БЕЗОПАСНОСТЬ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ЗАПРЕТИТЬ ИЛИ НАУЧИТЬСЯ РАБОТАТЬ?

А.М. ШОЙТОВ

МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ, СВЯЗИ
И МАССОВЫХ КОММУНИКАЦИЙ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Май 2025

2025 ГОД – БЕЗОПАСНОСТЬ ИИ

2024 год

Академией криптографии Российской Федерации в рамках реализации отдельного мероприятия федерального проекта «Искусственный интеллект» национальной программы «Цифровая экономика Российской Федерации» в 2024 году разработаны и направлены в заинтересованные федеральные органы исполнительной власти и организации «Предложения в проект требований по обеспечению информационной безопасности информационных систем, реализующих технологии искусственного интеллекта».

2025 год

На основании материалов Академией криптографии, Консорциум исследований безопасности технологий ИИ разрабатываем профильную матрицу угроз для объектов КИИ и ГИС и приступает к ее отработке с пилотными регионами России (респ. Татарстан, Новосибирская обл, Сахалинская обл.). В Консорциуме (Газинформсервис) разработана Матрица угроз безопасности при разработке ПО с применением ML-моделей с описанием 61 угрозы и 141 доступного решения.

КОНСОРЦИУМ ИССЛЕДОВАНИЙ БЕЗОПАСНОСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Сбербанк разработал и представил Модель угроз для кибербезопасности AI , для этапов сбора, подготовки данных, разработки модели обучения, эксплуатации модели и интеграции с приложениями. Документ описывает 70 угроз моделей генеративного (GenAI) и предиктивного (PredAI) искусственного интеллекта, актуальных для соответствующих сфер деятельности.

АТАКИ НА ИИ

HTTPS://OWASP.ORG/WWW-PROJECT-MACHINE-LEARNING-SECURITY-TOP-10/

С 2019 по 2024 год было опубликовано около 17000 научных статей,

которые описывают состязательные атаки на ИИ.

По запросу «adversarial examples AI» количество можно проверить

в scholar google ttps://scholar.google.com/

1. Инъекция (Prompt injection).

- Описание: Атака по смыслу схожа с sql-инъекцией. Но на текущий момент, это уже что-то вроде социальной инженерии для ИИ. В контексте атак на ИИ ее относят к jailbreak.

2.Инфекция (Infection).

- Заражение вредоносным ПО. В OWASP один из примеров – это атака на цепочку поставок.

3.Уклонение (Evasion).

- Данные на вход для ИИ модифицируют не значительно. Нарушитель стремиться заставить ИИ ошибиться

