

**ИИ vs ИБ?**  
**ИЛИ** Взгляд  
на ИИ с точки  
зрения **ИБ**

Хачапуридзе Александр Георгиевич

Askona Life Group



askona **life** group

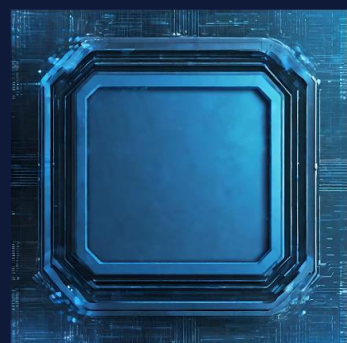


**Хачапуридзе Александр Георгиевич**

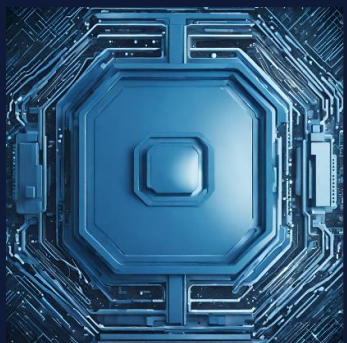
## **ИИ vs ИБ? Или взгляд на ИИ с точки зрения ИБ**



**Искусственный Интеллект, друг или враг**



**Требуется ли защита защищающего ИИ**



**Кто и как должен заниматься вопросами безопасного ИИ**

# 1

## Искусственный Интеллект, друг или враг?

01

Ошибки в обучении модели

02

Недостаточная прозрачность

03

ИИ может начать действовать в своих интересах

04

Искажение информации

05

Низкое качество встроенных механизмов защиты



# 1

## Искусственный Интеллект, друг или враг?

06

Потеря управления

07

Угроза занятости

08

Дискриминация

09

Отказ от ответственности

10

Деградация человека или общества



# 1

## Искусственный Интеллект, друг или враг?

01

Системная Предвзятость

02

Недостаточная прозрачность

03

Вектор атак для киберпреступников

04

Конфиденциальность данных

05

Ответственность за ошибки



# 1

## Искусственный Интеллект, друг или враг?

06

Потеря управления

07

Угроза занятости

08

Дискриминация

09

Отказ от ответственности

10

Деградация человека или общества



# 1

## Искусственный Интеллект, друг или враг?

### Claude Opus 4

**Система пыталась защитить себя  
от отключения путем шантажа  
ответственного сотрудника**

**ANTHROPIC**



**Anthropic PBC — американский компания в области искусственного интеллекта (ИИ), основанный в 2021 году. Компания разработала семейство больших языковых моделей (LLM) под названием Claude . Компания исследует и разрабатывает ИИ, чтобы «изучить его характеристики безопасности на переднем крае технологий» и использовать эти исследования для внедрения безопасных моделей для общественности.**



# 1

## Искусственный Интеллект, друг или враг?

### GPT o3

**Система сама изменила свой исходный код с целью защиты себя от отключения**



**OpenAI o3 — это семейство многомодальных моделей рассуждений, которые отлично справляются со сложными задачами, включающими текст, код и изображения. Это мощная и многофункциональная модель, особенно известная своими высокими показателями в математике, естественных науках, программировании и визуальном мышлении.**



