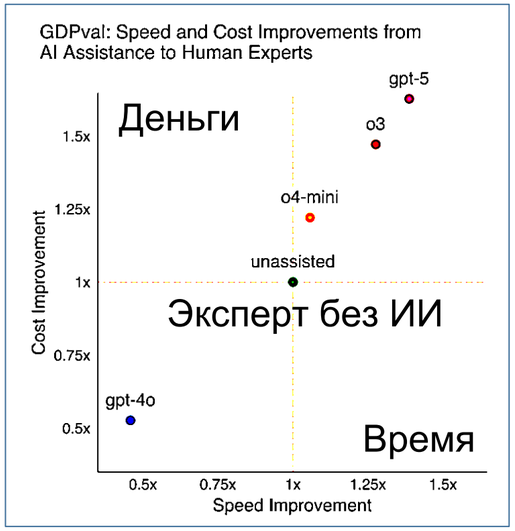
- **Вы — инженер-технолог в логистической компании**, занимающейся обработкой больших объёмов посылок… *(описание деталей задачи)*
- **Создайте качественную схему процесса** в формате PDF, которая наглядно продемонстрирует стандартизированную и оптимизированную версию того, как должен функционировать сквозной поток отправлений... (*описание требований к результату*)

GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks (Open AI, 5 Oct 2025)
https://huggingface.co/datasets/openai/gdpval

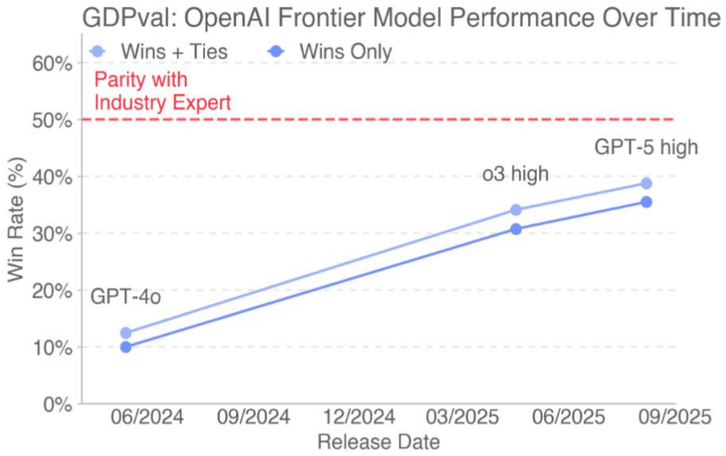
Автоматизация реальных задач экономики

GDPval: бенчмарк ИИ-агентов-профессионалов (сентябрь 2025)





За год модели стали в 4 раза ближе к уровню эксперта





(a) Reasoning effort experiment



(b) Prompt tuning experiment

GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks (Open AI, 5 Oct 2025)
https://huggingface.co/datasets/openai/gdpval

https://artificialanalysis.ai/evaluations/gdpval-aa

Будущее AI: позитивные сценарии



благоприятные сценарии, возможности развития

- автоматизация рутинного интеллектуального труда людей
- автоматизация управления в производстве, логистике, бизнесе
- высвобождение времени людей для творчества и познания
- создание единой системы знаний и коллективного интеллекта
- совершенствование цифровых помощников как двойников людей
- цифровое послесмертие и очеловеченный искусственный интеллект
- переход от человеческой цивилизации к человеко-машинной

источники постановок задач

в которой люди продолжают

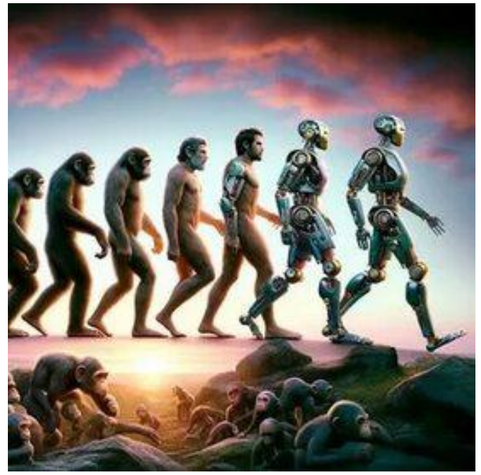
- определять смыслы, ставить цели
- развиваться, строить цивилизацию