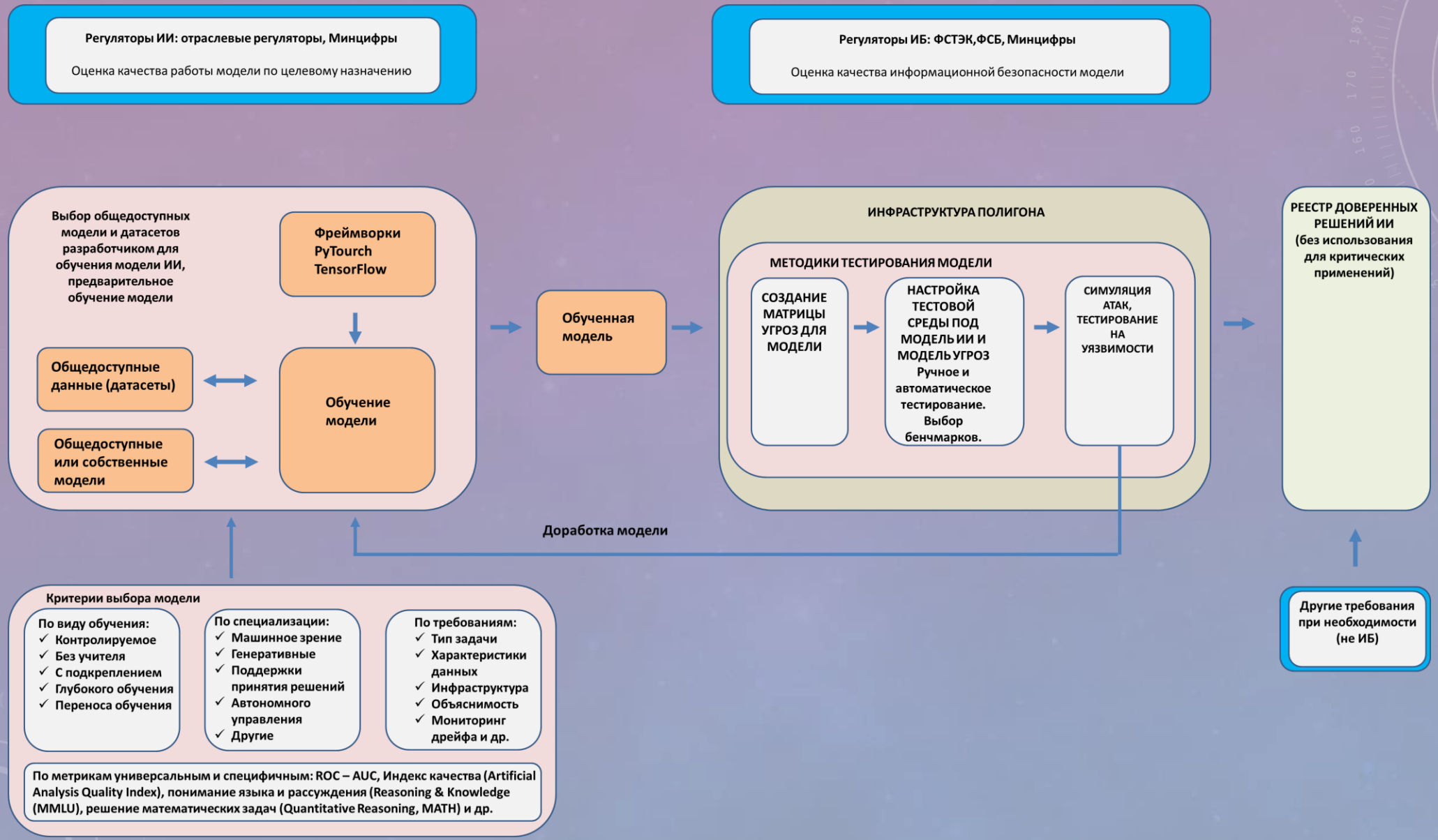
4. Отравление ИИ (Poisoning)

- Отравление данных, которые поступают в ИИ. Доказано, что достаточно внести 0,001 % ошибок в данные, чтобы результаты оказались аномальными и неверными

