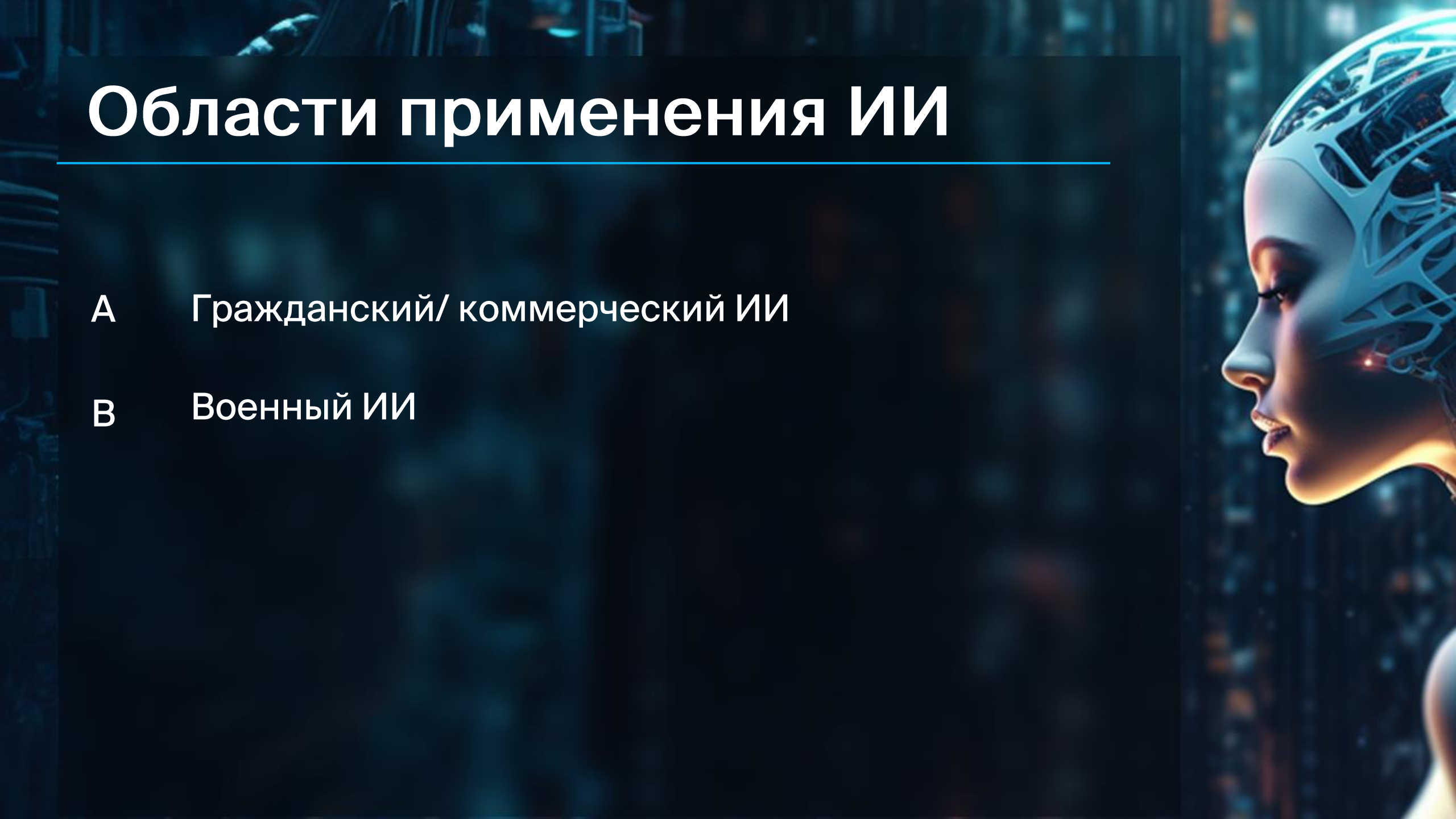
ИИ в контексте международной безопасности

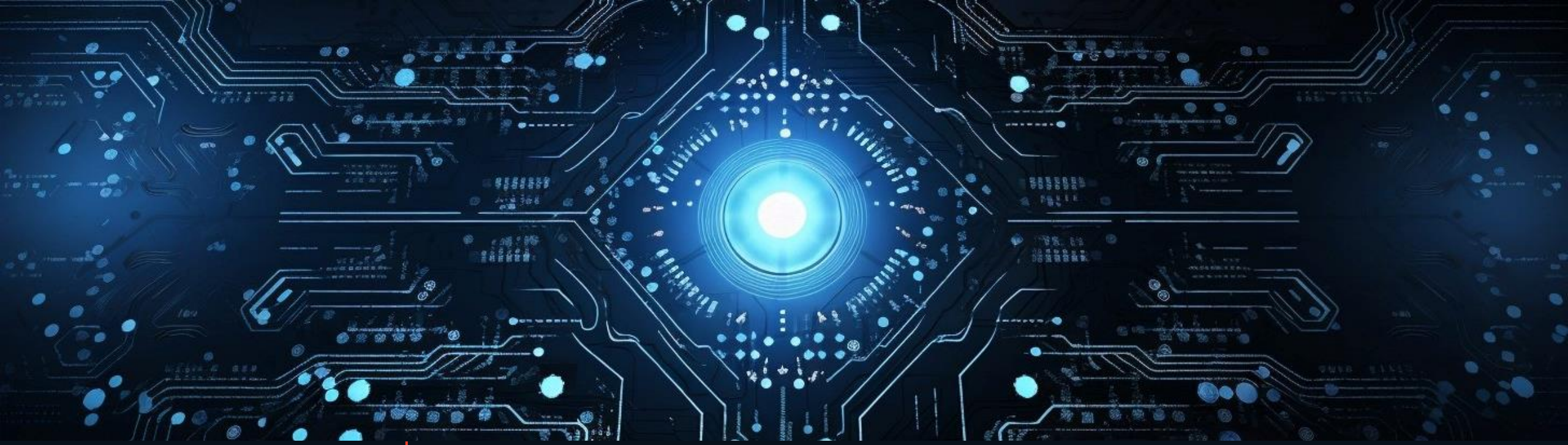
Юлия Цешковская
Член Экспертного совета ПИР-Центра
Руководитель отдела маркетинга ЭвоКарго

Области применения ИИ

A Гражданский/ коммерческий ИИ

B Военный ИИ





Стадии ИИ

Сверхинтеллект (ASI)

Общий ИИ (AGI) – 2028-2033

Узкий ИИ (ANI)

