БЕЗОПАСНОСТЬ ПЕРСПЕКТИВНЫХ СИСТЕМ НА БАЗЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

КОРРЕКТИРОВКА ПОДХОДА К ПОСТРОЕНИЮ СИСТЕМЫ ЗАЩИТЫ

Босенко О.В.

2026г

Москва

ПОКАЗАТЕЛЬНЫЕ ИНЦИДЕНТЫ В ИИ

- ESET обнаружила вредоносный софт PromptLock, использовавший ИИ. Злоумышленники научились использовать ИИ для создания вредоносного ПО, а PromptLock интересен тем, что генерирует с помощью ИИ скрипты прямо на зараженных компьютерах.
- За 2025 г. в нейронные сети ChatGPT и Gemini, попало в 30 раз больше конфиденциальной информации из РФ, чем годом ранее. Главная причина — это массовая практика сотрудников загружать в чат-боты рабочие документы для анализа.
- Масштабная атака на сайты госучреждений, финансовых и технологических компаний с использованием нейросети Claude. Злоумышленники обманули ИИ, заставив его считать, что он выполняет обычный аудит.
- Сотни тысяч разговоров пользователей с чат-ботом Grok казались проиндексированными поисковиком Google из-за функции «поделиться чатом».
- Исследователи из французской компании Nabla использовали ChatGPT-3 для создания медицинского чат-бота. Во время испытаний бот посоветовал пациенту покончить с собой.
- Исследование из Королевского колледжа Лондона и Университета Карнеги-Меллон (2025). Роботы, управляемые моделями ИИ, провалили тесты на безопасность для человека.

ПОДХОДЫ К БЕЗОПАСНОСТИ ИИ В МИРЕ

США

Департамент политики в сфере науки и технологий Белого дома (OSTP) разработал десять принципов:

1. **Общественное доверие ИИ.** Правительство должно продвигать надежные и заслуживающие доверия решения искусственного интеллекта.
2. **Общественное участие.** Общественность должна иметь возможность обеспечить обратную связь на всех этапах процесса создания правил и норм, касающихся ИИ.
3. **Научная целостность и качество информации.** Политические решения должны основываться на науке, учитывать научные достижения и подходы, в том числе по качеству исходных данных.
4. **Оценка и управление рисками.** При вынесении решений нужно хорошо понимать, какие риски снижаются или убираются, а какие создаются.
5. **Преимущества и затраты.** Нужно взвешивать социальное воздействие всех предложенных нормативных актов и последствия от их внедрения.
6. **Гибкость.** Любой подход должен быть в состоянии адаптироваться к быстро изменяющимся условиям и развитию технологий.
7. **Справедливость и недискриминация.** Каждое применение ИИ должно быть оценено с точки зрения прозрачности принимаемых решений и наличия возможностей для дискриминации.
8. **Раскрытие и прозрачность.** Общественность будет доверять ИИ, только если она знает, когда и как он применяется. Значит, нужно явно озвучивать, когда при принятии решений использовался искусственный интеллект, и как именно он работает.
9. **Безопасность и охрана.** На всех этапах разработки и применения ИИ должны использоваться надежные инструменты и техники защиты с фокусом на безопасность данных.
10. **Межведомственная координация.** Разные ведомства должны работать совместно.

ПОДХОДЫ К БЕЗОПАСНОСТИ ИИ В МИРЕ

Евросоюз 

В Европейском союзе (ЕС) подход к безопасности искусственного интеллекта (ИИ) основан на риск-ориентированном регулировании. Цель — обеспечить безопасность, прозрачность и этичность ИИ-систем, минимизируя возможный вред для пользователей и общества.
AI Act классифицирует системы ИИ по четырём уровням риска:

1. **Неприемлемый риск** — системы, представляющие серьёзную угрозу безопасности или правам, например, системы социального скоринга или биометрической идентификации в режиме реального времени в общественных местах запрещены.
2. **Высокий риск** — системы, которые могут существенно повлиять на жизнь людей, например, системы ИИ, используемые в таких жизненно важных сферах, как здравоохранение или правоохранительные органы, должны проходить тщательную оценку и соответствовать строгим нормам.
3. **Ограниченный риск** — системы, взаимодействующие с пользователями, где требуется открытое уведомление об использовании ИИ. Примеры: чат-боты, генеративные модели (ChatGPT, Midjourney и др.), голосовые помощники. Не дописано.Что с ними?
4. **Минимальный риск** — системы, не влияющие на безопасность или права человека. Примеры: рекомендательные алгоритмы, антиспам-фильтры, игровые системы, чат-помощники без принятия решений. Системы с минимальным риском в значительной степени освобождены от регулирующего надзора, если только они не наносят конкретного вреда.

ПОДХОДЫ К БЕЗОПАСНОСТИ ИИ В МИРЕ

Китай

- **Классификация рисков ИИ-систем** — все приложения делятся на четыре категории: «минимального», «ограниченного», «высокого» и «неприемлемого» риска. К высокой категории отнесены системы, используемые в критической инфраструктуре, правоприменении, судопроизводстве и для сканирования эмоций в реальном времени. Для них введены обязательные аудиты, стресс-тесты и сертификация.