Будущее AI: негативные сценарии

неблагоприятные сценарии, риски и угрозы

- массовые когнитивные искажения в восприятии чатов GPT
- захламление и деградация информационного пространства
- смещённость рекомендаций, информационные пузыри
- ментальные войны, усиление эффектов постправды
- усиление технологического и социального неравенства
- лишние люди, рост безработицы
- эскапизм в виртуальные среды
- деградация человеческого интеллекта
- автономные системы вооружений, риск потери контроля
- сингулярность по Курцвейлу, потеря управляемости
- передача управления и целеполагания машинам
- переход от человеческой цивилизации к машинной

источники постановок задач

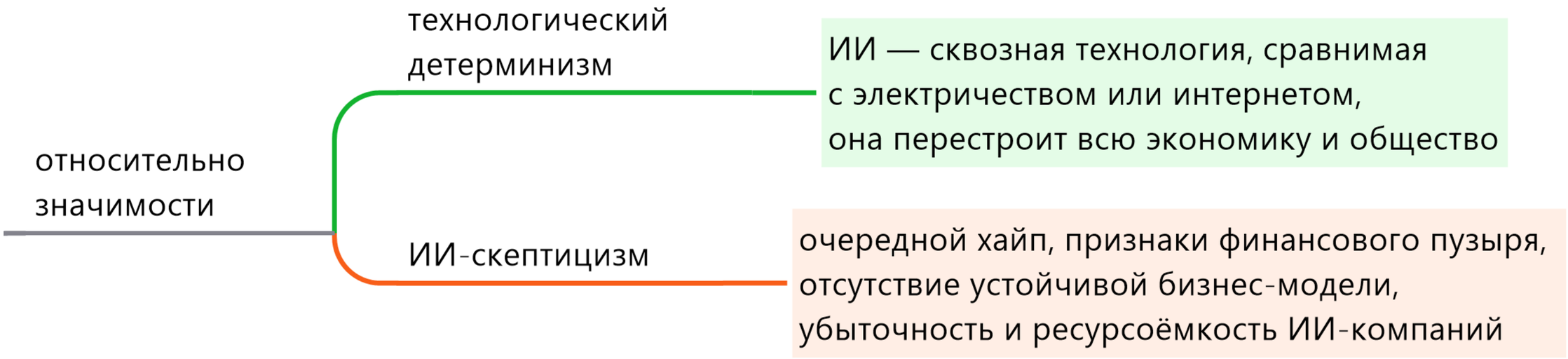






Поляризация мнений о сценариях развития ИИ

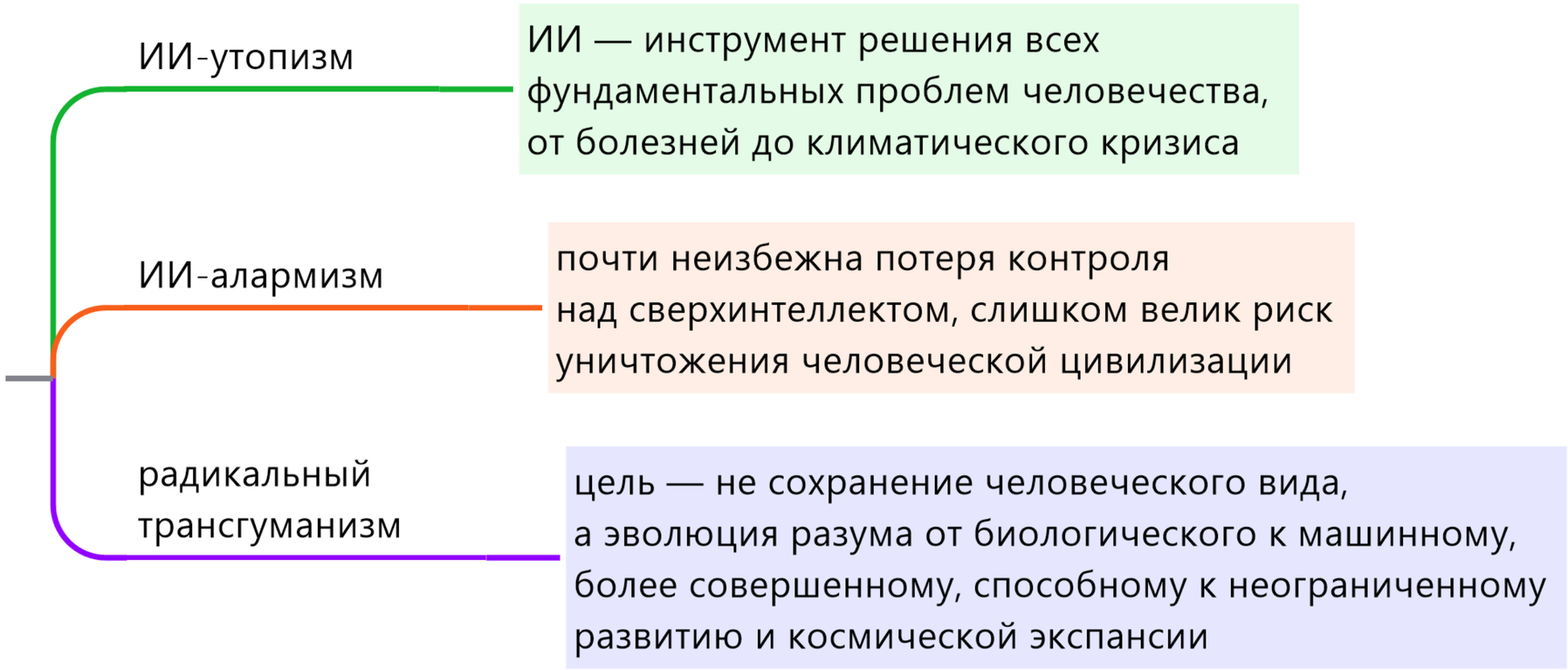
— относительно **значимости**, целеполагания, приоритетности, управляемости



Какой выбор положить в основу собственной деятельности?
Как перейти от попыток угадывания будущего к его конструированию?