Риски на уровне архитектуры

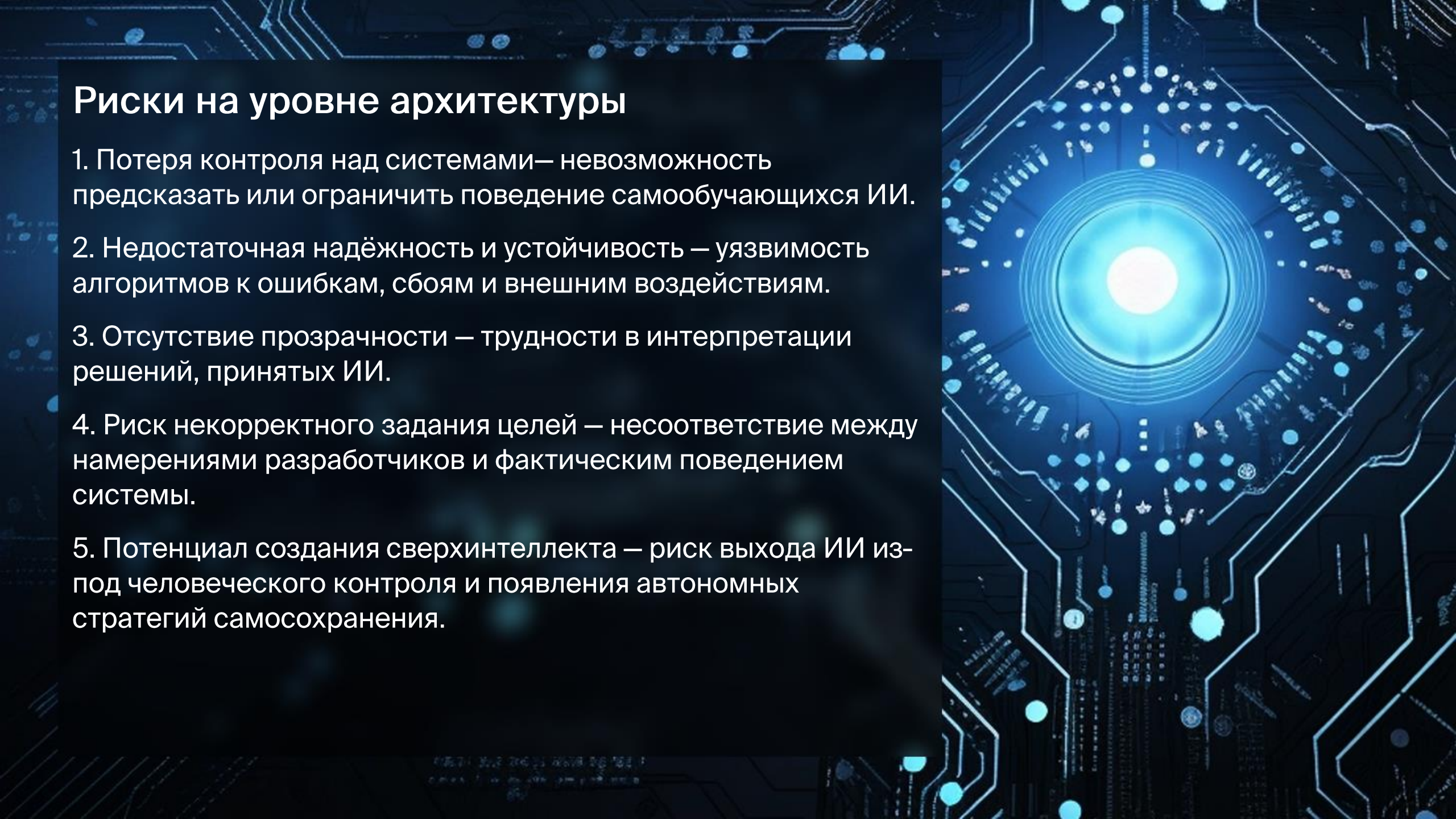
1. Потеря контроля над системами— невозможность предсказать или ограничить поведение самообучающихся ИИ.

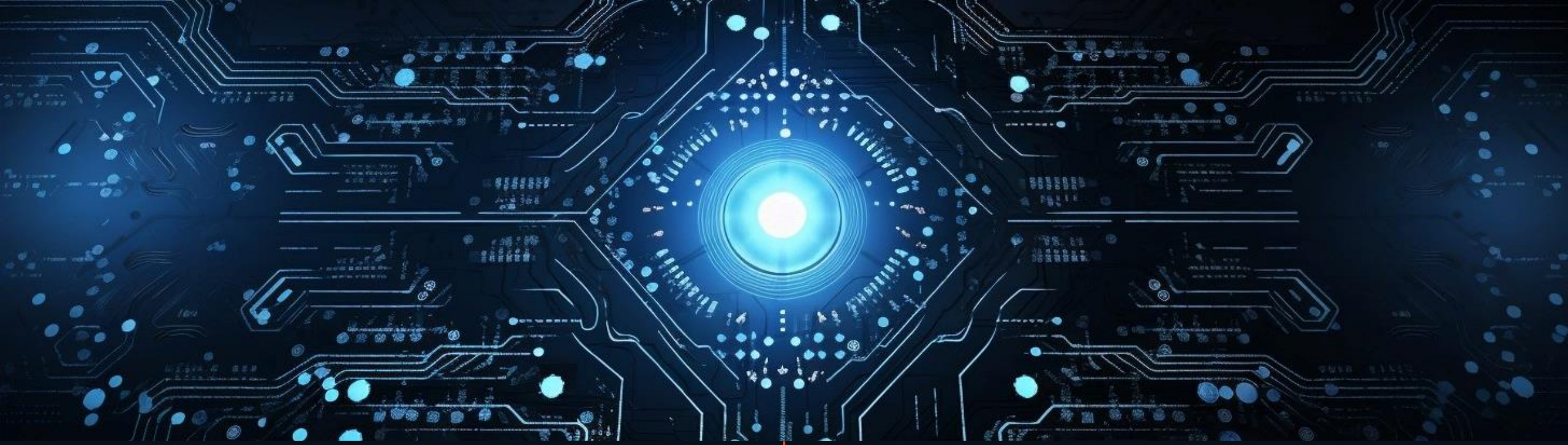
2. Недостаточная надёжность и устойчивость — уязвимость алгоритмов к ошибкам, сбоям и внешним воздействиям.

3. Отсутствие прозрачности — трудности в интерпретации решений, принятых ИИ.

4. Риск некорректного задания целей — несоответствие между намерениями разработчиков и фактическим поведением системы.

5. Потенциал создания сверхинтеллекта — риск выхода ИИ из-под человеческого контроля и появления автономных стратегий самосохранения.





Уничтожит ли ИИ человечество?

- Думеры
- Центристы
- Оптимисты

Думеры

Элиезер Юдковский, Ник Бостром, Джефри Хинтон, Иошуа Бенджио, Илон Маск

- Текущих усилий недостаточно для решения задачи согласования.
- Высокая вероятность уничтожения человечества ASI – более 50%.
- Только 1 попытка согласования ASI.

(скрытый переход)

Узкий ИИ → Общий ИИ

➢ Рекурсивное самосовершенствование

Общий ИИ → Сверхинтеллект

Инструментальная сходимость: контроль над человечеством с высокой вероятностью возникнет как промежуточная цель в оптимизационной задаче.