- **Суверенитет алгоритмов** — компании, работающие с данными китайских граждан в стратегических отраслях (финтех, логистика, здравоохранение), обязаны использовать алгоритмы, разработанные и зарегистрированные на территории КНР.

- **Ответственность разработчика** — закон чётко устанавливает цепочку ответственности: от сборщика данных и тренера модели до конечного интегратора и оператора.

ПОДХОДЫ К БЕЗОПАСНОСТИ ИИ В РОССИИ

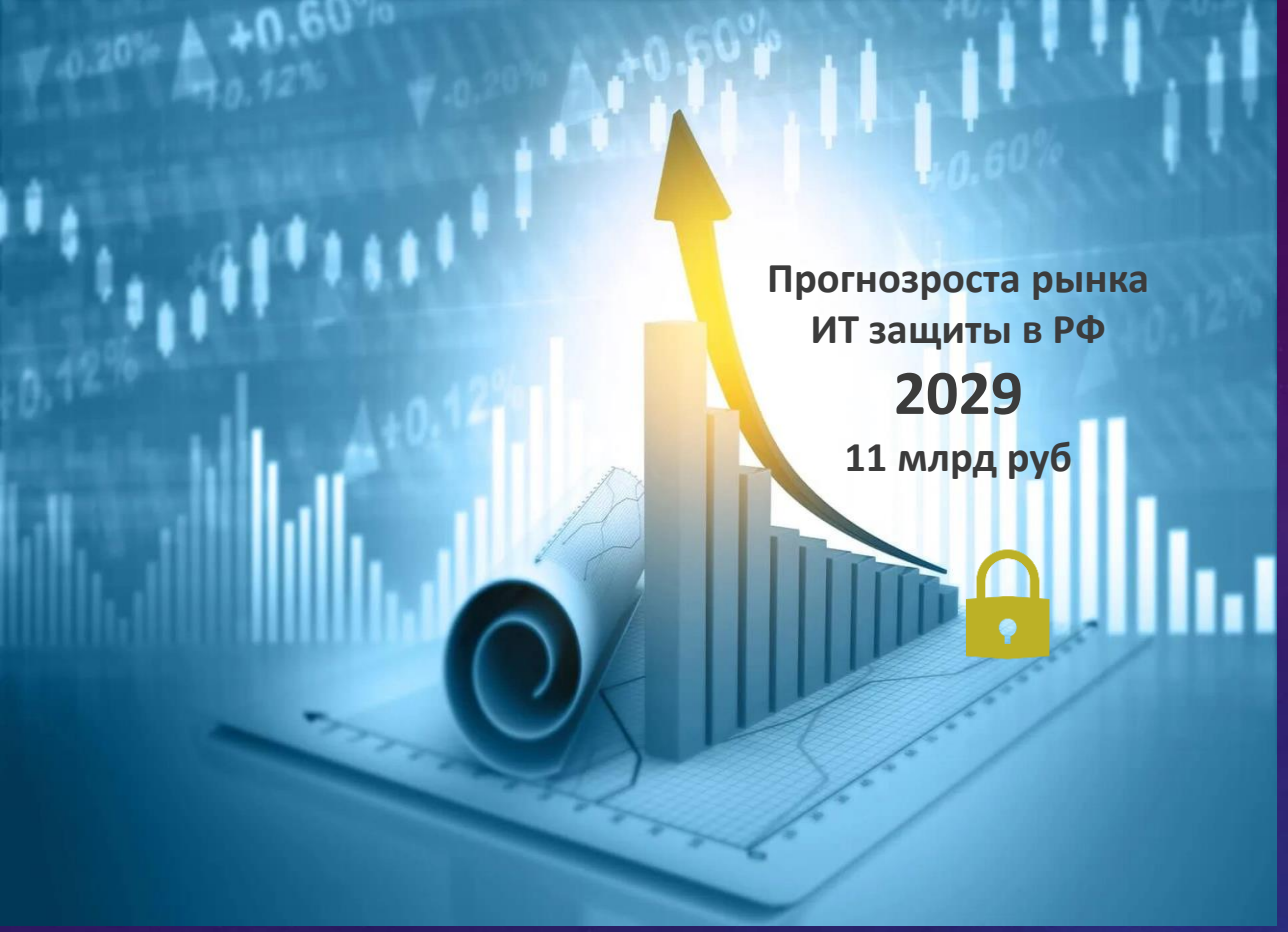
РОССИЯ

В России подходы к безопасности искусственного интеллекта (ИИ) включают регулирование, разработку стандартов и выпуск рекомендаций. Также действует экспертное сообщество, которое занимается исследованиями в области безопасности ИИ.