Поляризация мнений о сценариях развития ИИ

— относительно значимости, **целеполагания**, приоритетности, управляемости

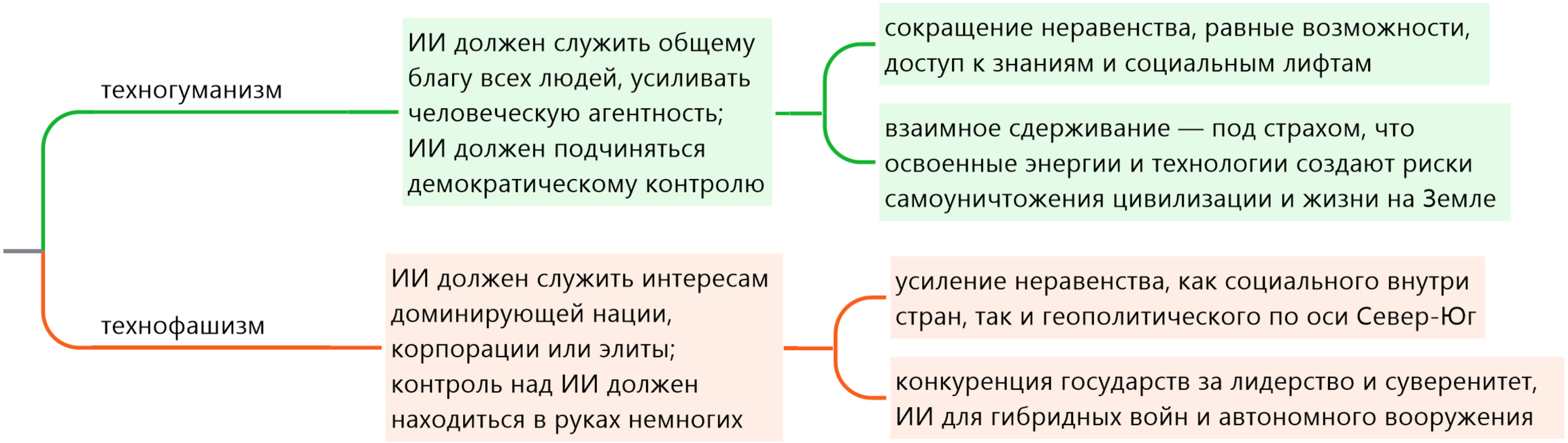


Какой выбор положить в основу собственной деятельности?

Как перейти от попыток угадывания будущего к его конструированию?

Поляризация мнений о сценариях развития ИИ

— относительно значимости, **целеполагания**, приоритетности, управляемости

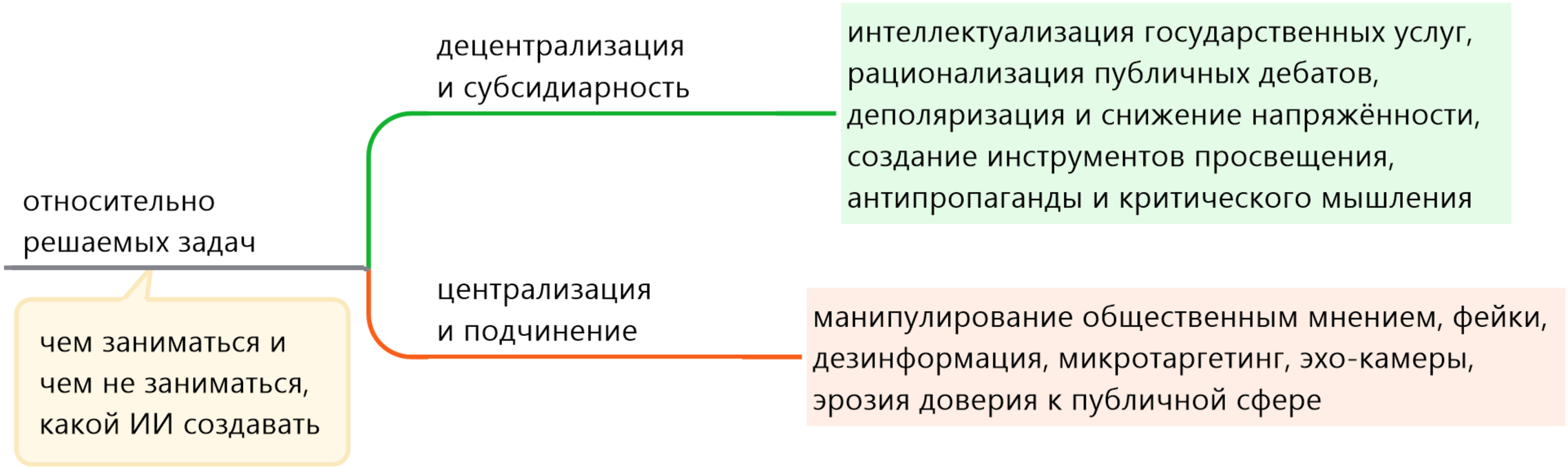


Какой выбор положить в основу собственной деятельности?

Как перейти от попыток угадывания будущего к его конструированию?