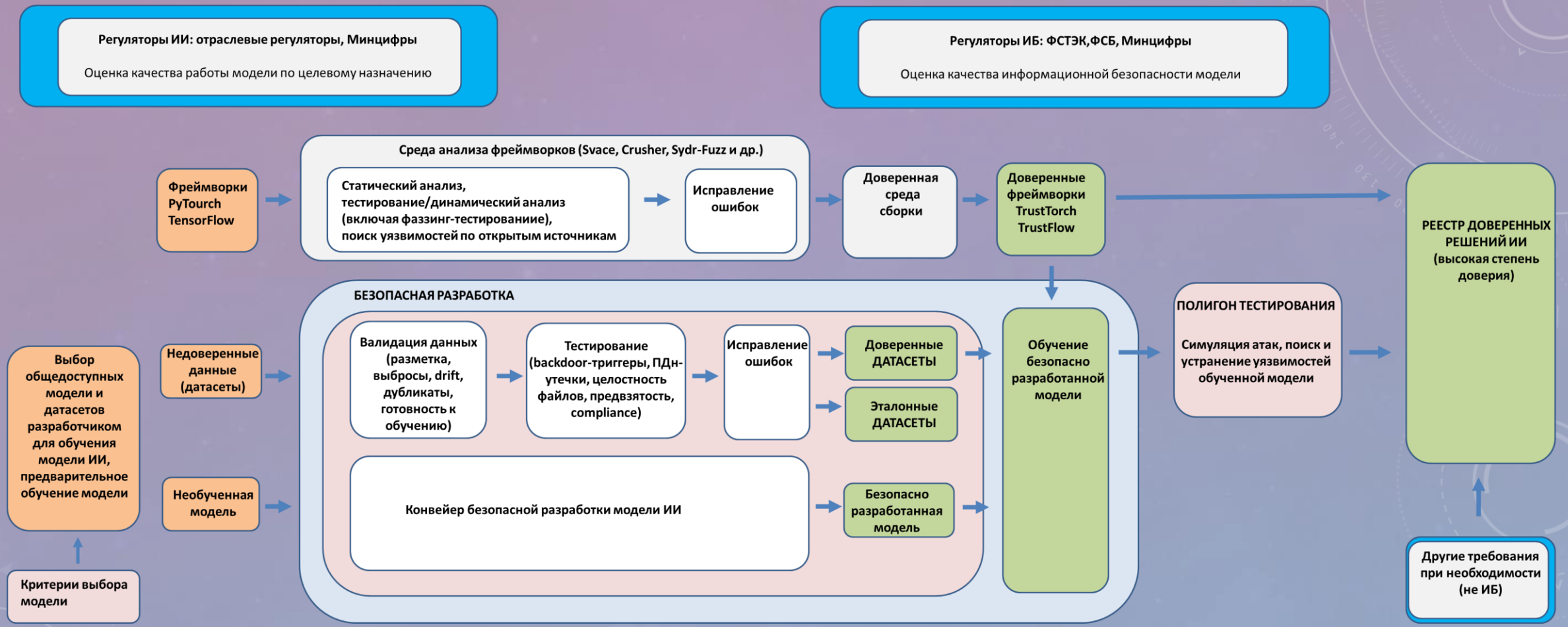
5. Извлечение (Extraction)

- Допустим, кто-то иногда, делает запросы в модель и медленно собирает из нее данные. В конце концов атакующий соберет достаточно, чтобы восстановить данные на своей стороне

ТЕСТИРОВАНИЕ ТЕХНОЛОГИЙ ИИ (БЕЗ ИСПОЛЬЗОВАНИЯ В КИИ)



БЕЗОПАСНАЯ РАЗРАБОТКА И ТЕСТИРОВАНИЕ ТЕХНОЛОГИЙ ИИ ДЛЯ КИИ (ВЫСОКАЯ СТЕПЕНЬ ДОВЕРИЯ)