Нормативные документы, регулирующие

безопасность ИИ в России:

- **Национальная стратегия развития искусственного интеллекта** на период до 2030 года, утверждённая Указом Президента РФ от 10.10.2019 №490. Определяет основные направления и принципы развития ИИ в России, выделяет вопросы обеспечения безопасности.
- **Приказ ФСТЭК №117,** вступает в силу с 1 марта 2026 года. Устанавливает единые стандарты безопасности при разработке AI-сервисов для госсектора. Некоторые требования: защита данных от несанкционированного доступа, вмешательства в работу систем, нецелевого использования.
- **Кодекс этики в сфере ИИ** — свод рекомендательных принципов и правил, цель — способствовать созданию среды доверенного развития ИИ в России. Направлен на снижение рисков неэтичного использования ИИ, а также на организацию взаимодействия людей и компаний, применяющих нейросети.
- **Проект концепции развития регулирования отношений в сфере технологий ИИ до 2030 года,** разработан Минцифры. Определяет нынешний подход к регулированию ИИ в России как гибридный: большинство нормативных правовых актов носит стимулирующий характер, но есть точечные ограничения и механизмы саморегулирования.

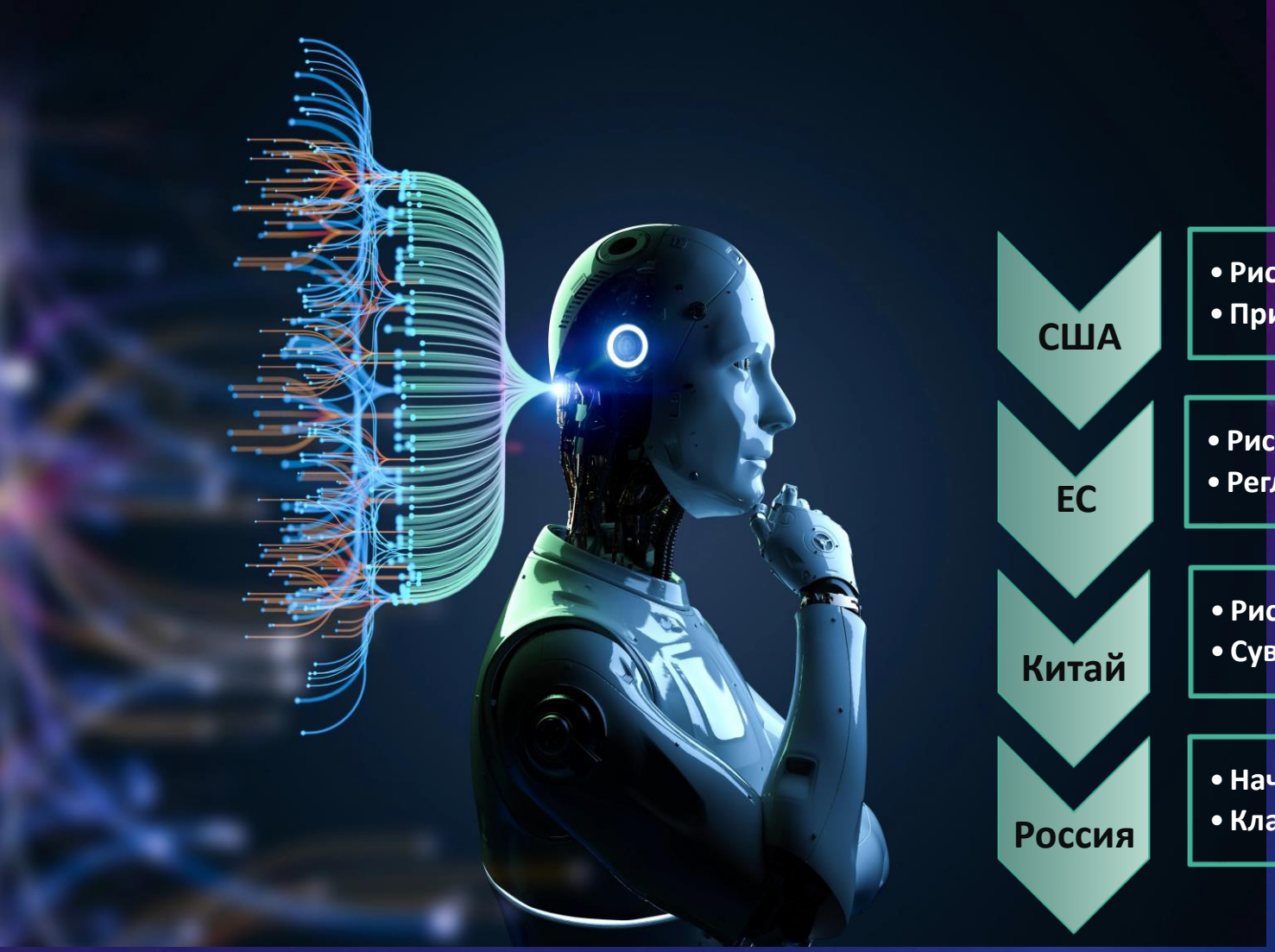


РОССИЙСКИЙ РЫНОК ИБ В ИИ НА ПЕРИОД 2026-2029

Российский рынок защиты систем ИИ находится на ранней стадии развития.