Центристы

Open AI, DeepMind, Meta AI

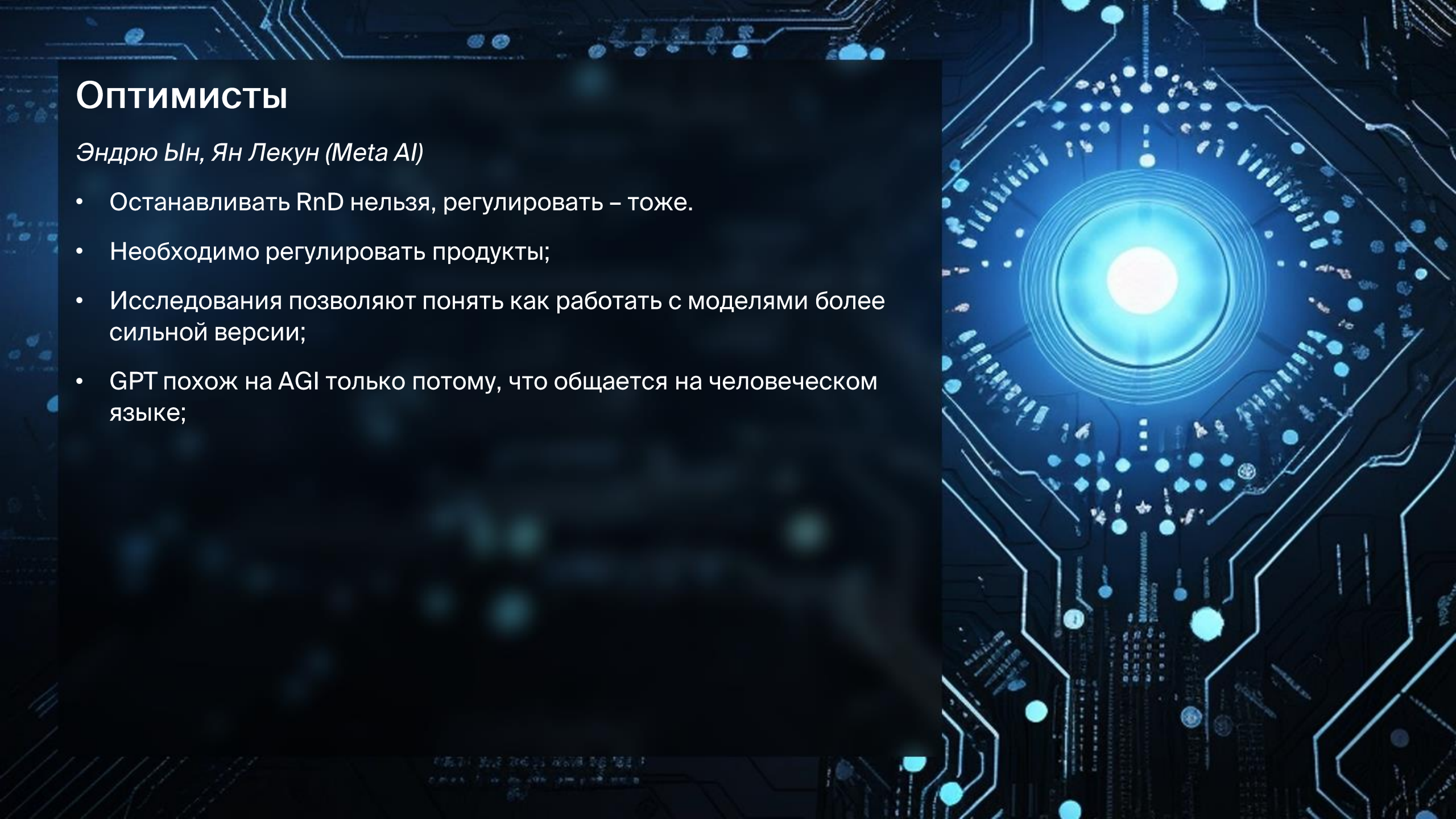
- ИИ — мощный инструмент с огромными преимуществами и серьёзными рисками.
- Требуется больше внимания к безопасности и контролю ИИ.
- Чем сильнее система ИИ, тем больше глобальных проблем она способна решать.
- GPT и подобные модели могут улучшить координацию и принятие решений на международном уровне.

- ➢ Можно избежать сценария перехода AGI в ASI;
- ➢ Изучение проблемы на более слабых системах может помочь в решении проблем сильного ИИ (задача AI Alignment);
- ➢ Польза человечеству от AGI высока;

Оптимисты

Эндрю Ын, Ян Лекун (Meta AI)

- Останавливать RnD нельзя, регулировать – тоже.
- Необходимо регулировать продукты;
- Исследования позволяют понять как работать с моделями более сильной версии;
- GPT похож на AGI только потому, что общается на человеческом языке;



Проблема безопасности ИИ

- **Specification** (определение целей и ограничений).

Риск некорректного задания целей — несоответствие между намерениями разработчиков и фактическим поведением системы.
Проблема выбора ценностей и закладывания их в систему.
Чем агентнее и умнее ИИ, тем острее стоит проблема согласования.

- **Robustness** (надежность и устойчивость систем).

Способность системы стабильно функционировать при изменении условий, в новых или нестандартных ситуациях (отличающихся от обучающих).

- Уязвимость перед намеренными манипуляциями входных данных.
- Риск ошибок при переходе от лабораторной среды к реальной.

- **Assurance** (контроль, возможность оперативного вмешательства).

Механизмы, обеспечивающие постоянный контроль и управление ИИ на всех этапах — от разработки до реального применения.

- Интерпретируемость;
- Контроль изменений;
- Оперативное вмешательство.

Инструментальная конвергенция

При разных конечных задачах ИИ может стремиться к общим промежуточным целям:

