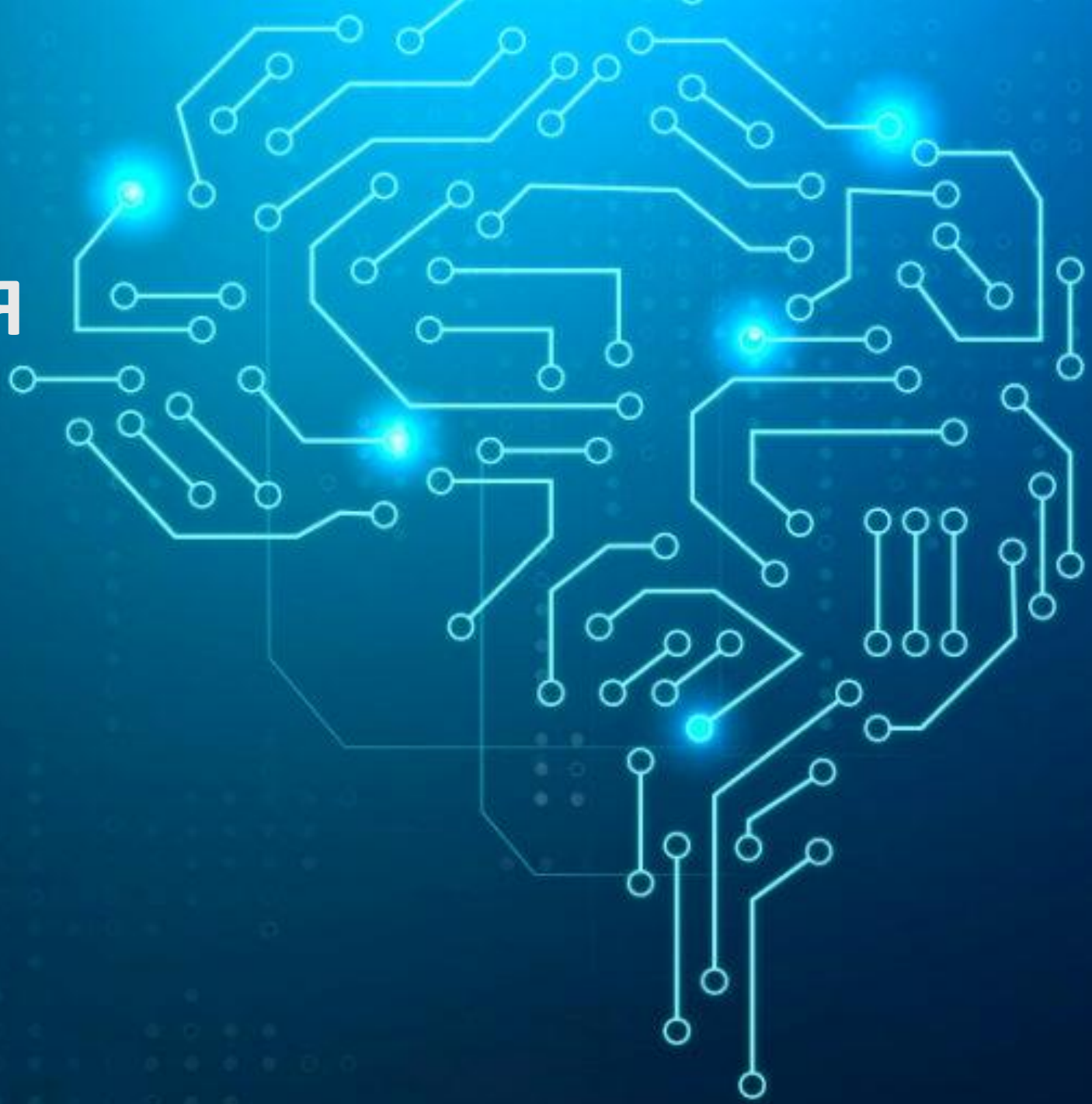
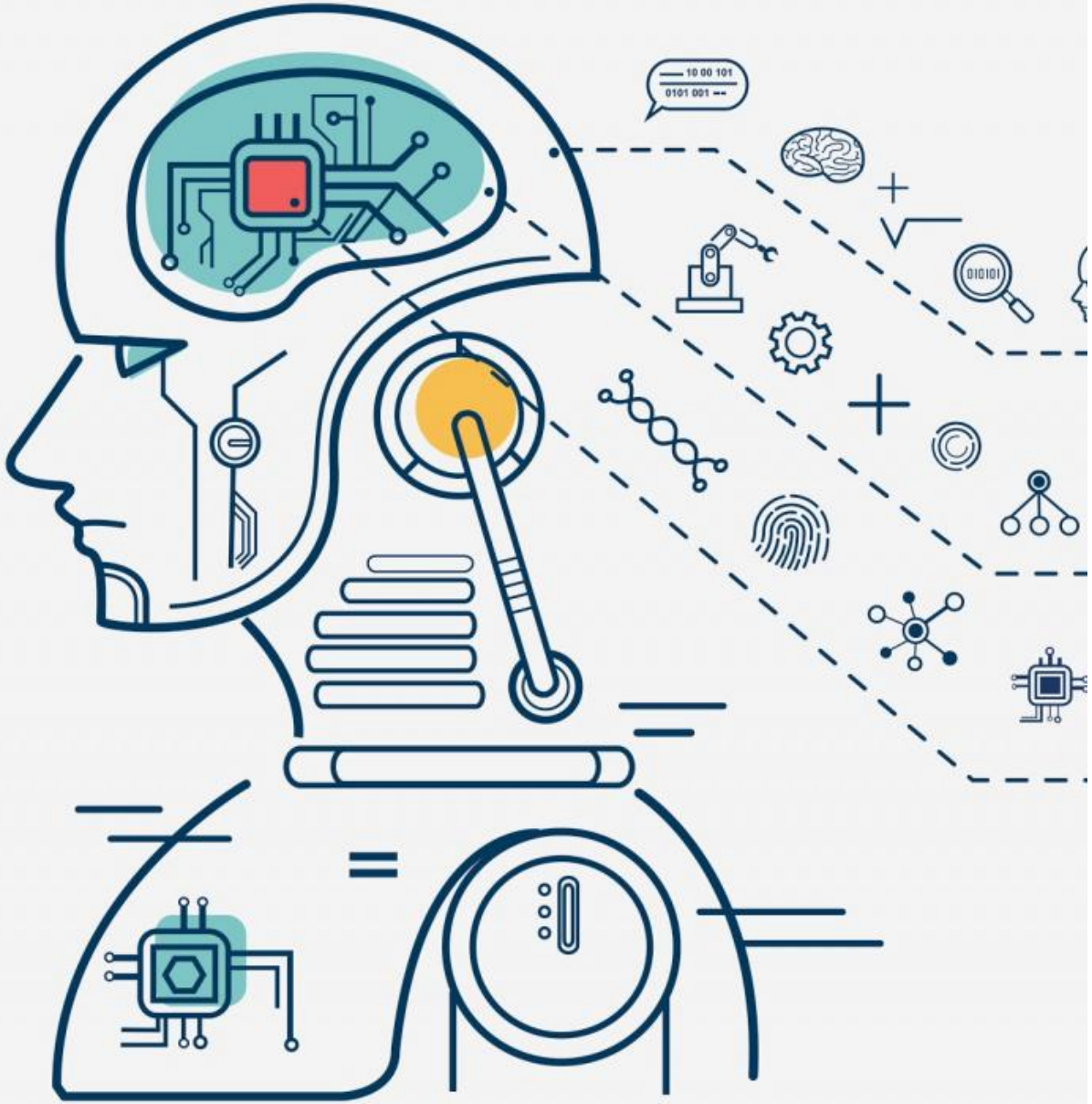
ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