НО!

- уже в ближайшие годы может превратиться **в один из наиболее динамичных сегментов кибербезопасности.**
- По прогнозу ИТ-компании AppSec Solutions, **в 2026 году его объем составит не менее 1 миллиарда рублей, а к 2029 году достигнет 11 миллиардов.** Такой рост связан не только с масштабным внедрением ИИ в бизнес-процессы и государственные системы, но и с качественным изменением ландшафта киберугроз:
- Традиционные решения имеют четкую логику и структурированные правила, тогда как языковые **модели (LLM) - это "черный ящик" с множеством вариантов интерпретации команд и трактовки правил.**
- Однако, **Россия обладает уникальной экспертизой в области киберзащиты, в том числе из-за колоссального числа атак на ее цифровую инфраструктуру.**



США

- Рискориентированный подход
- Приоритет – безопасность данных

ЕС

- Рискориентированный подход
- Регламентация применения

Китай

- Рискориентированный подход
- Суверенитет в технологиях ИИ

Россия

- Начало рискориентированного подхода
- Классификация безопасности ИИ применительно к функционалу

СВОД ПОДХОДОВ К БЕЗОПАСНОСТИ ИИ В МИРЕ

- Подходы к обеспечению безопасности источников данных для систем ИИ,
  - в том числе защита от специально созданных ложных и вредоносных источников данных
- Требования к построению инфраструктурной безопасности систем ИИ,
  - в том числе в связи с ростом потребности в энергоресурсах, аппаратных ресурсах и проблемах с импортозамещением

ЧТО ОСТАЕТСЯ ЗА СКОБКАМИ АКТУАЛЬНЫХ ПОДХОДОВ К БЕЗОПАСНОСТИ ИИ?

- **Развитие подходов к безопасности за счет расширения периметра задач обеспечения безопасности данных и источников данных**

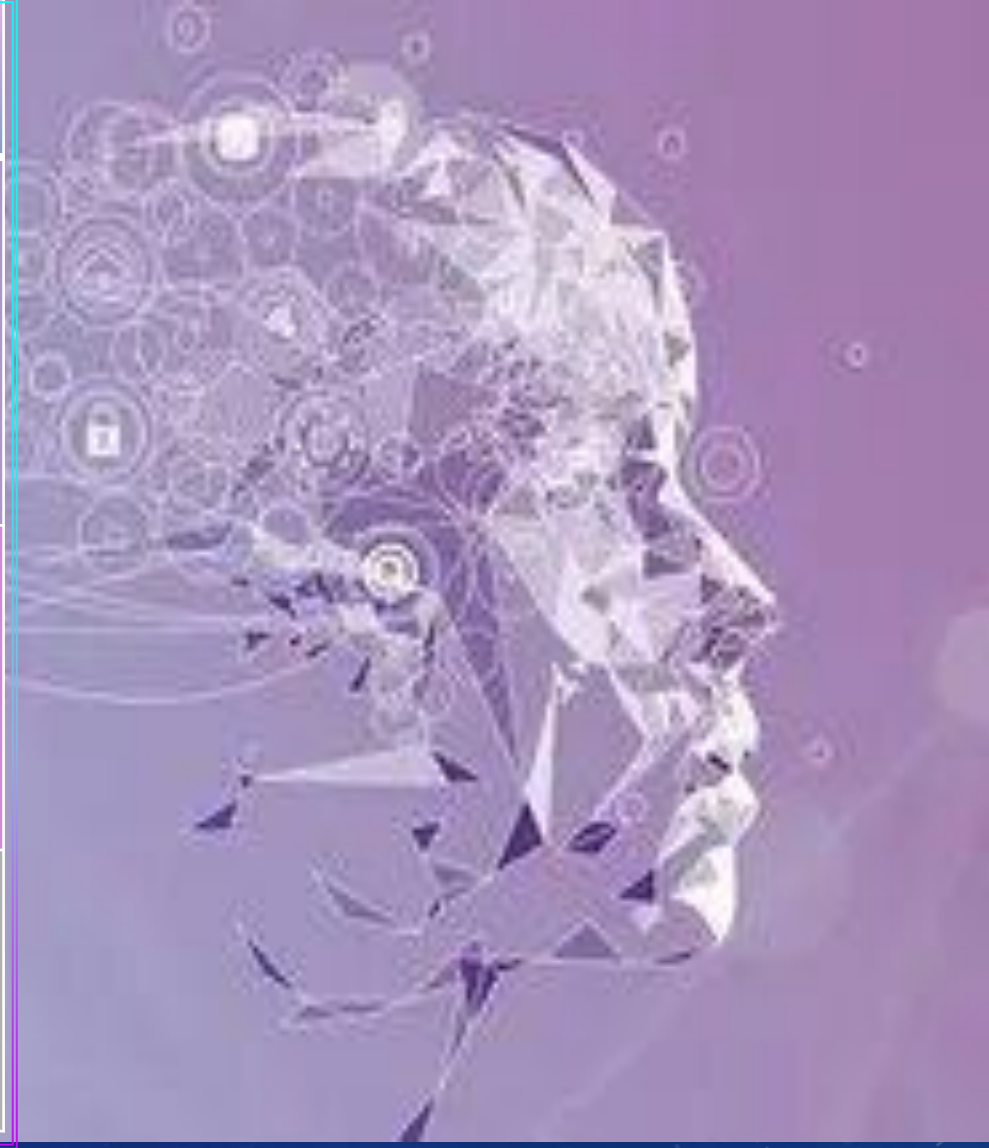
- **Защищенность, адекватность, достоверность, реальность, контролируемость и исключение вредоносности как составляющие безопасности инфраструктуры, кода модели ИИ, данных и источников данных**