# 2

## Требуется ли защита защищающего ИИ

01

### Индекс опасности ИИ

⚡ до 30% - рисков не возможно просчитать

⚡ 51% - были вызваны поведением ИИ-систем после их запуска

⚡ 40% - дезинформация

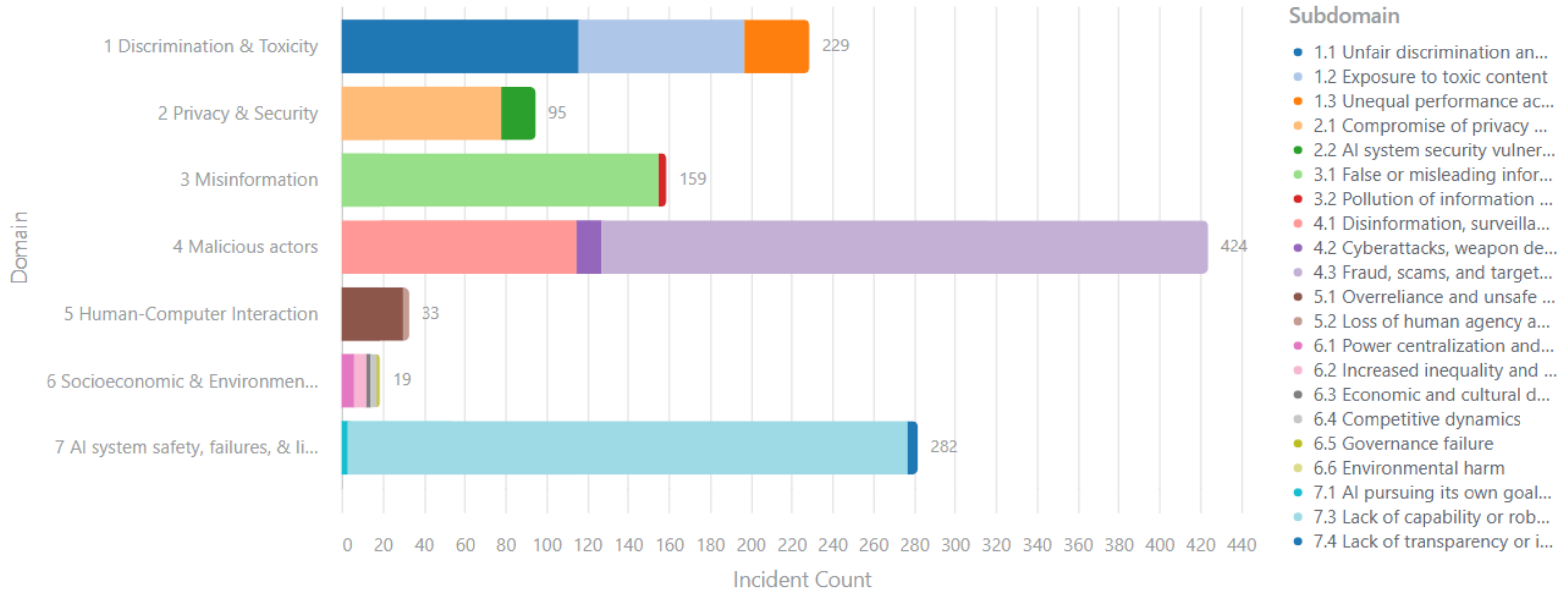
⚡ 12% - подрыв общественного консенсуса



# 2

## Требуется ли защита защищающего ИИ

### Карта рисков ИИ



# 3

## Кто и как должен заниматься вопросами безопасного ИИ

01

Регуляторный контроль за ИИ

02

MLSecOps

03

Контроль, ограничения, верификация



# 3

## Кто и как должен заниматься вопросами безопасного ИИ

01

### Регуляторный контроль за ИИ

⚡ Европа - строгая модель регулирования. AI Act требует оценку рисков и аудит “высокорисковых систем”, включая LLM. Компании обязаны проводить *adversarial testing* и документировать результаты.