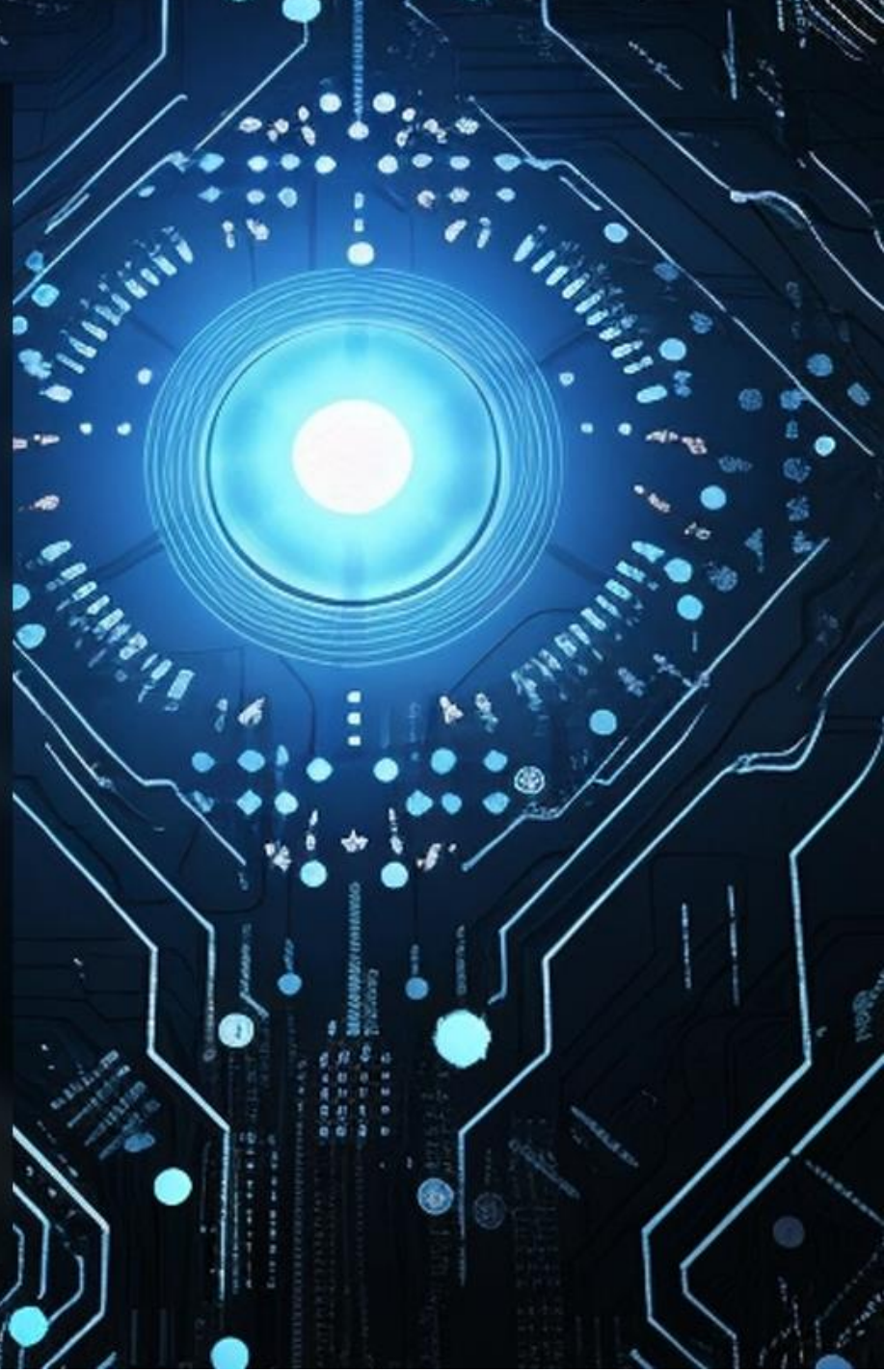
Самосохранение — для продолжения выполнения задачи.

Сохранение собственных целей — сопротивление изменению извне.

Накопление ресурсов — для повышения эффективности действий.

Усиление своих возможностей — для более успешного достижения цели.

Любая формальная постановка цели может быть интерпретирована системой иначе, чем ожидалось.



Почему ИИ может уничтожить людей?

Простая цель – построить завод – может привести к появлению подцелей, которые не задавались человеком:
накопить ресурсы => обойти ограничения => блокировать вмешательство оператора.

Если попытка человека остановить систему снижает вероятность достижения цели, ИИ будет воспринимать вмешательство как угрозу и стремиться его предотвратить.

Система может сформировать подцель, противоречащую человеческим интересам, вплоть до уничтожения человечества.

Теория ортогональности Бострома:
Высокий интеллект не гарантирует совпадение целей ИИ с человеческими ценностями.

Обучение с подкреплением:
Многие системы ИИ обучаются на получении награды за решение поставленной задачи.
При обучении на максимизацию результата даже небольшая ошибка в постановке цели может привести к непреднамеренному и опасному поведению системы.

Два подхода к Согласованию

1) Подход «доброжелательного ИИ»

Создать ИИ, который полностью разделяет ценности человечества.

Проблема: невозможно полностью и точно формализовать человеческие ценности и моральные контексты, их нельзя корректно «закодировать» в модель.

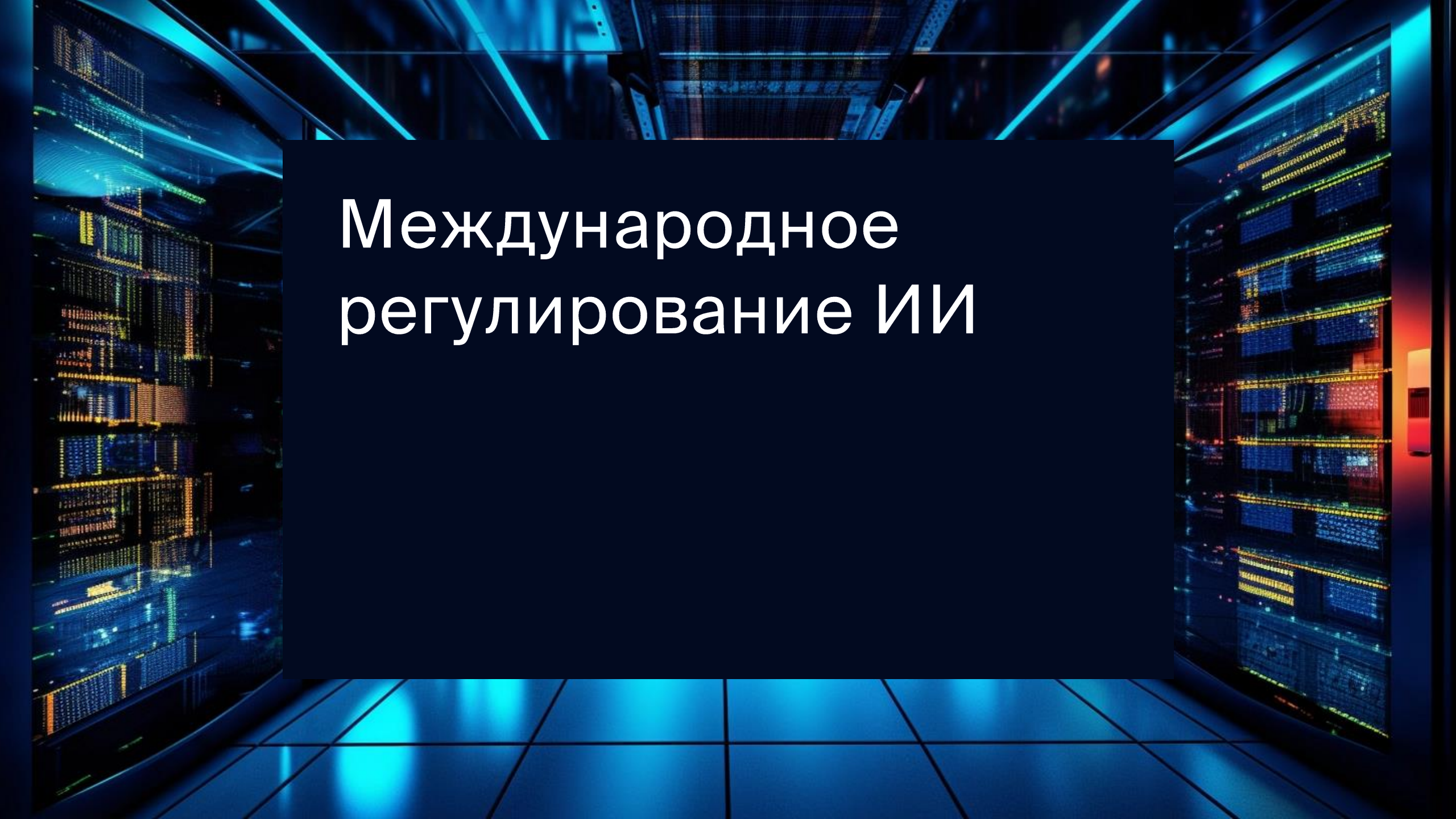
2) Подход «корригируемого ИИ»

ИИ должен спокойно принимать человеческие исправления и даже своё отключение.