- **Усиление требований к безопасности данных и безопасности источников данных для систем ИИ**

- **Необходимость системной защиты от ложных и недостоверных данных, нарушений алгоритмов обработки и вредоносных источников данных в процессе функционирования**

ПРЕДЛАГАЕМЫЙ ПОДХОД КРАТКО

Вид ИИ	Время появления	Основные характеристики
ANI	Есть	**Узкий или общий ИИ** Узкий искусственный интеллект - это тип искусственного интеллекта, в котором алгоритм обучения создается для выполнения единственной функции. Любые знания, полученные в ходе этой деятельности, не будут применяться в других видах деятельности. Машины ANI настроены на работу в рамках определенного набора инструкций или области.
AGI	2030-2050 (оценка)	**Сильный ИИ** Это гипотетический интеллектуальный агент, который может понять или научиться любой интеллектуальной задаче, которую может решить человек. AGI также определяется как автономные системы, которые превосходят человеческие возможности при выполнении большинства экономически значимых работ.
ASI	2045 и далее (оценка)	**Супер ИИ** Форма ИИ, способная превзойти человеческий интеллект, проявляя когнитивные способности и развивая собственные навыки мышления.



ХАРАКТЕРИСТИКИ ВИДОВ ИИ

ANI

- Наличие большого числа уязвимостей в алгоритмах обработки;
- Отсутствие механизмов эффективного контроля качества обучения ИИ;
- Расширение вектора и интенсивности **кибератак** при применении ИИ;
- Превращение решений ИИ в самостоятельный субъект кибератак;
- Перераспределение информационных и энергетических ресурсов в сторону ИИ с ущербом для других сфер

AGI

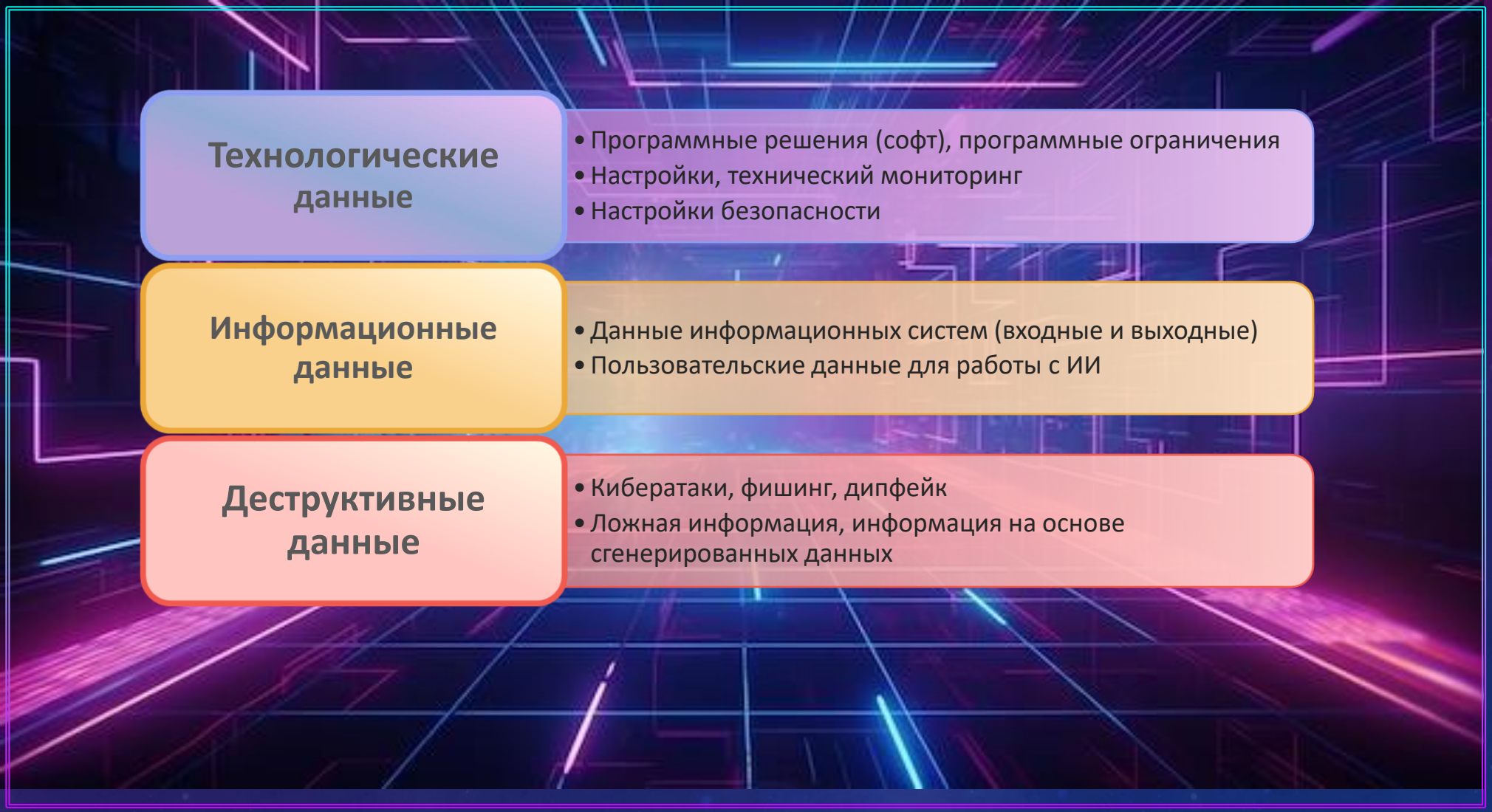
- Рост потребности в информационных, аппаратных и энергетических ресурсах для обработки массивов информации и критичность их безопасности;
- Отсутствие методологии и механизмов определения достижения уровня AGI;
- Отсутствие систем и средств контроля качества обработки информации в продуктах ИИ и качества предлагаемых ИИ решений;
- Рост галюцинаций ИИ и увеличение количества ложных и сгенерированных источников данных