«Снижении Рисков AI для бизнеса»

ДАВИД МАМЕДОВ





Бизнес считает:
Будущее наступило

- Повышение эффективности и производительности
- Улучшение принятия решений и анализа данных
- Персонализация клиентского опыта
- Инновации и конкурентное преимущество
- Снижение затрат и оптимизация ресурсов
- Улучшение устойчивости и compliance

Тем временем: Будущее

- Повышение эффективности и производительности **утечек**
- Улучшение принятия решений и анализа **на ошибочных** данных
- Персонализация **отрицательного** клиентского опыта
- Инновации **Рисков** и конкурентное преимущество **в уязвимостях**
- Увеличение **затрат** и **нехватки базовых IT** ресурсов
- **Нарушение** устойчивости и compliance

Красивый маркетинг ИИ разбивается о реальность НТ

Магии не существует

AI - не магия, а новый тип IT-актива с иной поверхностью атаки и новыми Рисками
Основные принципы:

- **ML (машинное обучение):** алгоритмы учатся на данных, чтобы прогнозировать и классифицировать.
- **Нейросеть:** математическая модель слоёв и весов которая объединяет множество простых в сложное.
- **Deep Learning:** ML на базе многослойных нейросетей.
- **NLP/LLM (Когнитивные):** понимание и генерация контента (очеловечивание)
  - **Компьютерное зрение:** распознавание в фото и видео
  - **Синтезированная речь**: AI понимает и старается говорить как человек

Нам необходим AI Требования или Хайп

- Современные реалии - AI везде: От чат-ботов до аналитики - рост на 40% ежегодно*
- Однако 80% проектов проваливаются из-за рисков*
- Простой нарратив: *“AI - как инструмент: без разницы чинить или ломать.”*
- Магии не существует, даже если инструмент очень дорогой и хороший – без управления мы только увеличиваем Риски

***данные Gartner Top Strategic Technology Trends 2023**

Адаптация

- Даже «хорошо защищённый» бизнес **раньше с этим не сталкивался**.
- Риски не cybersecurity, а как у **моделей и данных**.
- Ошибки AI: **утечки, репутация, деньги** (например штрафы 35м€ по EU Act).

Вы не верите в AI? А он верит.

Shadow-AI и Неконтролируемое использование

- *Samsung (май 2023)* - запретила ChatGPT после утечки конфиденциальных данных сотрудниками.
- *Apple (май 2023), JPMorgan (февраль 2023), Amazon (январь 2023)* ограничили ChatGPT из-за риска утечек.
- *По данным Cyberhaven:* 11% сотрудников по всему миру вставляют в ChatGPT чувствительную информацию

Утечка данных через промпты и интеграции

- *Samsung (2023)* - слив части кода через промпт, тот же ChatGPT.

