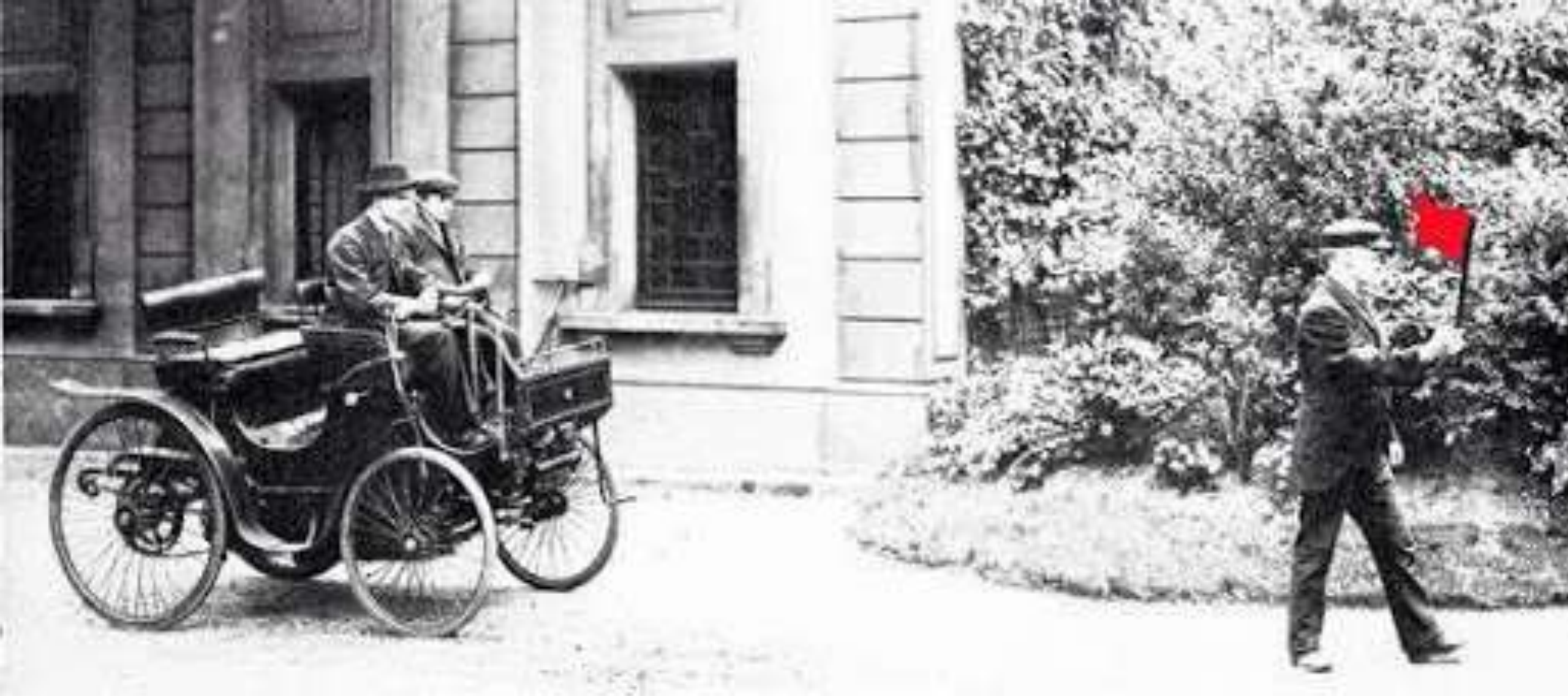
SBER AI

Подходы к регулированию искусственного интеллекта

Незнамов Андрей Владимирович
Председатель российской Комиссии по реализации Кодекса этики в сфере ИИ
Управляющий директор Центра регулирования ИИ, Сбер



Регулирование новых технологий на примере дорожного движения

- Затормозить развитие *(Англия)*
- Не мешать или способствовать развитию *(Россия, Германия)*
- Отставать от развития *(США)*
- Опережать развитие *(Франция, Англия)*

Общепризнанные правила для новых технологий всегда отстают от реальности

Самолеты:

Чикагская конвенция о международной гражданской авиации, 1944 г.

Организация Международной гражданской авиации, 1947 г.

Первый полет – в 1903 г.

Ядерное оружие:

Договор о нераспространении ядерного оружия, 1968 г.

Международное агентство по атомной энергии, 1957 г.

Карибский кризис – в 1962 г.

Клонирование:

Азиломарская конференция о создании рекомбинантных ДНК, 1975 г.

Директива 90/219 ЕС «Об Ограниченном использовании генетически изменённых микроорганизмов», 1990 г.

Декларация ООН о клонировании человека, 2005 г.

Искусственный интеллект, робототехника и другие сквозные цифровые технологии (?)

Как будет регулироваться искусственный интеллект?

