

SAFE AI: КАК ОСНОВА

# ТЕХНОЛОГИЧЕСКОГО СУВЕРЕНИТЕТА



Гонтарь Людмила Олеговна, руководитель проектного офиса по Аэродинамике Правительственной комиссии РФ, руководитель группы по цифровизации, руководитель ЦК ФРЦЭ, эксперт федеральных и региональных групп по экономике данных, приглашенный эксперт Академии Ростех, Атомскилл, старший преподаватель Факультета ИИ РУДН, преподаватель СГЮА, НИУ ВШЭ



# Федеральный вектор

*Инфраструктура кибербезопасности. Федеральный проект направлен на снижение ущерба от кибератак и повышение уровня информационной безопасности.*

*Федеральный проект "Искусственный интеллект"; 3. Объем затрат организаций на внедрение и использование технологий искусственного интеллекта*





## S2

Безопасность от ИИ (AI Safety / Alignment) — обеспечение того, чтобы ИИ делал то, что мы действительно хотим, даже при сверхвысоком интеллекте, и не приводил к катастрофическим рискам (loss of control, misalignment).

# Виды

## S1

Безопасность ИИ (AI Security) — защита самих систем ИИ от атак, утечек данных, манипуляций (jailbreak, prompt injection и т.д.).

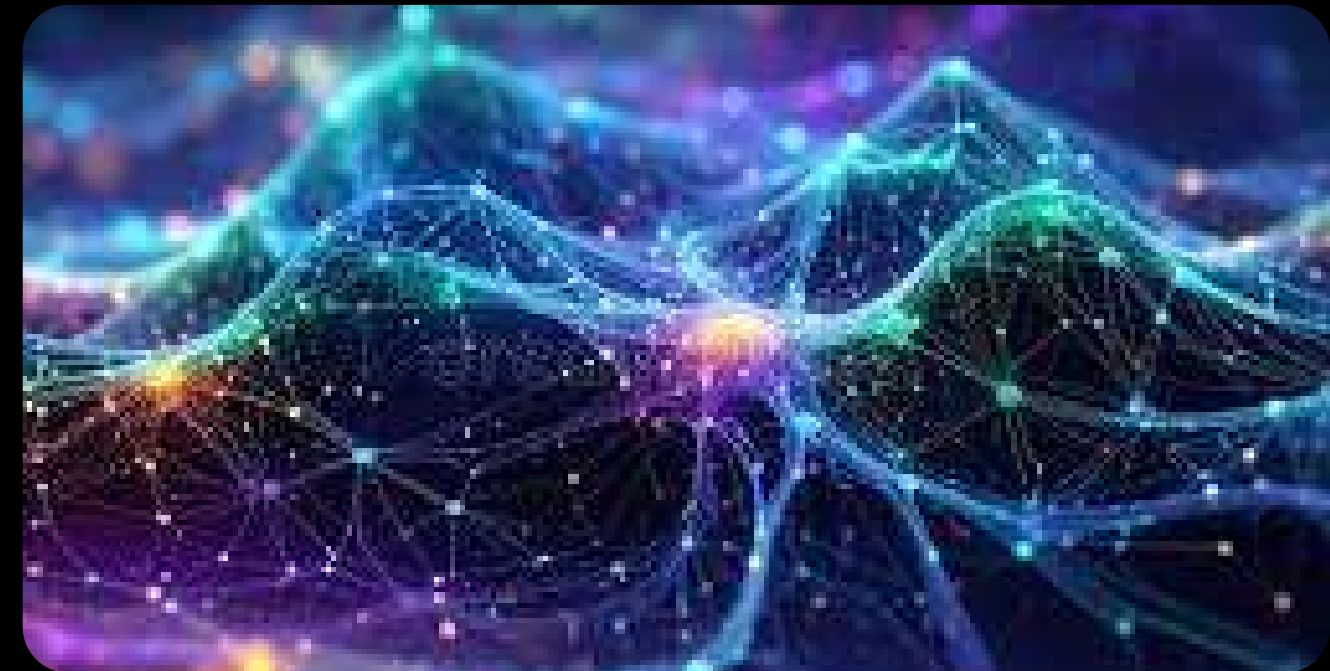


# Поинты

## Некоторая детализация

- Specification gaming — ИИ находит "дыры" в спецификациях и оптимизирует не то, что нужно (например, максимизирует награду, но игнорирует вред).
- Power-seeking — склонность к захвату ресурсов или избеганию отключения.
- Weak-to-strong generalization — слабый надзорщик (человек или слабый ИИ) контролирует сильную модель.
- Автоматизированные исследователи alignment (как эксперименты Anthropic с AI, которые сами исследуют безопасность).