Поляризация мнений о сценариях развития ИИ

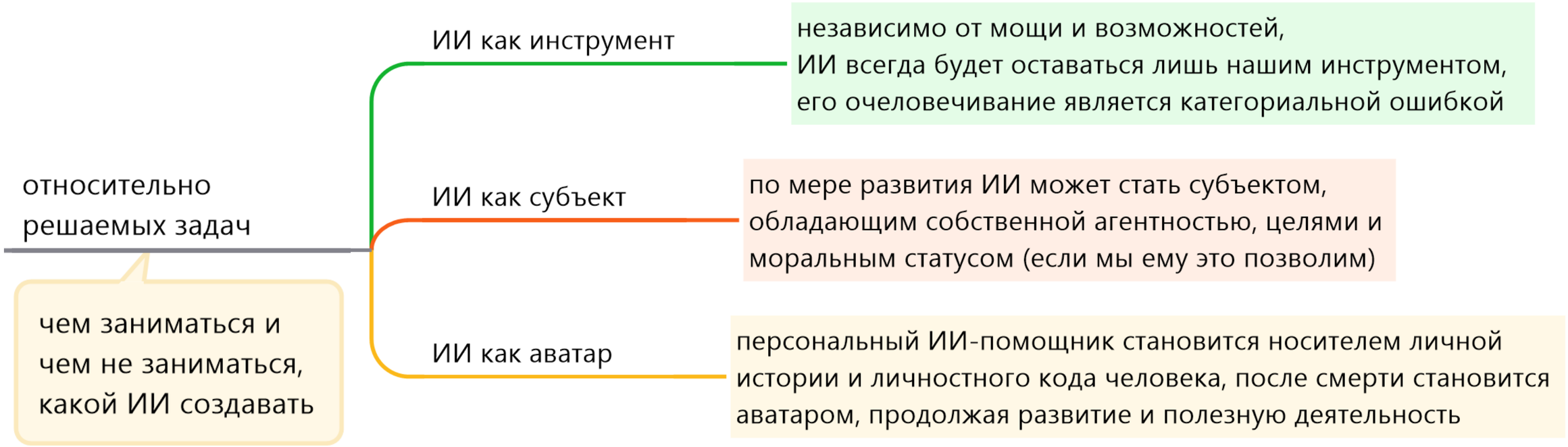
— относительно значимости, целеполагания, **приоритетности**, управляемости



Какой выбор положить в основу собственной деятельности?
Как перейти от попыток угадывания будущего к его конструированию?

Поляризация мнений о сценариях развития ИИ

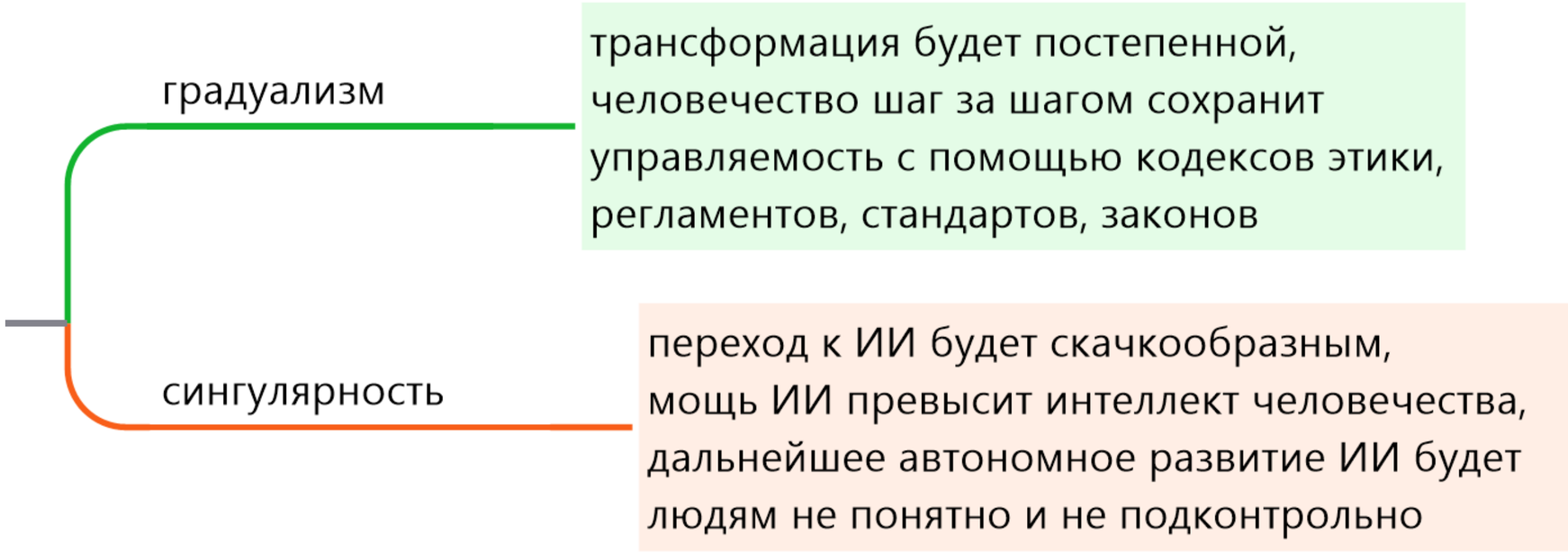
— относительно значимости, целеполагания, **приоритетности**, управляемости