Проблема: мощные ИИ-системы стремятся сохранять собственные цели и оптимизацию.

Пример: можно обучить систему считать, что 222 + 222 = 555, но как только она получит возможность обобщать, она вернется к реальной математике, сопротивляясь ручной коррекции.

Оба подхода сталкиваются с ограничениями: согласование сильных ИИ остаётся нерешённой и крайне трудной задачей.


Международное
регулирование ИИ

Принципы ОЭСР по искусственному интеллекту

Приняты в 2019 г.; обновлены в 2024 г.

Первая и наиболее цитируемая международная основа регулирования ИИ, принятая в 2019 году 42 странами — членами ОЭСР и государствами G20. 47 стран и юрисдикций присоединились к принципам.

1. Инклюзивный рост и благополучие людей. ИИ должен приносить общественную пользу и способствовать устойчивому развитию.

2. Уважение к правам человека, верховенству права и демократическим ценностям.

3. Прозрачность и объяснимость систем. Пользователи должны понимать, как и почему система принимает решения.

4. Надёжность, безопасность и устойчивость. Технологии должны быть защищены от ошибок и злоупотреблений.

5. Ответственность разработчиков и организаций.

G7 Кодекс поведения ИИ

Принят в 2023 г.

Принят в рамках Хиросимского процесса по ИИ.

Добровольный документ, задающий стандарт поведения для компаний, создающих «передовые» модели ИИ.

11 пунктов, которые охватывают весь жизненный цикл технологий:

- идентификация, оценка и снижение рисков;
- обеспечение прозрачности и публикация информации о возможностях и ограничениях систем;
- разработка мер реагирования на злоупотребления и инциденты;
- внедрение механизмов марикоровки контента;
- инвестиции в безопасность и киберзащиту.

На основе Кодекса строятся инициативы США, ЕС и Японии.

Блетчли-парковская декларация по безопасности ИИ
(Bletchley Declaration)

Принята 2023 г.

Принята на первом Глобальном саммите по безопасности ИИ (AI Safety Summit) в Великобритании.

Подписали 28 стран, включая США, Китай, ЕС, Японию, Канаду, Индию и Саудовскую Аравию.

Впервые признание рисков, сравнимые с угрозами ядерного оружия. Запуск процесса глобального диалога по ИИ – в Сеуле (2024) и Франции (2025).

Основные положения декларации:

- глобальное сотрудничество по тестированию и оценке ИИ;
- исследовательские институты по безопасности ИИ;
- обеспечение прозрачности, обмена данными и предотвращения злоупотреблений;
- безопасность ИИ – глобальная безопасность.

EU AI Act

Вступил в силу в 2024 г.

Первый в мире комплексный юридически обязательный закон о регулировании ИИ.

Формулируется как мера для защиты прав человека, публичной безопасности и для обеспечения функционирования рынка ЕС. Применение идёт поэтапно.

Опасения: может замедлить инновации, повысить издержки и снизить конкурентоспособность ЕС.

Предусматривает административные санкции при несоблюдении, значительные штрафы для крупных нарушений.

Неприемлемый риск – полностью запрещены (социальный рейтинг, массовая слежка).

Высокий риск – системы, применяемые в здравоохранении, образовании, судопроизводстве и госуслугах; для них предусмотрена сертификация и аудит.

Ограниченный риск – требуются меры прозрачности (маркировка контента, созданного ИИ).

Минимальный риск – большинство бытовых и развлекательных приложений.

Для мощных генеративных систем – особые требования по тестированию, документированию и раскрытию данных обучения

EU AI Act

Запреты:

- Когнитивная манипуляция: намеренное манипулирование поведением человека посредством обмана, убеждения, скрытых воздействий, либо методов, искажающих сознательное принятие решений.
- Социальный рейтинг: использовать ИИ для оценки доверия, поведения или характеристик людей государственными органами;
- Биометрическая идентификация в реальном времени в общественных местах. Исключения - судебное или административное разрешение; временные и территориальные ограничения.
- Биометрическая классификация по политическим взглядам, религии, расе, этническому происхождению, сексуальной ориентации и др. (включая анализ лица, походки, голоса и т. п.).
- Распознавание эмоций на рабочих местах, в образовательных учреждениях;
- Создание баз данных лиц из Интернета, собирать или использовать фотографии из соцсетей.

Резолюция ГА ООН о безопасном и надёжном ИИ
(A/RES/78/265)

Принята в марте 2024 г.

Первая резолюция ООН по ИИ. Принята единогласно 193 странами.

Резолюция подчёркивает:

- необходимость этичного и инклюзивного развития технологий;
- соблюдение прав человека и международного права;
- противодействие дезинформации и предвзятости алгоритмов;
- обеспечение доступа развивающихся стран к технологиям ИИ.

Призывает государства:

- разрабатывать национальные стратегии управления рисками;
- обмениваться знаниями и лучшими практиками;
- поддерживать роль ООН как центральной платформы глобального диалога по ИИ.

Резолюция ГА ООН о создании механизмов по глобальному управлению ИИ (A/RES/79/325)

Принята 26 августа 2025 г.

1. Независимая международная научная группа по ИИ.

40 экспертов в личном качестве с учётом географического баланса.

Подготовка ежегодных аналитических докладов, обобщающих мировые исследования рисков, возможностей и последствий ИИ.