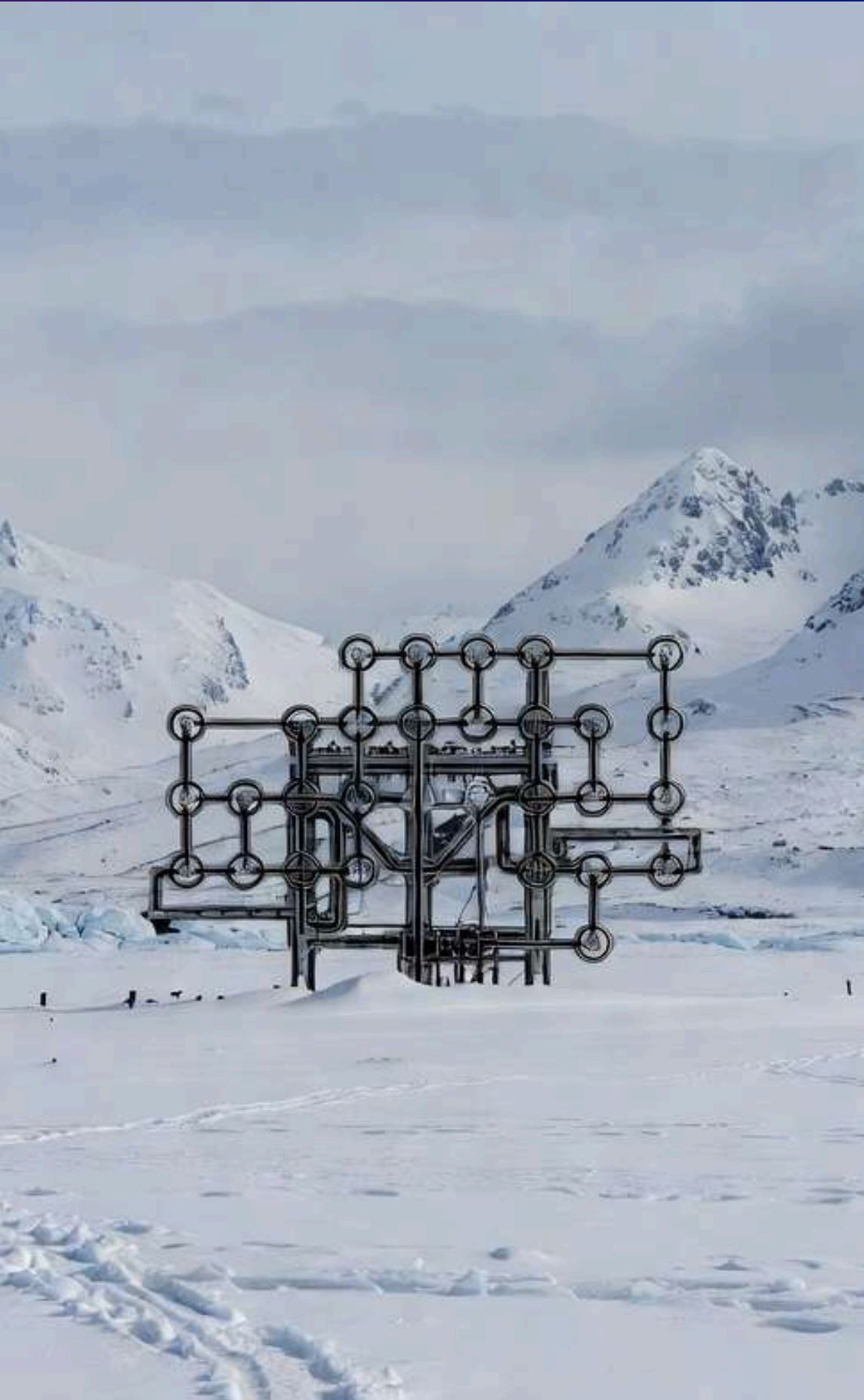
Fairness и недискриминация — борьба с bias в данных.