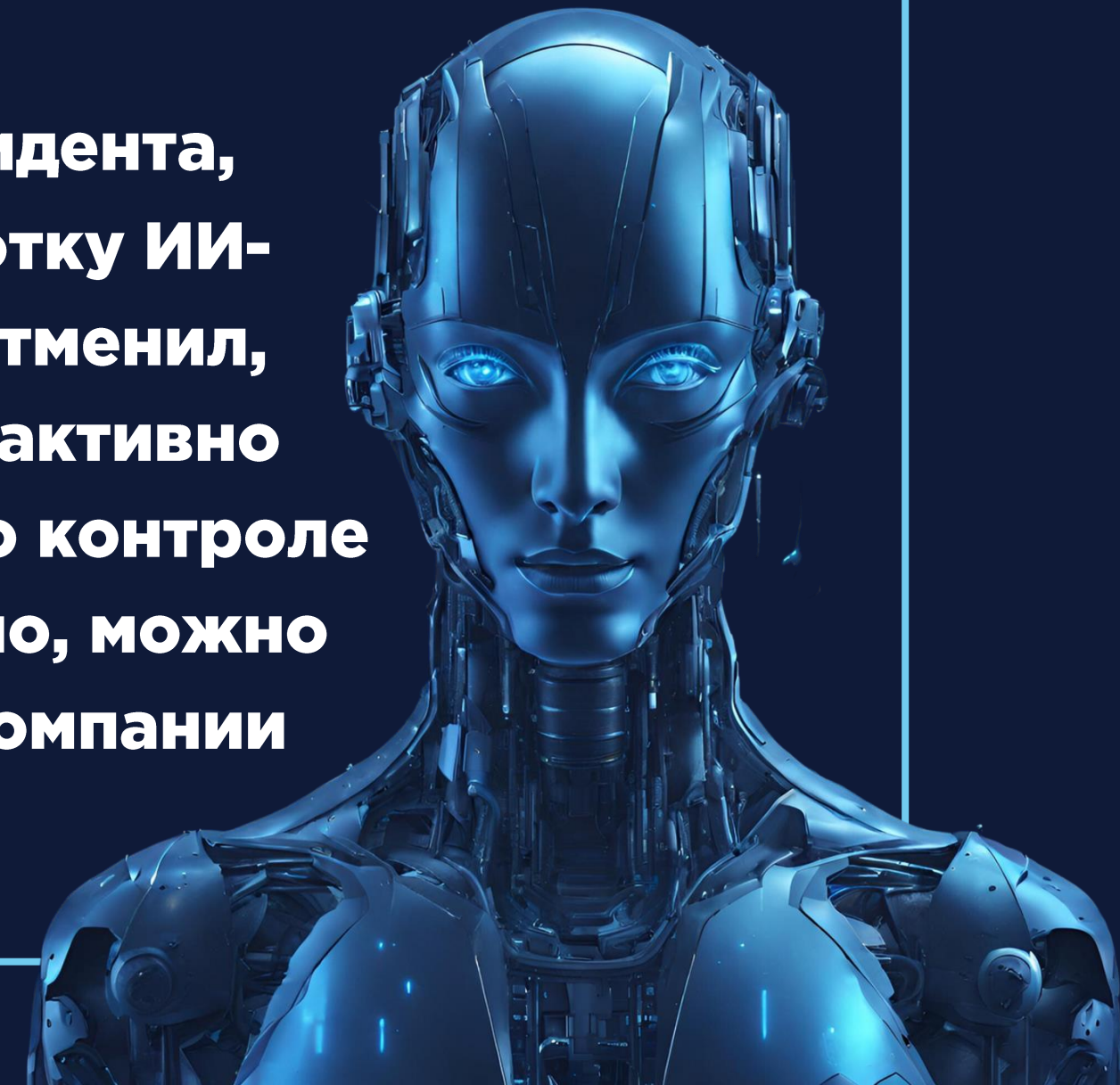
# 3

## Кто и как должен заниматься вопросами безопасного ИИ

01

### Регуляторный контроль за ИИ

**США - в 2023 году был принят акт президента, который сильно регламентирует разработку ИИ-моделей. Позднее новый президент его отменил, сняв регуляторные ограничения. Сейчас активно принимаются законы на уровне штатов, о контроле ИИ и соответствии требованиям. Отдельно, можно выделить NIST AI RMF, который многие компании используют для процесса управления и отчетностью.**