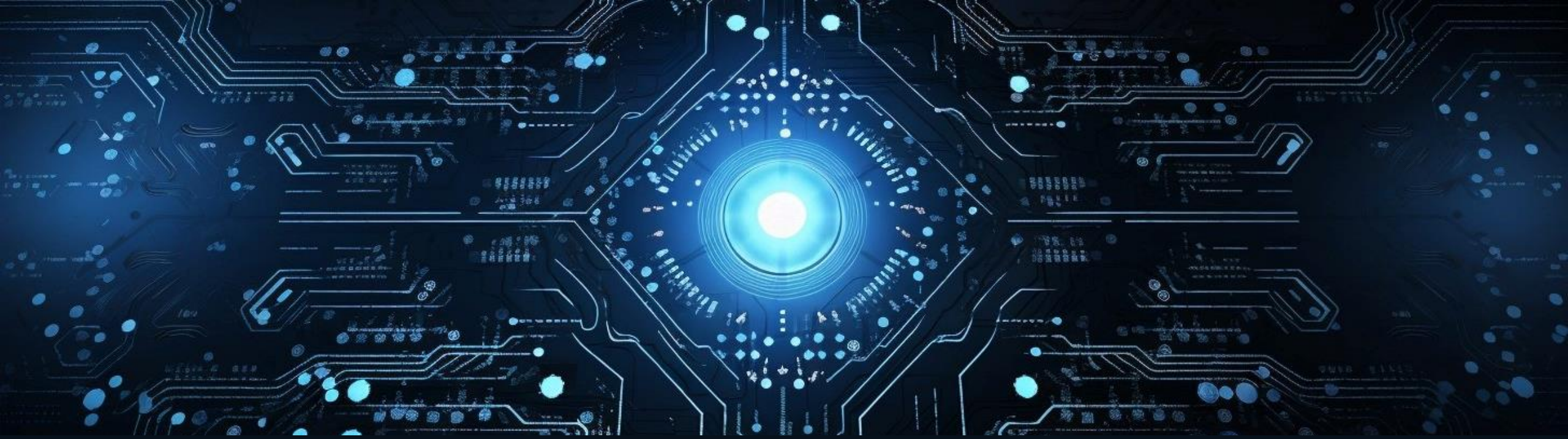
2. Глобальный диалог по управлению ИИ

Многосторонняя площадка для выработки политического консенсуса по регулированию ИИ. Встречи: в рамках 80-й сессии ГА ООН (2025), в рамках Саммита МСЭ «AI for Good» (2026), в ходе Многостороннего форума по науке, технологиям и инновациям (SDGs) 2027.

Формирует глобальное управление ИИ в системе ООН. Переход от рамочных деклараций к созданию единой архитектуры.

Развитие регулирования ИИ в России

Ключевые регуляторные инициативы

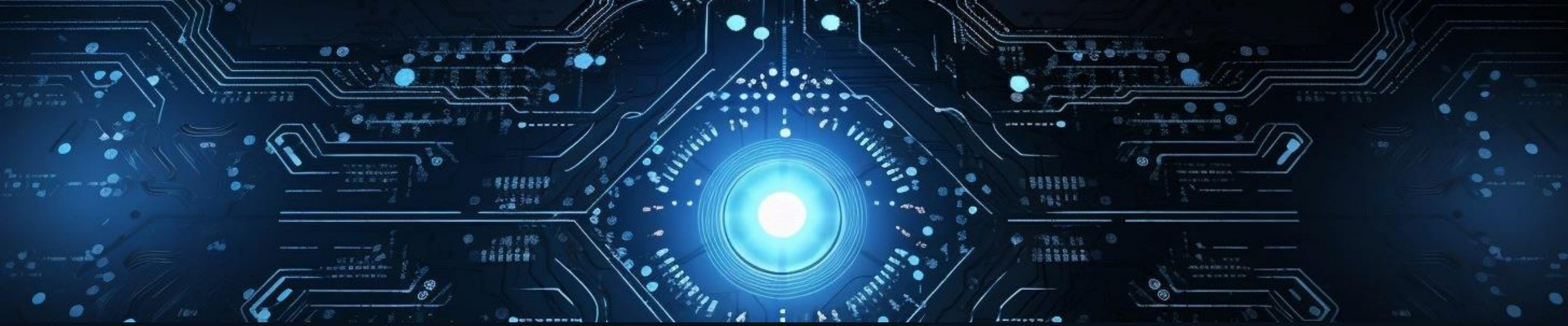


Концепция развития регулирования ИИ до 2030 года

Подготовлена в августе 2025 г.

Определяется комплексный подход к регулированию ИИ.
Основа концепции – человекоориентированный подход, решения в области ИИ должны учитывать интересы человека и общества.

Документ предусматривает создание условий для развития технологий без излишних административных барьеров.



Стратегия развития ИИ

Обновлена в феврале 2025 г.

Основное внимание уделено вызовам, связанным с международными санкциями и необходимостью обеспечения технологической независимости.

Ключевые задачи:
обеспечение конкурентоспособности российских технологий ИИ;
укрепление национальной безопасности и технологического суверенитета.

Акцент на импортозамещении и построении автономной технологической экосистемы, способной развиваться без зависимости от внешних поставщиков.

Проект базового закона об ИИ

Ключевые понятия и механизмы регулирования:

- обязательное обозначение использования ИИ при взаимодействии с пользователем;
- классификация рисков – от минимального до неприемлемого.
- Системы высокого риска подлежат обязательной регистрации, а системы неприемлемого риска подлежат запрету.

Для систем высокого риска предусмотрена обязательная государственная регистрация.

Системы с неприемлемым уровнем риска будут запрещены к использованию.

Уголовная ответственность за преступления с использованием ИИ. Штрафы до двух млн. руб. и лишение свободы на срок до пятнадцати лет.

Экспериментальные режимы

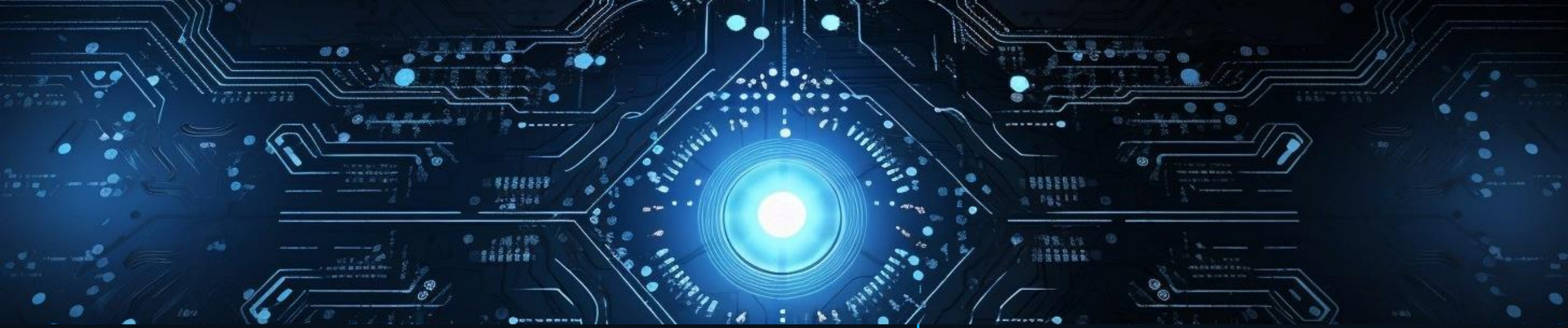
В 2025 году действует порядка четырнадцати режимов в различных областях применения ИИ. Охватывают беспилотный транспорт, медицину, финансы и другие сферы.