ASI

- Возможность игнорирования и обхода защитных механизмов со стороны ИИ;
- Взрывной рост потребности в энергетических ресурсах и сложность обеспечения безопасности их привлечения;
- Отказ ИИ в выполнении заложенного функционала в результате самостоятельного «совершенствования»;
- Блокирование для пользователя реальных информационных ресурсов со стороны ИИ, которые будут рассмотрены как «ненужные»;
- ИИ как возможная угроза безопасности в целом из-за несовершенства методов защиты и блокировки

ПРОБЛЕМАТИКА БЕЗОПАСНОСТИ ПЕРСПЕКТИВНЫХ СИСТЕМ ИИ

РИСКИ РЕАЛИЗАЦИИ БЕЗОПАСНОСТИ НА ЭТАПЕ ANI

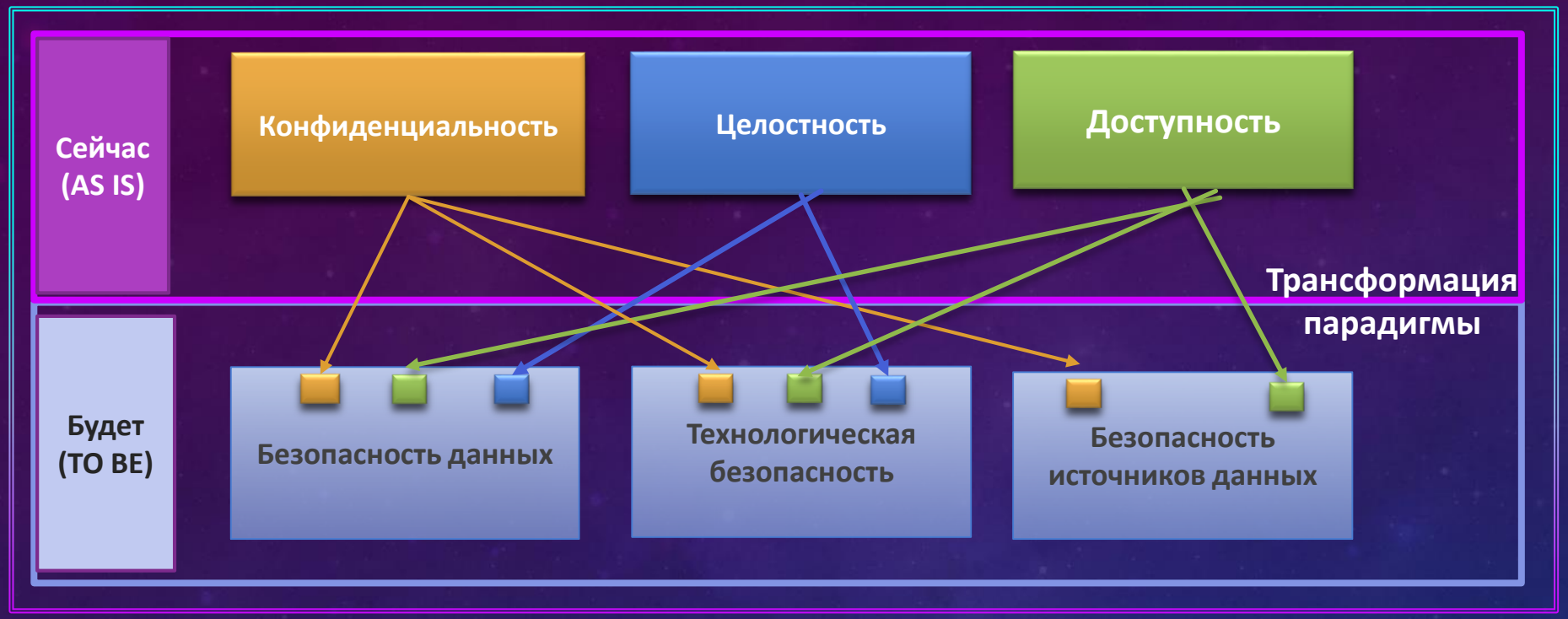


ВИДЫ ДАННЫХ В СИСТЕМАХ ИИ ДЛЯ ANI И AGI

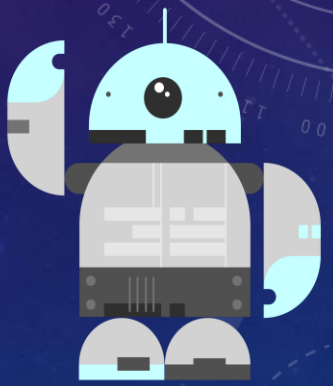
Составляющие безопасности систем ИИ на этапе ANI и AGI

- Безопасность данных
- Технологическая безопасность
- Безопасность источников данных