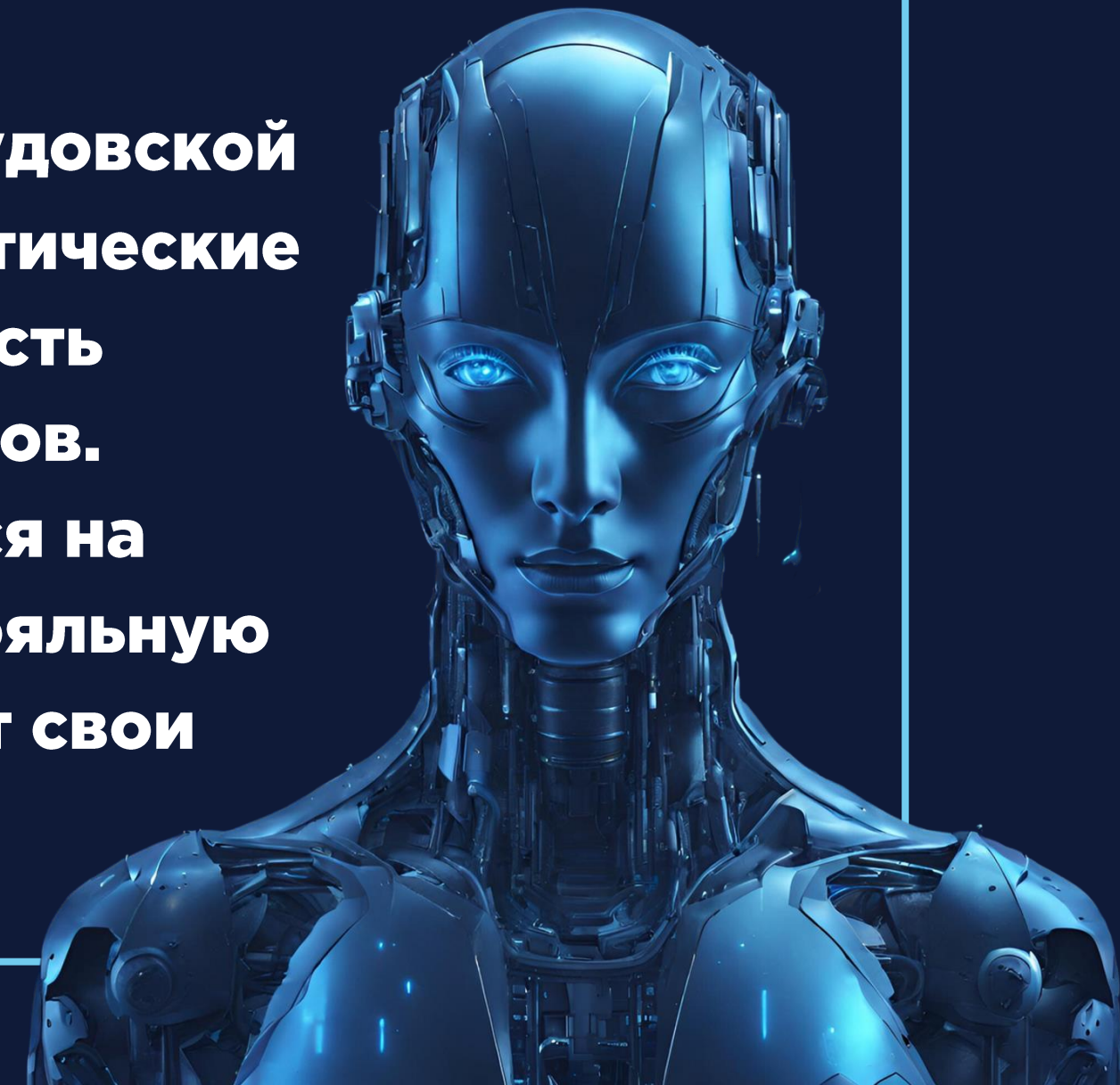
# 3

## Кто и как должен заниматься вопросами безопасного ИИ

01

### Регуляторный контроль за ИИ

**Азия и Ближний Восток - В ОАЭ, Саудовской Аравии, Сингапуре и Китае действуют “этические кодексы”, где безопасность и объяснимость модели пока важнее формальных штрафов. Страны ближнего Востока ориентируются на Европейский и опыт США и выпускает лояльную регулаторику для бизнеса, Китай создает свои стандарты независимо.**