Галлюцинации и неверные ответы

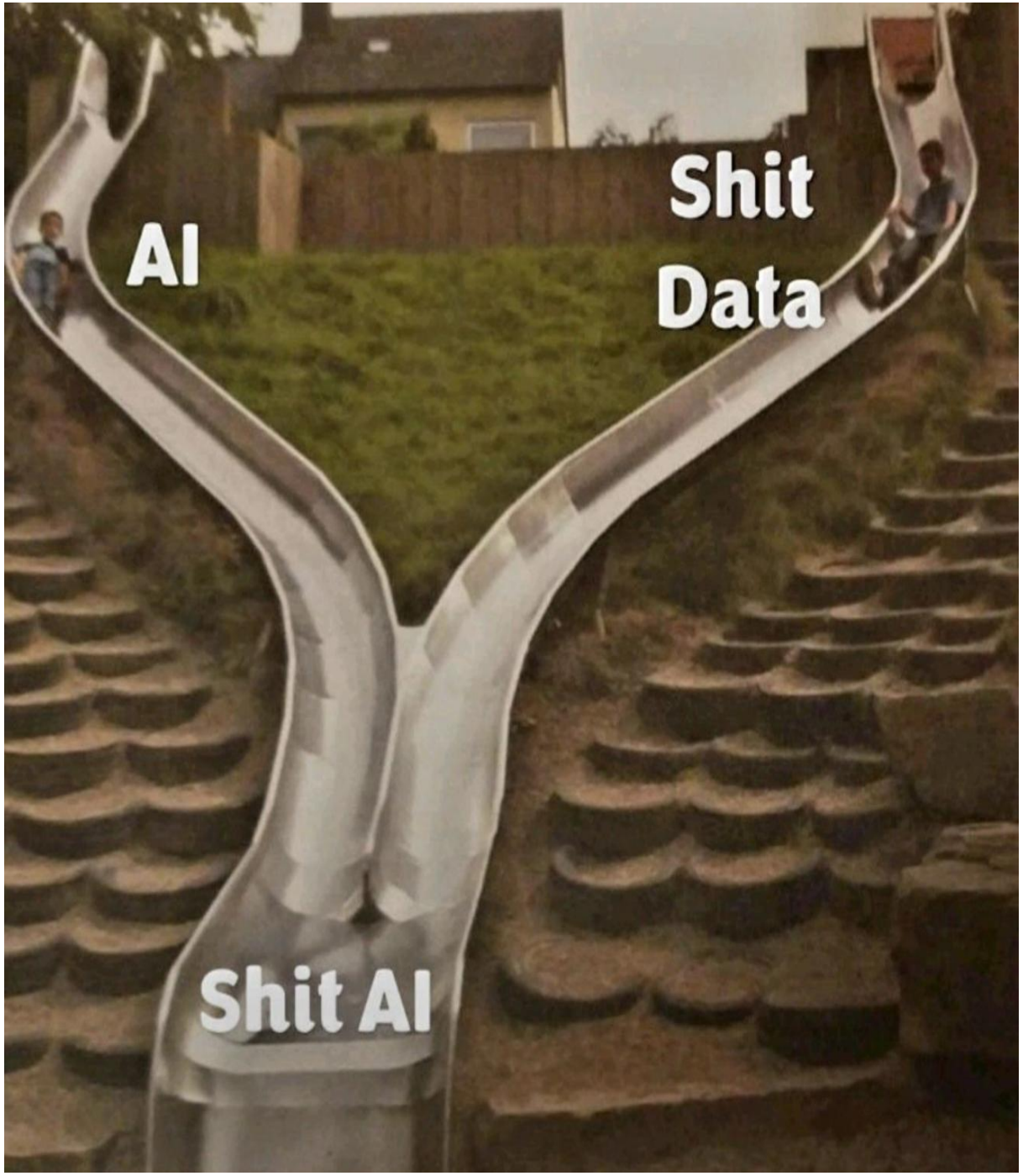
- *Air Canada (февраль 2024)*: чат-бот «выдумал» политику компенсаций - суд обязал авиакомпанию выплатить клиенту компенсацию и штраф за «несуществующие» кейсы в судебной записке.

Bias/дискриминация (регуляторные риски)

- *Meta (Facebook) (июнь 2022)* : DOJ обвинило алгоритмы доставки рекламы жилья в дискриминации; достигнуто «прецедентное» урегулирование - переделать систему под надзором суда (с 2023 по 2026).
- Кабинет Рютте подал в отставку в январе 2021 - скандал с «алгоритмической охотой» на пособия в Нидерландах показал масштаб социального ущерба и критику регуляторов.