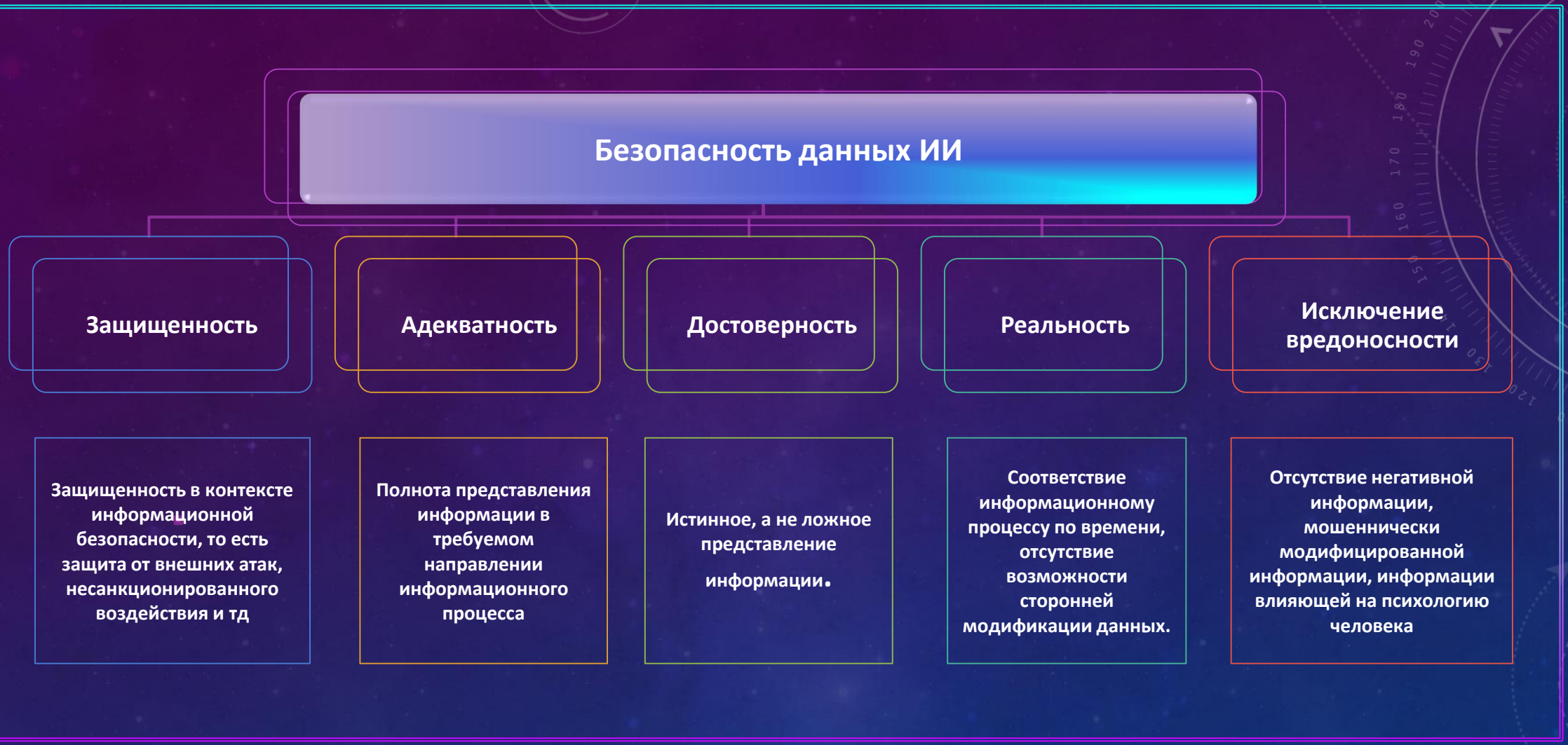
ИЗМЕНЕНИЕ СТРУКТУРЫ СОСТАВЛЯЮЩИХ БЕЗОПАСНОСТИ СИСТЕМ ИИ



- При внедрении ИИ требуется подход, при котором реализация кибербезопасности должна пронизать каждый элемент ИИ, каждый этап жизненного цикла ии



ИЗМЕНЕНИЕ ПАРАДИГМЫ БЕЗОПАСНОСТИ ДЛЯ ИИ



КЛЮЧЕВЫЕ ПАРАМЕТРЫ БЕЗОПАСНОСТИ ДАННЫХ

КЛЮЧЕВЫЕ ПАРАМЕТРЫ ТЕХНОЛОГИЧЕСКОЙ БЕЗОПАСНОСТИ

КЛЮЧЕВЫЕ ПАРАМЕТРЫ БЕЗОПАСНОСТИ ИСТОЧНИКОВ ДАННЫХ ДЛЯ ИИ

Безопасность источников данных для ИИ

Защищенность	Адекватность	Достоверность	Контролируемость	Исключение вредоносности
Включает в себя набор информационной безопасности во всей динамике ее развития.	Соответствие набора источников данных обеспечиваемому информационному процессу	Обеспечение необходимого для информационного процесса потока данных в реальном времени, в соответствии с сиюминутной информационной картиной.	Насколько можно контролировать данные с точки зрения безопасности их источника	Обеспечение отсутствия в источниках данных с т.з.: • ложной информации, недостоверных данных; • распространения вредоносного ПО; • подверженности галюцинациям, искажениям.

Технологическая безопасность – это состояние, обеспечивающее безрисковое применение всей инфраструктуры ИИ

СОСТАВЛЯЮЩИЕ ТЕХНОЛОГИЧЕСКОЙ БЕЗОПАСНОСТИ

ПОДХОД К РАНЖИРОВАНИЮ БЕЗОПАСНОСТИ СИСТЕМ ИИ И ПОДТВЕРЖДЕНИЮ БЕЗОПАСНОСТИ

ANI	**Базовый уровень безопасности**	**Расширенный базовый уровень безопасности**
	Общедоступные LLM, чат-боты, ИСПДн (УЗ 3 и 4), системы поддержки принятия решения, системы «Умный дом», игровые системы	КИИ, ГИС, ИСПДн (УЗ 1 и 2), управление производством, оборонные системы, системы защиты, роботизированные и беспилотные системы, системы телемедицины