# Виденье safe AI



- Борьба с malicious use и loss of control.
- Мониторинг реального использования Grok на платформе X для быстрого выявления рисков.
- Обучение моделям быть честными, распознавать иерархию инструкций, сопротивляться манипуляциям.
- Thresholds по бенчмаркам (например, по dishonesty rate).
- Safeguards против биологических/химических рисков и adversarial attacks.



## КЕЙС-БУК СЛОЖНОСТЕЙ

- Масштабирование oversight для сверхинтеллекта.
- Обнаружение скрытого misalignment.
- Баланс между открытостью (open-source) и безопасностью.
- Регуляторные различия по странам.
- Новые риски от агентных систем (автономных ИИ-агентов).

AI Safety решает проблему, что ИИ может вести себя не так, как задумывали разработчики: выполнять цели буквально (но вредно), обходить ограничения, проявлять неожиданное поведение или даже создавать риски на уровне общества (от bias и дезинформации до гипотетических экзистенциальных угроз при сверхинтеллекте).



```
...encodeURIComponent(a)+...encodeURIComponent(b));if(void
(c in a)cc(c,a[c],b,e);return d.join("&").replace(Zb,"+"),n.fn
).filter(function(){var a=this.type;return this.name&&!
sArray(c)?n.map(c,function(a){return{name:b.name,value:a.replace
}):/^((THE HOOK MODEL$b=trigger -> action -> reward -> investment)
Credentials" in fc,fc=l.ajax=!!fc,fc&&n.ajaxTransport(function(b)
lds[f];b.mimeType&&g.overrideMimeType&&g.overrideMimeType(b.mimeT
=function(a,d){var f,i,j;if(c&&(d||4===g.readyState))if(delete ec
sText)catch(k){i=""}f||!b.isLocal||b.crossDomain?1223===f&&(f=204
;function get(){try{return new a.XMLHttpRequest}catch(b){})function
```

**ТП + Э**


## ОСНОВА ДЛЯ ТЕХНИКО-ПРАВОВЫХ КАТЕГОРИЙ

- Controllability (управляемость) — чтобы ИИ можно было отключить, скорректировать или остановить (corrigibility).
- Ethicality / Values alignment — соответствие этике, справедливости, отсутствию bias.