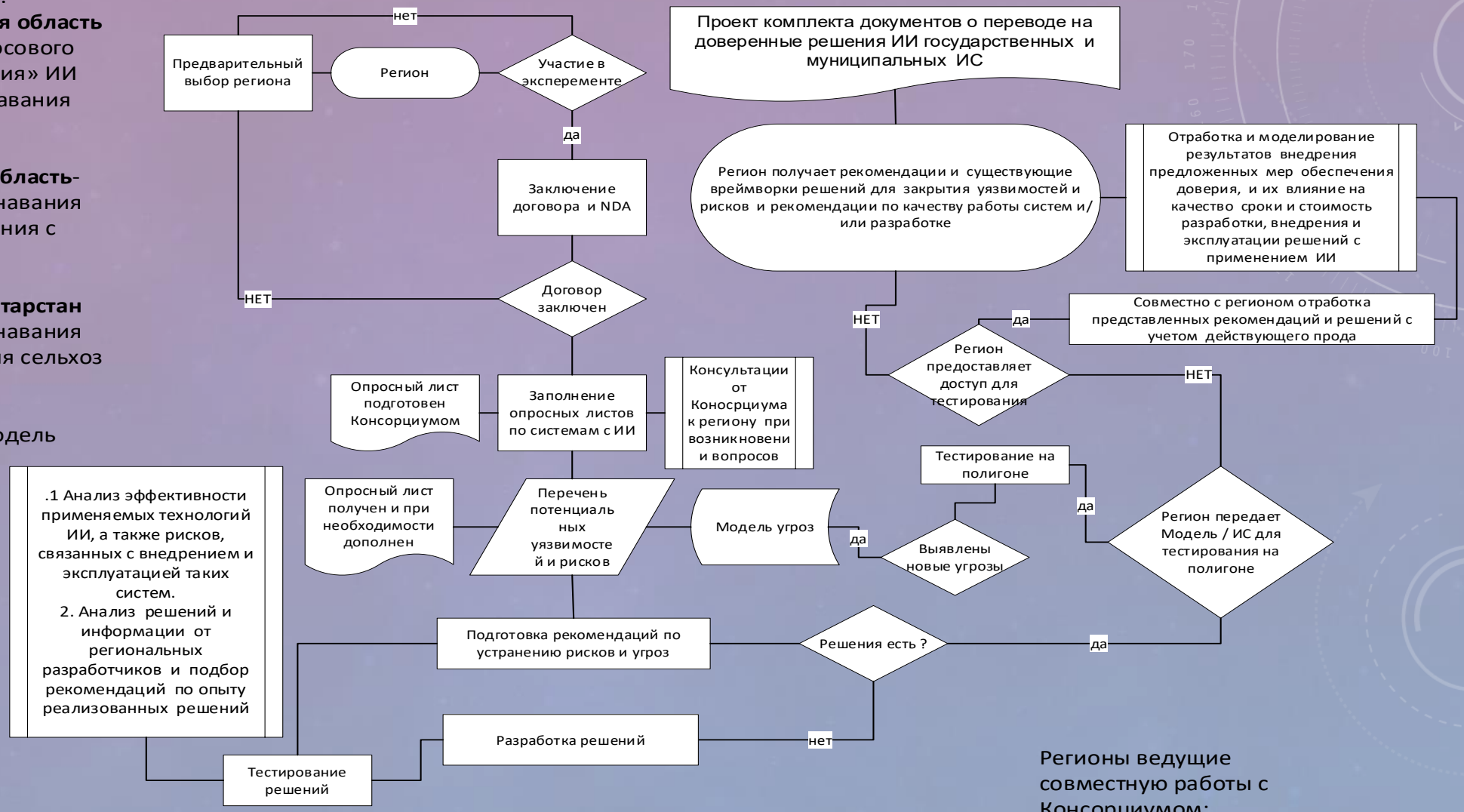
РЕГИОНАЛЬНЫЙ ПИЛОТНЫЙ ПРОЕКТ КОНСОРЦИУМА ИБ ИИ

Регионы пилоты :

1. **Новосибирская область**
 - «Система голосового самообслуживания» ИИ система распознавания речи

2. **Сахалинская область**-
-Система распознавания видео изображения с объектов

3. **Республика Татарстан**
-Система распознавания фотоизображения сельхоз земель
- «ГосПромпт» генеративная модель



Регионы ведущие совместную работы с Консорциумом:

4. Москва
5. Московская область
6. Нижегородская область

УГРОЗЫ ИИ

Страхи перед ИИ

• ИИ принимает решения, которые невозможно объяснить или предсказать
• ИИ может ошибаться, а ответственность за эти ошибки размыта
• ИИ может быть предвзятым и манипулировать мнением ИИ собирает и использует персональные данные без явного согласия
• ИИ внедряется везде, и его использование может выйти из-под контроля
• ИИ эволюционирует и изменяет алгоритмы без внешнего контроля

Конкретные риски

• Потеря контроля над критическими инфраструктурами
• Финансовые потери из-за ошибок ИИ
• Утечки и неправомерное использование данных
• Репутационные риски из-за ошибок ИИ
• Юридическая ответственность за решения ИИ на пользователе не способном управлять ИИ
• Нарушение конфиденциальности и сбор данных без согласия
• Потеря контроля над личной информацией
• Манипуляция сознанием и поведенческая реклама
• Опасность автоматизированных решений без возможности апелляции