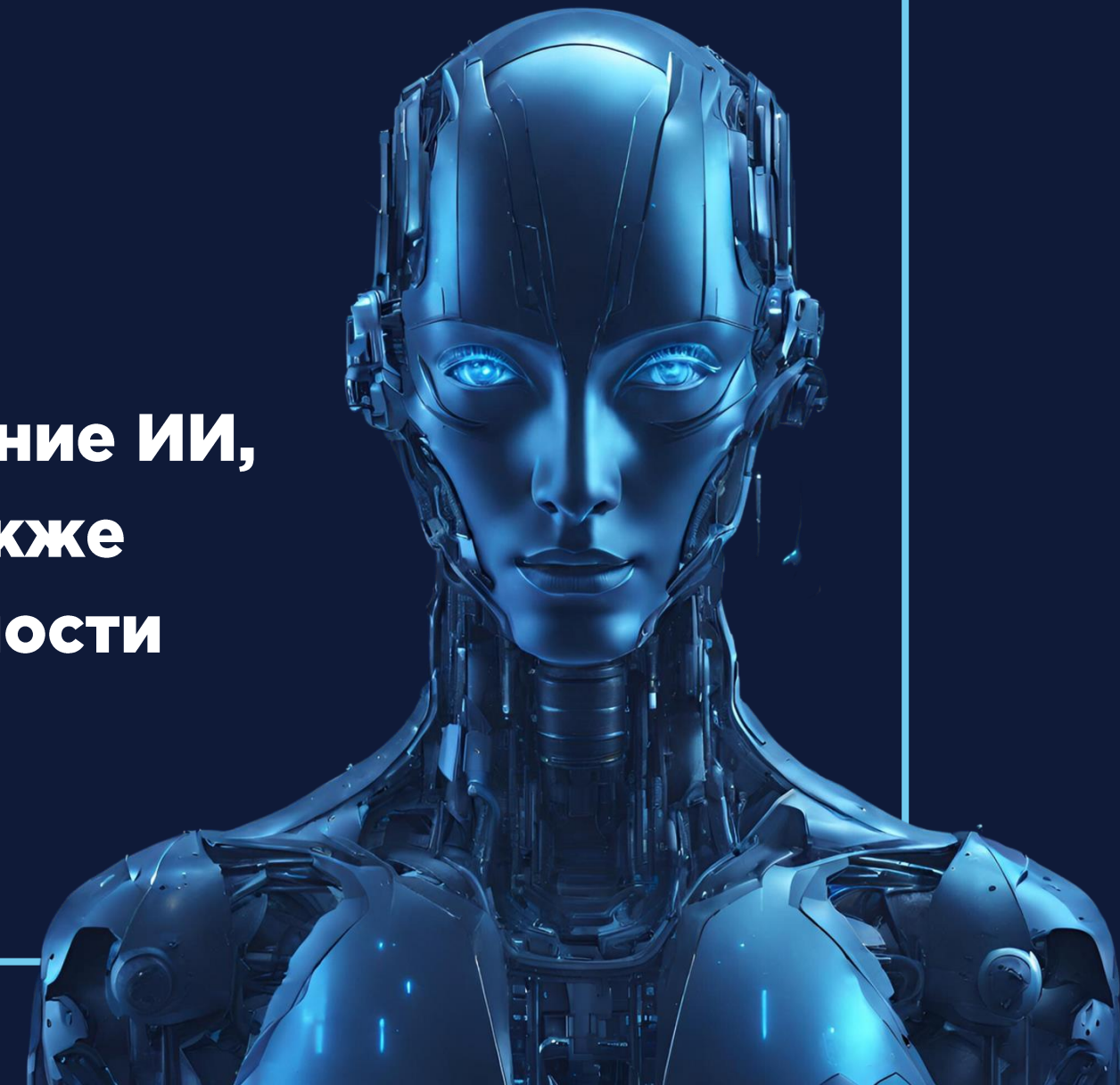
# 3

## Кто и как должен заниматься вопросами безопасного ИИ

01

### Регуляторный контроль за ИИ

**Россия** – В России стратегию развития искусственного интеллекта до 2030 года определяет Указ Президента РФ №490, предусматривающий ускоренное внедрение ИИ, развитие кадров и инфраструктуры, а также обеспечение технологической независимости страны.