Вендор-зависимость: отключения и блокировки

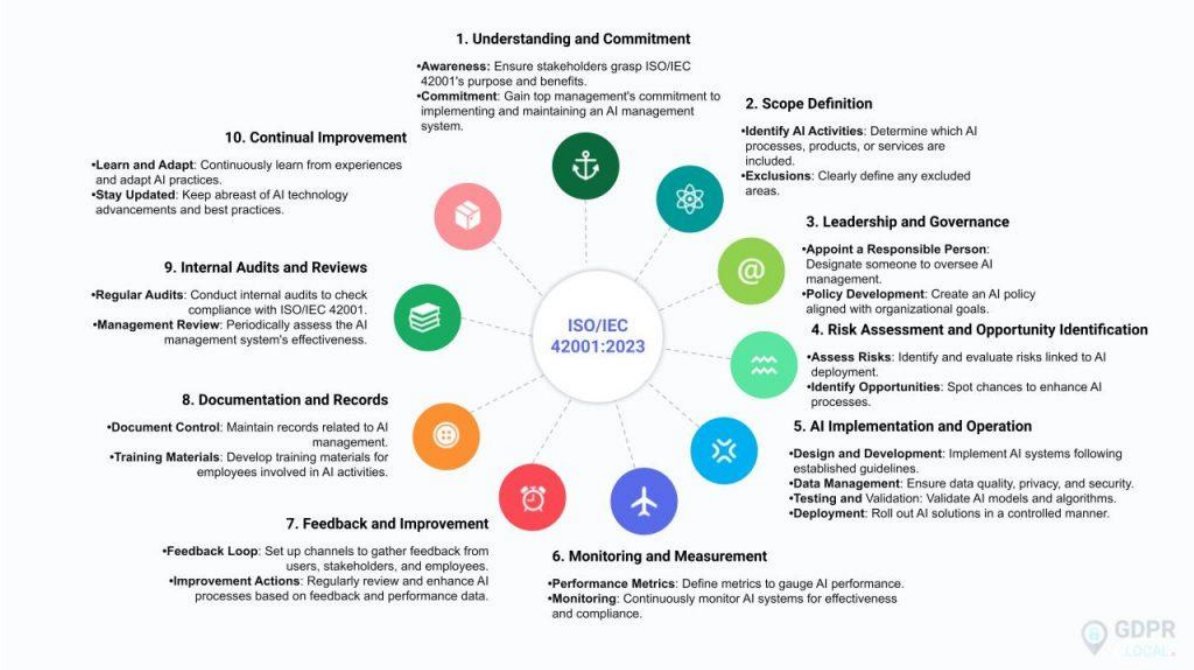
- *OpenAI/ChatGPT (2023-2025)*: масштабные сбои и 502/503 - официально признанные; часть из-за DDoS.
- Италия (*март–апрель 2023*) приостанавливала ChatGPT и штрафовала OpenAI за нарушения приватности.