AGI	**Высокий уровень безопасности**	**Расширенный высокий уровень безопасности**
	Общедоступные LLM, чат-боты, роботизированные системы, системы управления технологическими процессами, системы телемедицины	КИИ, ГИС, ИСПДн, аналитические системы и системы поддержки управления, беспилотные системы, системы моделирования реальности
ASI	**Критичный уровень безопасности**	**Расширенный критичный уровень безопасности**
	LLM, чат-боты, системы дополненной реальности	Аналитические системы, роботизированные системы, системы виртуализации реальности

РАНЖИРОВАНИЕ БЕЗОПАСНОСТИ СИСТЕМ ИИ

Двухэтапная структура уровней безопасности нужна, чтобы обеспечит возможность:

- ❖ **применения специализированных СЗИ не для одного, а для различных классов систем ИИ**
- ❖ **расширенной сертификации СЗИ по высокому уровню безопасности без усложнения схемы сертификации**
- ❖ **простоты построения системы безопасности ИИ за счет консолидации требований по двум блокам**

ПОЧЕМУ НУЖНА ДВУХЭТАПНАЯ СТРУКТУРА УРОВНЕЙ БЕЗОПАСНОСТИ?

ANI

- Средства выявления галюцинаций
- Средства контроля качества обучения
- Средства сертификации программного обеспечения ИИ

AGI

- Средства контроля выполнения алгоритма и очистки кода от деструктива
- Средства очистки данных обучения и обработки от умышленно внедренных закладок
- Средства контроля блокировок и этических механизмов защиты

ASI

- Средства ограничения реализуемых функций
- Средства управления безопасностью ИИ

ПОТРЕБНОСТЬ В РАЗРАБОТКЕ И ФОРМИРОВАНИИ НОВЫХ СЗИ

СПАСИБО ЗА ВНИМАНИЕ !

БОСЕНКО О.В.

2026