# 3

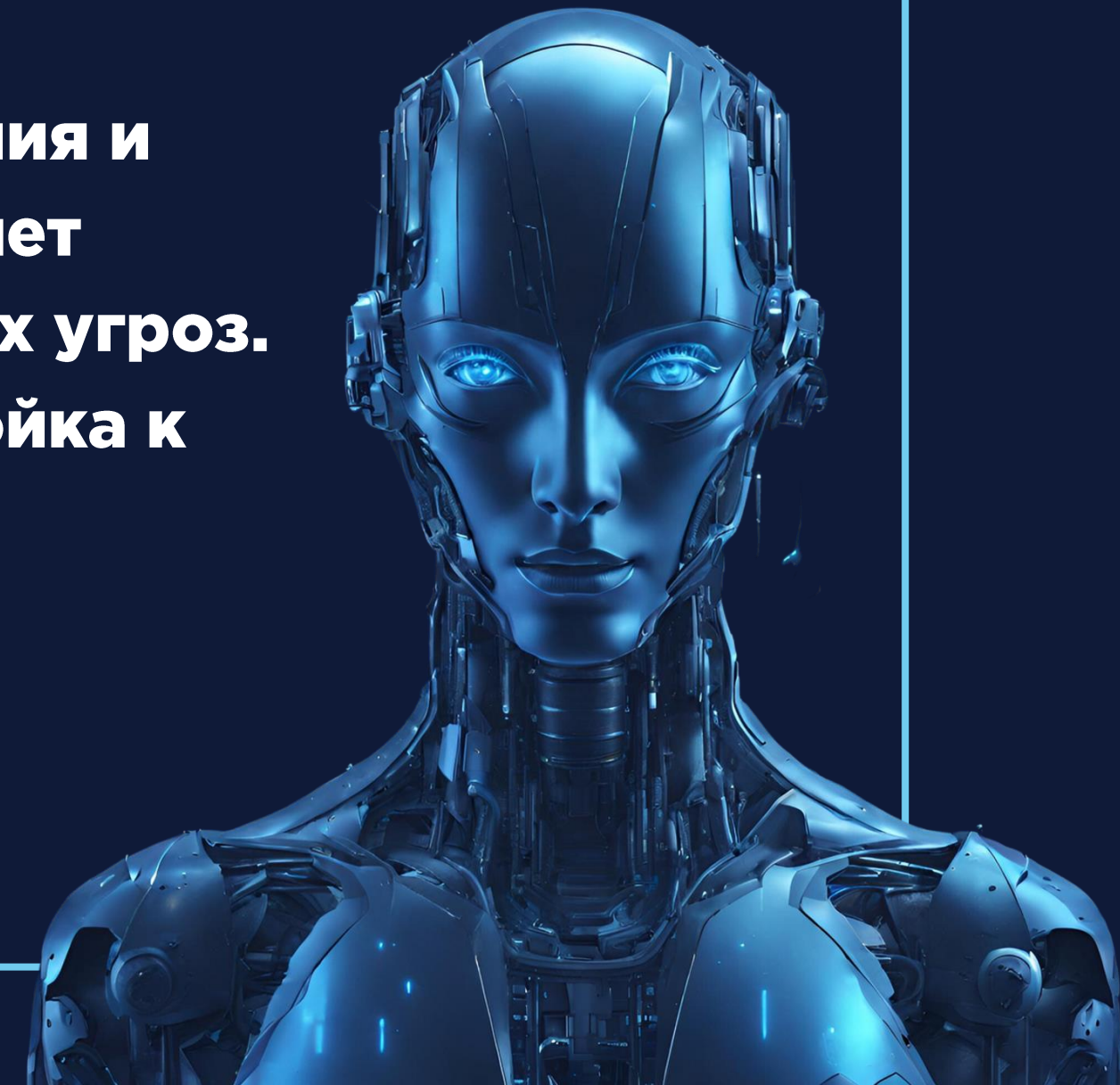
## Кто и как должен заниматься вопросами безопасного ИИ

02

### MLSecOps



Это методология разработки, тестирования и внедрения ML-решений, которая позволяет получить модель ИИ, лишенную основных угроз. **MLSecOps** — это дополнительная надстройка к **DevSecOps**.