Какой выбор положить в основу собственной деятельности?
Как перейти от попыток угадывания будущего к его конструированию?

Поляризация мнений о сценариях развития ИИ

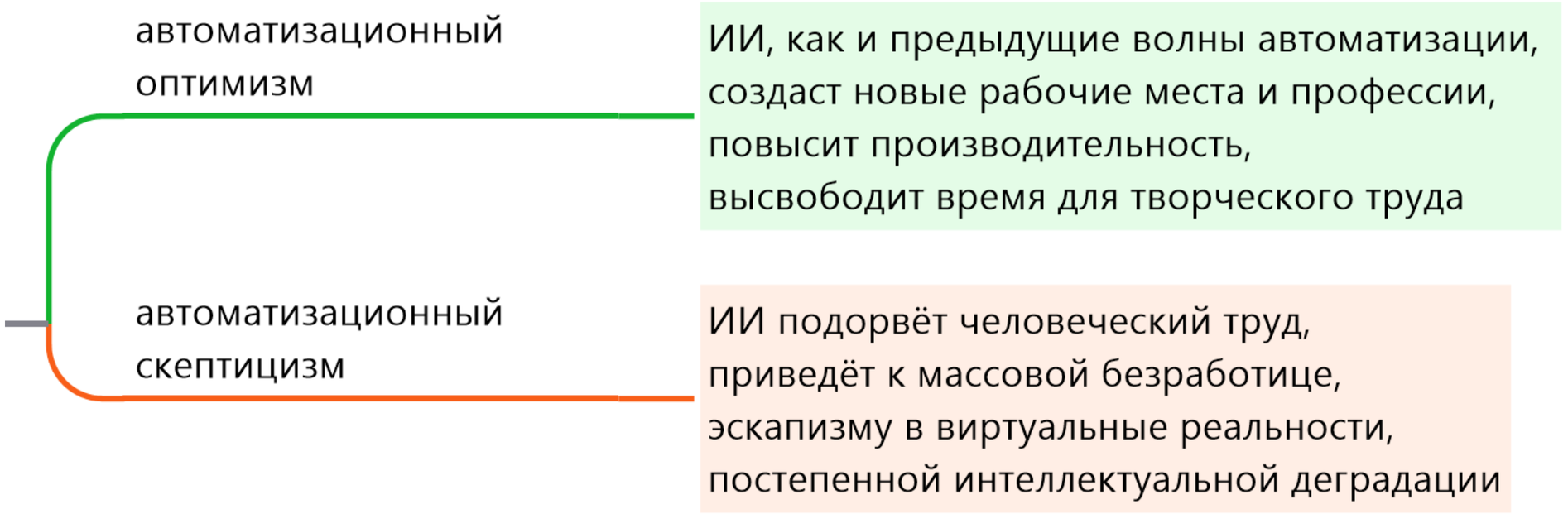
— относительно значимости, целеполагания, приоритетности, **управляемости**



Какой выбор положить в основу собственной деятельности?
Как перейти от попыток угадывания будущего к его конструированию?