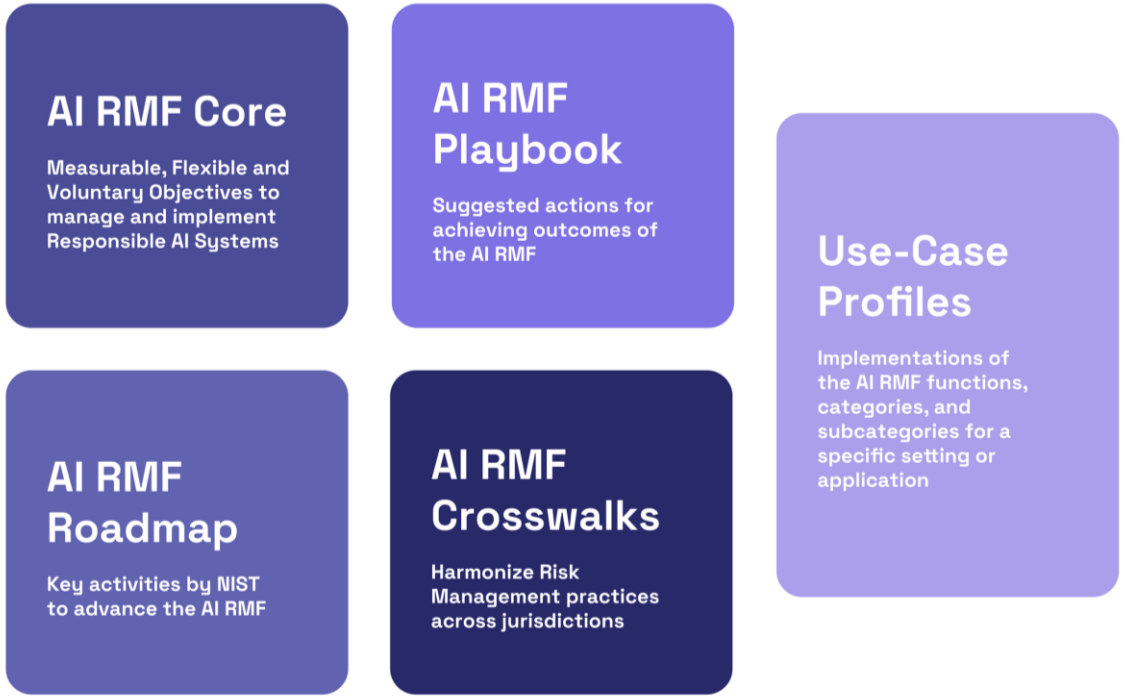


Карта Рисков смещается

- **Данные:** утечки через промпты и API; риск для ПДн и коммерческой тайны.
- **Модели:** атаки через промпт, «отравление» данных, дрейф качества и галлюцинации.
- **Поставщики:** зависимость от облачных вендоров, реальность SLA, где физически хранятся данные.
- **Комплаенс и этика:** дискриминация (bias), требование объяснимости решений, аудит следов.
- **Операционные:** отсутствие логов, непрозрачные инциденты AI.

Сейчас безопасность требует других процедур, чем классическая ИБ.





Принципы безопасности для AI

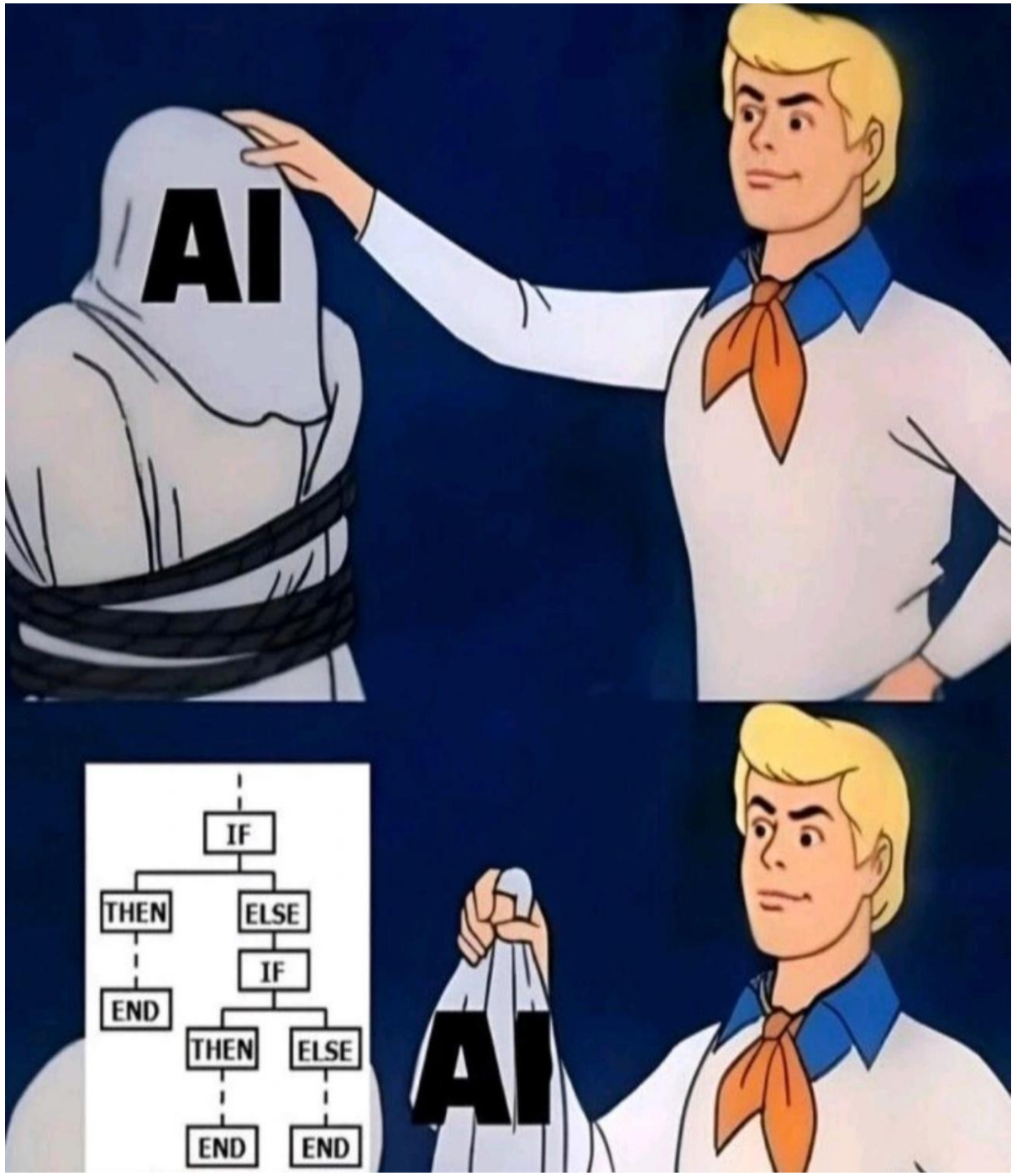
- Принципы **ISO 42001 (AIMS)**: Контроль через построение системы менеджмента AI.
- **NIST AI RMF** (Управление Рисками AI): Govern (Глобальное управление); Map (Картирование); Measure (Измерение); Manage (Управление lifecycle)
- **Gartner AI TRiSM**: наборы практик и технологии которые делают AI управляемым.
- **EU AI Act** и **EU CRA** (вступают поэтапно (2025-2027): Регуляции требуют контроля для избежания вреда; Без compliance AI - вредный инструмент.