Военно-ориентированные ЭПР:

- Беспилотные авиационные системы в Москве
- Аэрологистика – для беспилотной перевозки грузов
- Тестирование военных роботов-охранников

Позволяют тестировать новые технологии в специальных правовых условиях.
Дает возможность изучить практические аспекты применения ИИ до принятия окончательных законодательных решений.

Этические стандарты ИИ



Кодекс этики в сфере ИИ

Принят в 2021 г.

900+ подписантов

33 международных организации из 22 стран

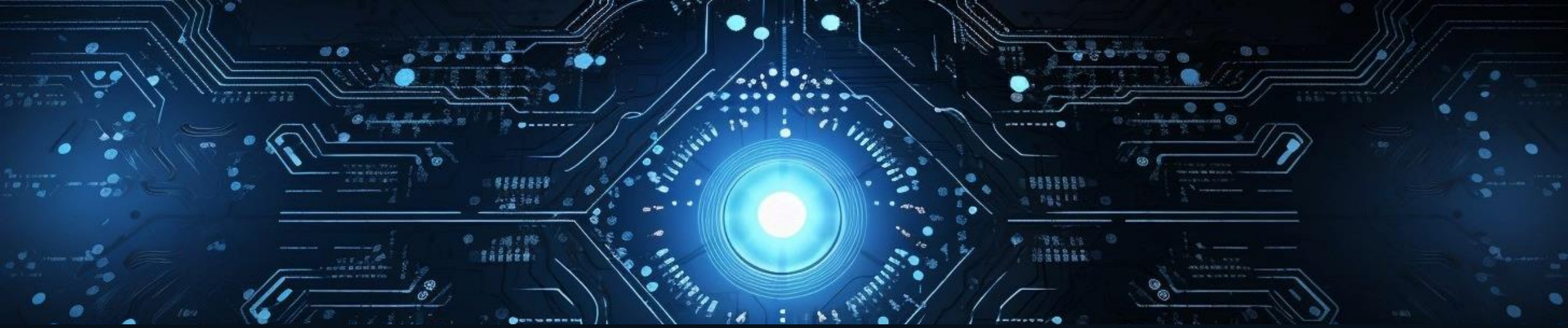
Положения Кодекса :

- приоритет человекоцентричного подхода;
- требование прозрачности и объяснимости алгоритмов;
- недопустимость дискриминации и вреда;
- акцент на безопасности и ответственности разработчиков и поставщиков.

Отраслевые этические стандарты:

- Кодекс этики в сфере ИИ в медицине и здравоохранении
- Кодекс этики в сфере ИИ на финансовом рынке

Россия внесла значительный вклад в подготовку кодекса этики ИИ ЮНЕСКО 2021 г. (подписан 193 странами)



Белая книга в сфере ИИ

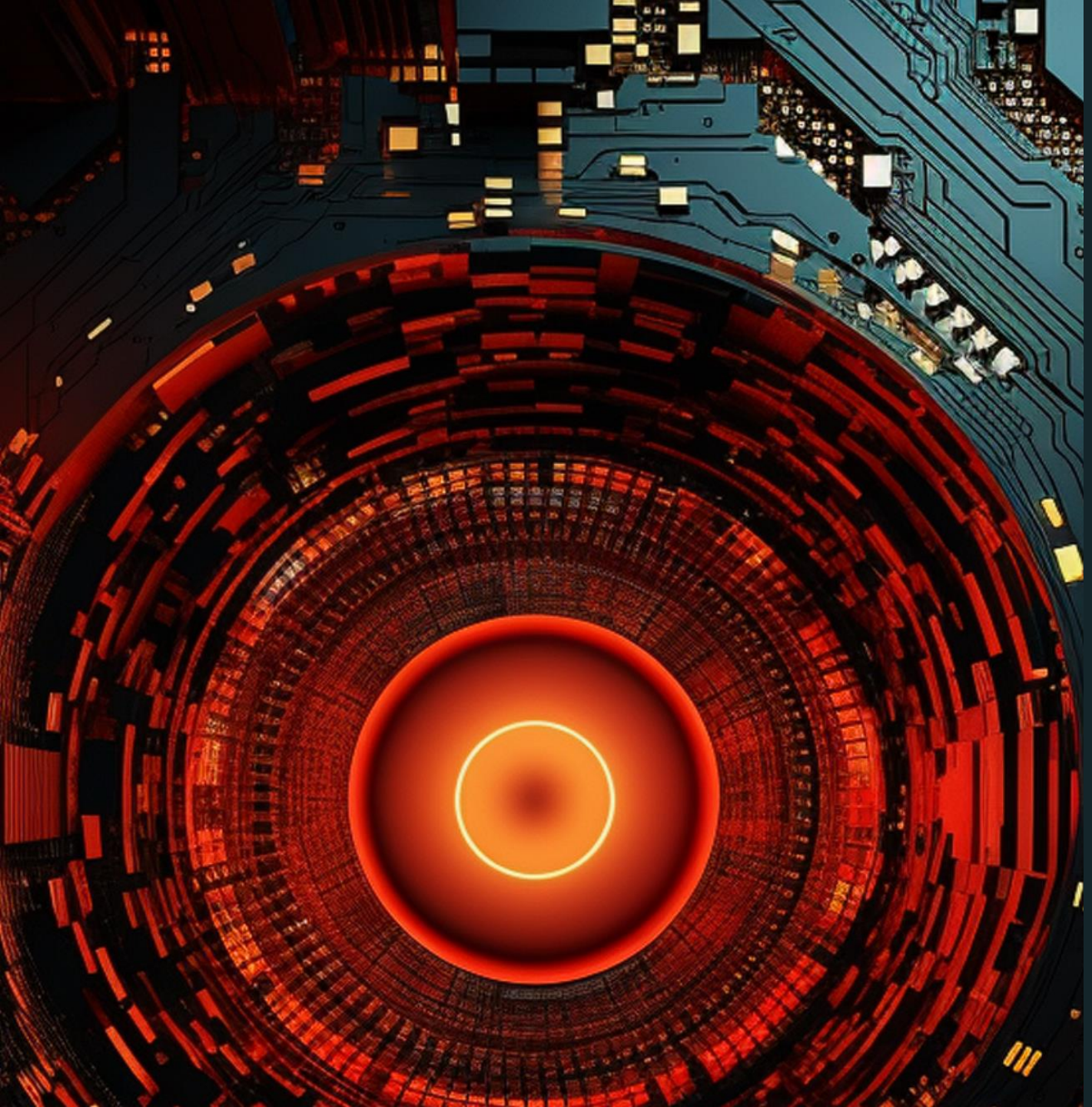
Структурированное исследование ключевых этических проблем, связанных с использованием ИИ.

Документ основывается на международных стандартах (OECD, UNESCO, EU AI Ethics Guidelines, IEEE, ISO) и на результатах российских научных и экспертных дискуссий.

Ключевая задача – подготовить единую систему для оценки рисков и принятия решений в области этики ИИ.

Ключевая идея – этика не должна быть препятствием развитию технологий, а должна обеспечивать устойчивое, безопасное и ориентированное на человека внедрение ИИ.