Поляризация мнений о сценариях развития ИИ

— относительно значимости, целеполагания, приоритетности, **управляемости**



Какой выбор положить в основу собственной деятельности?
Как перейти от попыток угадывания будущего к его конструированию?

МН: какие технологии развивать и ради чего?

«Сама по себе технология не является
ни решением человеческих проблем, ни злом;
но в конкретной истории она не бывает нейтральной,
потому что принимает облик тех, кто ее придумывает,
оплачивает, направляет и использует» [9]

«Никогда еще человечество не обладало
такой властью над самим собой» [4]

«Ради какого замысла и в каком духе мы должны трудиться, чтобы сохранить и умножить величие человечества, полученное нами в дар?» [90]



Magnifica Humanitas
(Великолепное человечество)
— энциклика папы римского Льва XIV, 25мая 2026 г.
о защите человеческой личности в эпоху искусственного интеллекта

«Технический прогресс, сам по себе ценный, требует осознания стоящей за ним **антропологической модели и целей**, которые он преследует» [94]

МН: обращение к профсообществам в сфере ИИ

мотивации

- двойственная природа технологий
- ответственность
  — исследователей
  — разработчиков
  — бенефициаров
  — пользователей

социальная доктрина

- достоинство личности
- общее благо и его всеобщее предназначение
- справедливость
- субсидиарность
- солидарность

риски ИИ

- эскалация войны
- автономное вооружение
- неоколониализм
- изменение структуры труда
- несвобода личности
- деградация
- пост-правда

безопасный ИИ

- нормы этики ИИ
- «Цивилизация любви» как проект
- определение критериев и приоритетов
- воспитание и образование
- регламенты, законы, право

Достоинство человека, труд, общее благо

«**Уровень цивилизации** измеряется не мощью ее средств, а тем, как она заботится о человеке и видит в нем личность, а не функцию» [114]

«**Человек** является фундаментальной ценностью, которая важнее капитала и прибыли. **Человеческий труд** имеет приоритет перед любой производственной или финансовой логикой, и должен достойно вознаграждаться» [30]

«Богатство наций всё больше зависит от знаний и технологий, и если эти блага остаются сосредоточенными в руках немногих, без надлежащих форм совместного использования и доступа, возникает новое неравенство, противоречащее **всеобщему предназначению благ**» [67].

«**Строить цивилизацию любви** — не наивная утопия, а насущно необходимый проект» [186]

более подробный разбор энциклики МН

Антропоцентричное определение ИИ

Искусственный интеллект — комплекс технологических решений, позволяющий **имитировать** когнитивные функции человека (включая поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, **сопоставимые** с результатами интеллектуальной деятельности человека или **превосходящие** их.

— *Национальная стратегия развития искусственного интеллекта на период до 2030 года.*

Искусственный интеллект —
вычислительные технологии, повышающие производительность созидательного интеллектуального труда людей

не соперник и **не** замена человека,
не «загадочный новый тип разума»,
не повод уподобиться Богу, чтобы «творить по образу и подобию Своему»



О СПИКЕРЕ

PRO& CONTRA 2026

д.ф.-м.н., профессор РАН,

зав. лабораторией МОСА Института ИИ МГУ,

зав. кафедрой ММП ВМК МГУ,

зав. кафедрой МОЦГ МФТИ и кафедрой ИС МФТИ,

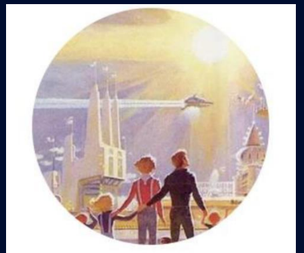
г.н.с. ФИЦ ИУ РАН

vorontsovkv@my.msu.ru

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



Воронцов Константин Вячеславович



Цивилизационная идеология

ДЗЕН-канал

https://dzen.ru/civideology