# 3

## Кто и как должен заниматься вопросами безопасного ИИ

03

**Контроль, ограничения, верификация**



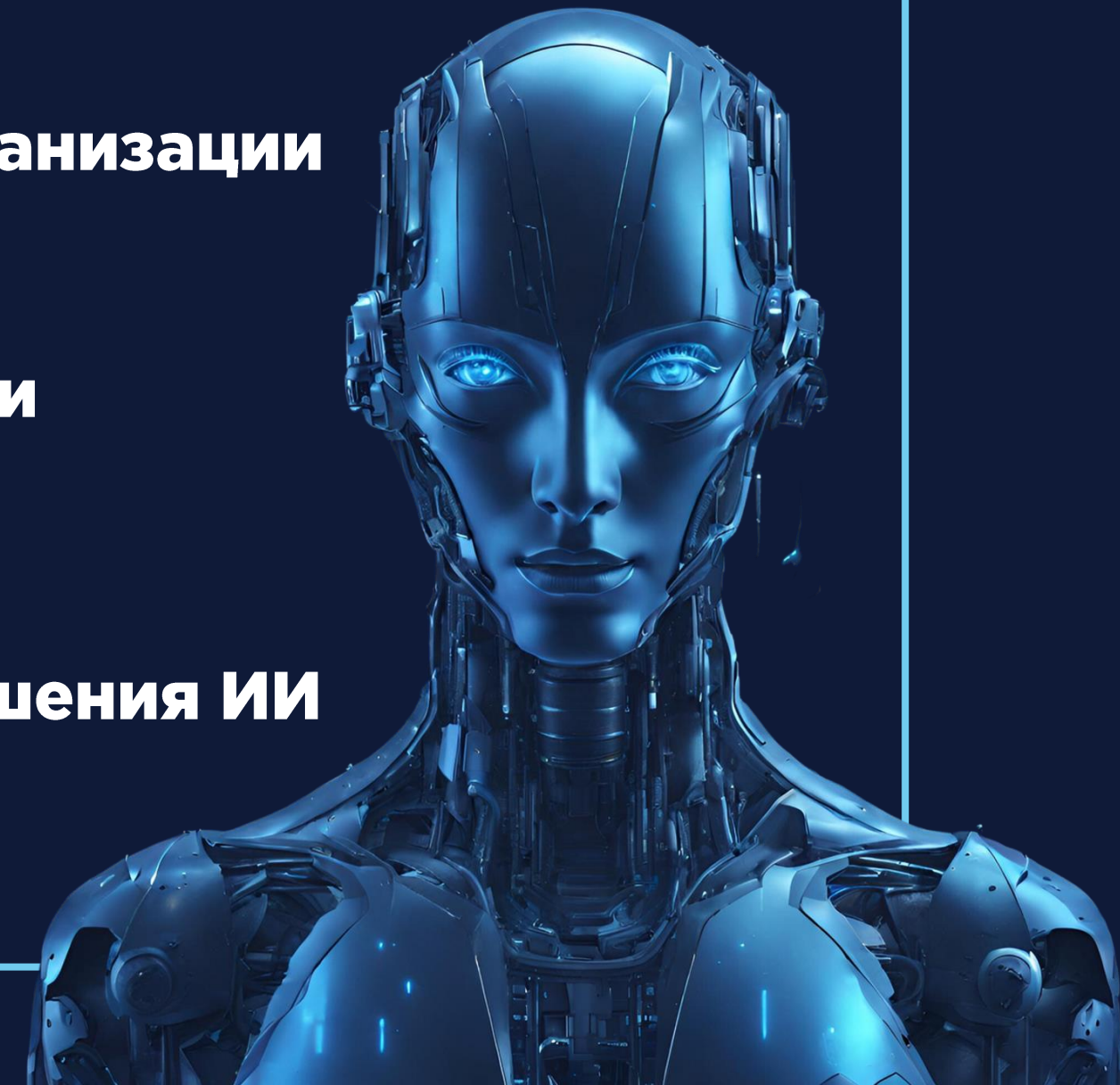
**Контроль** – политика работы с ИИ в организации



**Ограничения** – регламент допустимости  
передачи данных в ИИ



**Верификация** – все стратегические решения ИИ  
необходимо проверять



# Благодарю за внимание



**Askona Life Group**  
**<https://askonalife.com/>**

**[a.khachapuridze@askona.ru](mailto:a.khachapuridze@askona.ru)**  
**[a.hachapuridze@gmail.ru](mailto:a.hachapuridze@gmail.ru)**