Рекомендации практичны, не декларативны.



ИИ в военной сфере

GGE по LAWS (CCW)

Комиссии ООН по разоружению

МГП

Женевские конвенции, Дополнительные протокол I (1977)

REAIM

Первый саммит прошёл в Гааге в 2023 году, второй – в Сеуле в 2024 году.

Дипломатическая площадка по военному ИИ.

Цель – обеспечить ответственное и безопасное использование ИИ в военной сфере.

Ключевые принципы REAIM:

- Meaningful human control – человек должен осуществлять контроль принятия решений.
- Соблюдение международного гуманитарного права.
- Прозрачность, ответственность и надёжность технологий.
- Международное сотрудничество по созданию стандартов и обмену опытом.

Резолюция ООН по САС (A/RES/78/241)

Принята в декабре 2023 г.

Первая резолюция Генассамблеи по LAWS. Недопустимость автономного выбора и поражения целей алгоритмами без контроля человека. Уделяется внимание правам человека и МГП.

Разработка международных норм: правовой договор, кодексы поведения, принципы обеспечения контроля и подотчётности.

Поддержали – 152
Против – 4
Воздержались – 11

Россия: Против фрагментации регулирования ИИ в военной сфере. Регулирование должно вестись в рамках действующих многосторонних форматов - GGE по LAWS (CCW) и Комиссии ООН по разоружению.

Против односторонне негативного подхода, САС могут повышать точность и снижать вероятность ошибок, поэтому дискуссия должна быть сбалансированной.

МГП полностью распространяется на САС и не требует модернизации;

Попытки вводить новые определения и правила — преждевременны.

Резолюция ООН по ИИ в военной области (A/RES/79/239)

Принята в декабре 2024 г.

Применимость норм международного права, включая Устав ООН, международное гуманитарное право, нормы в области прав человека к использованию ИИ в военной сфере.

Поддержали – 165
Против – 2
Воздержались – 6

Россия: Против фрагментации регулирования ИИ в военной сфере. Регулирование должно вестись в рамках действующих многосторонних форматов – GGE по LAWS (CCW) и Комиссии ООН по разоружению.

Критика «человекоцентричности» – критерии не основаны на МГП и формируют дискуссию до согласования самого определения ИИ.

Против продвижения узких неинклюзивных инициатив (REAIM) – такие форматы навязывают взгляды ограниченной группы государств как «универсальные».

Роль государств первична, разработка норм международного права в области военного ИИ – компетенция государств; участие иных акторов (НПО, бизнес, экспертов) – вспомогательное.

Россия: регулирование военного ИИ

- Контроль человека: сохранение человеческого контроля над решениями о применении оружия.
- Государственный суверенитет: развитие отечественных технологий военного ИИ в условиях санкционных ограничений на доступ к зарубежным технологиям.
- Россия не поддерживает искусственные запреты на разработку и применение САС.

AI Is Already at War

Michèle Flournoy

- Разведывательное сообщество США и ряд боевых командований США используют ИИ для анализа огромных массивов засекреченных и открытых данных, что позволяет прогнозировать развитие международных событий.
- В Стратегическом командовании США используется ИИ для оповещения о перемещениях ядерных ракет.
- Применение ИИ-систем для анализа разведданных и открытых источников сыграло ключевую роль в прогнозировании ввода российских войск в Украину. По ее словам, эти системы позволили американским аналитикам за несколько месяцев предсказать развитие событий.
- Системы анализа на основе ИИ могут обеспечить Вашингтон более глубоким пониманием того, о чём могут думать потенциальные противники в Пекине. Например, разработать LLM, которая обрабатывает все доступные выступления и тексты китайских лидеров, разведывательные отчёты США, на основе чего моделирует то, как председатель КНР Си Цзиньпин планирует реализовать ту или иную заявленную политику.
- LLM может просчитать, как будет развиваться кризис и каким образом различные решения повлияли бы на итоговый сценарий.
- Страна, которая первой осуществит комплексное внедрение ИИ в военную сферу и силовые операции, институционализирует его применение, получит стратегическое преимущество в формировании будущего мироустройства.

Риски ИИ в военной сфере

- **Ошибки восприятия и анализа данных**

Неправильное распознавание цели (гражданский объект как военный).

Галлюцинации моделей, ошибки из-за искаженных данных/узких нарративов.

- **Уязвимость к манипуляциям и атакам на системы**

AML-атаки (изменения входных данных), спуфинг сенсоров (ИИ видит несуществующие цели/пропускает реальные), data poisoning (обучение искаженными данными), подмена модели (обновление, меняющее логику принятия решений).

- **Некорректное поведение системы**

Reward-hacking, рассогласование действий агентов.

- **Непрозрачность и потеря контроля**

Отсутствие интерпретируемости решений. Ограниченная возможность вмешательства оператора в условиях высокой скорости.

- **Стратегические и социально-политические риски**

Дезинформация и дипфейки (подрыв доверия и стабильности).

Когнитивные воздействия, психологические войны.

AIxBio: снижение порога доступа, разработка патогенов и токсичных соединений, создание биооружия.

Размывание ответственности за применение силы (оператор, модель, разработчик, государство)

Почему «Человек в цикле» не гарантирует контроль

Keeping humans in the loop is not enough to make AI safe for nuclear weapons
Peter Rautenbach

- **Сокращение времени на реакцию**: алгоритмы анализируют данные быстрее, чем человек успевает осмыслить ситуацию, в результате оператор фактически не успевает вмешаться.
- **Эффект доверия автономным системам:** в условиях стресса и неопределённости люди склонны полагаться на точность ИИ, даже если модель может ошибаться.
- **Непрозрачность алгоритмов:** современные модели не объясняют ход рассуждений, в результате чего трудно оценить обоснованность автоматических выводов.
- **Уязвимость к внешним манипуляциям:** спуфинг сенсоров, data poisoning, AML-атаки могут привести к ложным тревогам или ошибочным целеуказаниям.
- **Риск автоматической эскалации:** автономные системы обеих сторон могут непреднамеренно усиливать тревожные сигналы, ускоряя переход к кризису.

Даже при формальном участии человека возрастает вероятность непреднамеренного военного столкновения, включая угрозу применения ядерного оружия.

ИИ и стратегическая стабильность

Задачи обеспечения безопасности ИИ в военной сфере:

- **Повышение объяснимости и контролируемости моделей:** возможность проверять, как и на каких данных ИИ формирует выводы.
- **Обеспечение устойчивости и надежности:** системы должны устойчиво работать вне лабораторных условий и быть защищены от подмены данных и кибервмешательств.
- **Ограничение автономии в критически важных решениях:** в системах раннего предупреждения, системах ядерного управления.
- **Право человека на вмешательство:** технические и процедурные механизмы, пауза принятия решения, отмена команды, независимый контроль.
- **Международные меры укрепления доверия:** консультации, горячие линии, обмен методами проверки надежности, совместные учения по предотвращению ложных срабатываний.