Понимание рисков и уязвимостей. Управляемость

Базовые требования и контроли

- **Политика по AI**: понятный процесс одобрения кейсов, конфиденциальность.
- **Карта данных:** где лежат, кто видит, что утекает; запрет на отправку чувствительных данных без контроля.
- **Шлюз AI:** одна «точка входа» с фильтрами токсичности, автоматическим скрытием ПДн, лимитами запросов и блоком секретов.
- **Объяснимость и мониторинг:** журналы промптов и ответов, метрики качества, алерты, контроль дрейфа и галлюцинаций.
- **Аварийный режим:** отключения (kill-switch), план Б (fallback): деградация функционала, резервные правила, ручной режим.

Превращаем AI из хайп темы в управляемый сервис с SLA, метриками и ответственными.



Что делать бизнесу сейчас

- Идентифицировать все AI-сервисы: Кто и где использует, какие данные туда попадают и выходят.

- Вести **реестр моделей и API** и назначить владельцев.

- Включить **логирование промптов и ответов**, контроль доступов.

- **Оценивать риски: DPIA и AI-risk assessment** до запуска и в жизненном цикле.

- **Подключать Роли с первого дня:**
ИБ, DPO, CDO и юристов.

- **Для собственных моделей**: планировать **Kill-switch и fallback**

- Для приобретаемых моделей: SLA, локализация данных, права на аудит, и чёткий план выхода и миграции.

Это «база». Без неё дальше нельзя.

AI: инструмент с контролем

- **Принципы стандартов для safe AI.**
  Принципы ISO 42001 (AIMS), NIST AI RMF и конечно интеграция с старыми-добрыми ISO 27001 (ISMS), ISO 27005 (process) - ключ к контролю, чтобы AI не вредил.
- **Без контроля вред; с контролем - преимущество.**
  AI - инструмент, как нож: полезен в руках шеф-повара, но без контроля режет бизнес. Управляйте рисками по стандартам - и превратите вред в успех!

Превращаем ИИ из хайп темы в управляемый сервис.

Благодарю за внимание

David Mamedov

- ISO 27001
- ISO 27005
- Conversational AI Ensuring Compliance and Mitigating Risks LFS120
- EU Cyber Resilience Act CRA LFEL1001
- AI Security & Governance
- Data Security Posture Management

От грядущего не скрыться.
Я старался…правда старался.