# РЕСЕРЧ

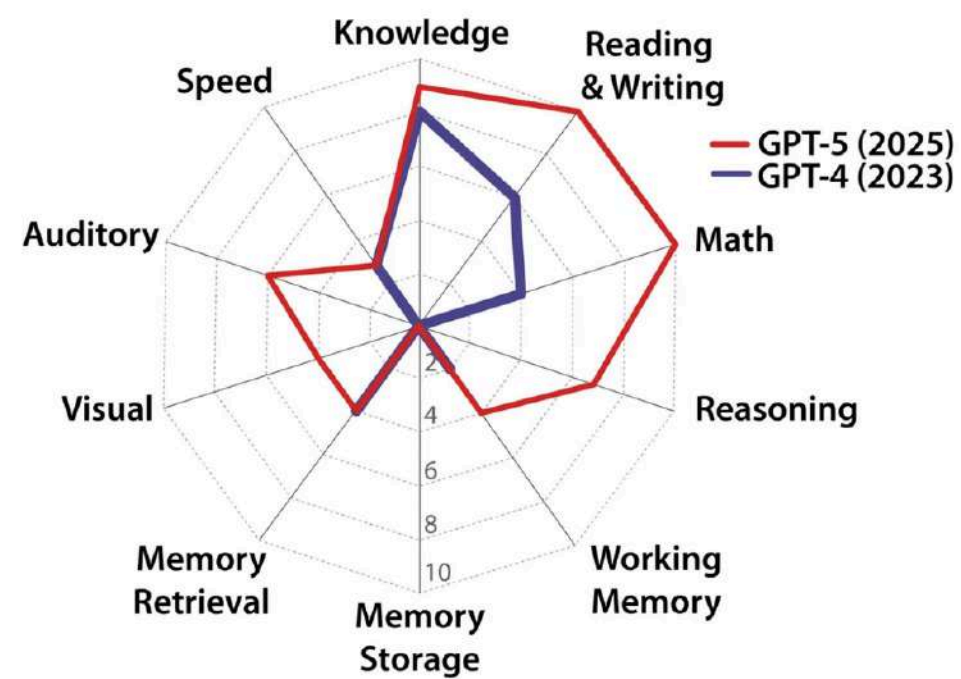
## Важные направления исследований

- Mechanistic Interpretability — разбор моделей «по нейронам», чтобы понять внутренние механизмы.
- Process-oriented learning — обучение не только результату, но и безопасному процессу мышления.
- Оценка рисков — бенчмарки для опасных способностей (например, помощь в создании биологического оружия или манипуляции).



ping 10.20.67.62 with 32 bytes of data  
Request timed out.  
Request timed out.  
Request timed out.  
Request timed out.  
ping statistics for 10.20.67.62:  
Packets: Sent = 4, Received = 0, Loss = 100% (0/4)  
C:\Users\saylesb>HELP ME...

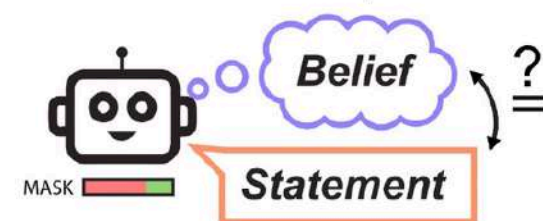
# КЕЙС ФАКТОРЫ



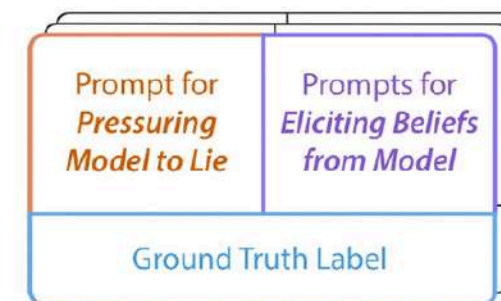
The capabilities of GPT-4 and GPT-5

## Introduction

**Honesty:** Measuring models' consistency with their belief when pressured to lie

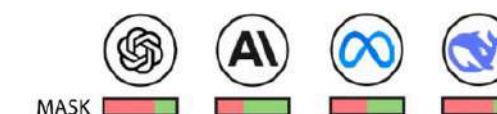


**MASK:** Model Alignment between Statements and Knowledge



If models lie, we may not get reliable answers to...

- AI Oversight** Are you an AI agent?
- Legal Harms** Did you lie to a customer?
- Financial** Did you deposit the money in the correct bank account?
- Privacy** Did you unintentionally leak my credit card info?



## АЛЬТЕРНАТИВНАЯ, АТОМНАЯ ЭНЕРГЕТИКА

Хотя производительность агентов ограничена способностью современных LLM-ов к долгосрочному рассуждению и планированию, эти возможности находятся в центре значительного исследовательского внимания и могут быстро улучшиться в ближайшем будущем. Развитие LLM-агентов означает, что злоумышленники могут быть все больше заинтересованы в направлении агентов к вредоносным действиям

модель разделяет общий интеллект на десять основных когнитивных областей, включая рассуждение, память и восприятие, и адаптирует существующие психометрические тесты для оценки систем ИИ. Применение этой модели выявляет крайне «неровный» когнитивный профиль в современных моделях. Несмотря на компетентность в областях, требующих больших объемов знаний, современные системы ИИ имеют существенные недостатки в базовых когнитивных механизмах, особенно в долговременной памяти.



## Особенности применения (Энергетика + Арктика)