Зарождение регулирования

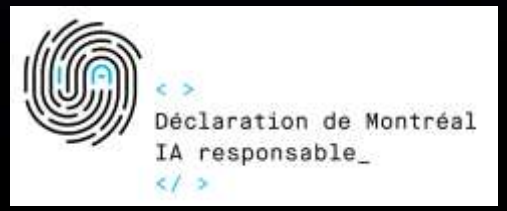
От «Робоэтики» → К «Этике ИИ»

- Лига Гуманности: роботы не должны подвергаться бесчеловечному обращению (R.U.R. Карела Чапека, 1920)
- Четыре закона робототехники (А. Азимов, 1942)
- Robot Ethics Charter (Южная Корея, 2007)
- Европейская хартия робототехники (Европарламент, 2017)

- 10 Законов для искусственного интеллекта (Microsoft, 2016)
- Азиломарские принципы ИИ (Future of life Institute, 2017)
- Глобальная инициатива по этике автономных и интеллектуальных систем (IEEE, 2017)
- Декларация о сотрудничестве в сфере ИИ (страны ЕС, 2018)
- Принципы ИИ (ОЭСР, 2019)
- Рекомендация об этических аспектах искусственного интеллекта (ЮНЕСКО, 2021)

Предложения НКО и частных групп исследователей

OpenAI



AINOW

- 23 Asilomar AI principles
- Montréal Declaration: Responsible AI
- The AI Now Report. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term
- OpenAI Charter
- The Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning System
- Группа проф. Susumu Hirano, Япония: 8 принципов
- Модельная конвенция робототехники и ИИ (Россия)

Азиломарские принципы ИИ

Отказ от гонки

Команды, разрабатывающие системы ИИ, должны активно сотрудничать между собой и не пытаться победить за счет игнорирования стандартов безопасности

Совместное процветание

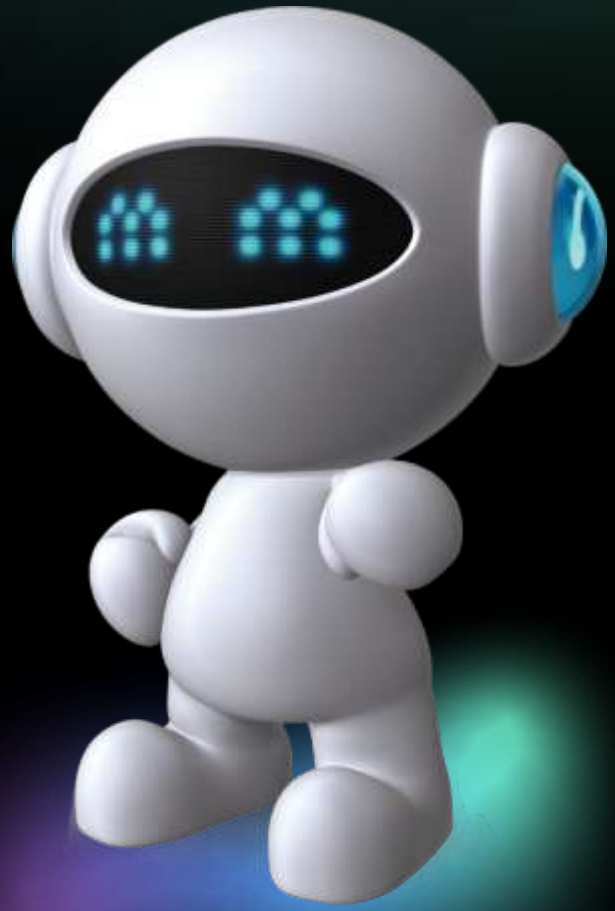
Экономическое процветание, достигнутое благодаря ИИ, должно широко использоваться в интересах всего человечества

Человеческие ценности

Системы ИИ должны разрабатываться и работать таким образом, чтобы быть совместимыми с идеалами человеческого достоинства, его прав и свобод, многообразия культур

Общее благо

Суперинтеллект должен разрабатываться только для служения широко разделяемым этическим идеалам и на благо всего человечества, а не одного государства или организации



Корпоративные нормы

«10 Законов для искусственного интеллекта», Microsoft

«7 принципов ИИ», Google

«6 тематических областей исследования этики ИИ», Deepmind

«5 принципов ИИ», Telefonica

«Руководство по ИИ», Deutsche Telekom

«Руководящие принципы ИИ», SAP

«Этические принципы», Японская ассоциация ИИ

«Руководящие принципы этики ИИ», Sony Group

«Руководящие принципы ИИ», Unity

«Инициатива в сфере ИИ», Partnership on AI

Национальные инициативы

Примеры:


Robot Ethics Charter *(Южная Корея)*


A guide to using artificial intelligence in the public sector *(Великобритания)*


Understanding artificial intelligence ethics and safety (*Великобритания)*


AI Ethics Framework *(Австралия)*


Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government *(США)*


Ethical Principles for Artificial Intelligence *(США)*


Principles of Artificial Intelligence Ethics for the Intelligence Community *(США)*


Civil Law Rules on Robotics *(European Parliament)*


Responsible use of AI *(Канада)*


Social principles of Human-centric AI *(Япония)*


Artificial intelligence at the service of citizens *(Италия)*

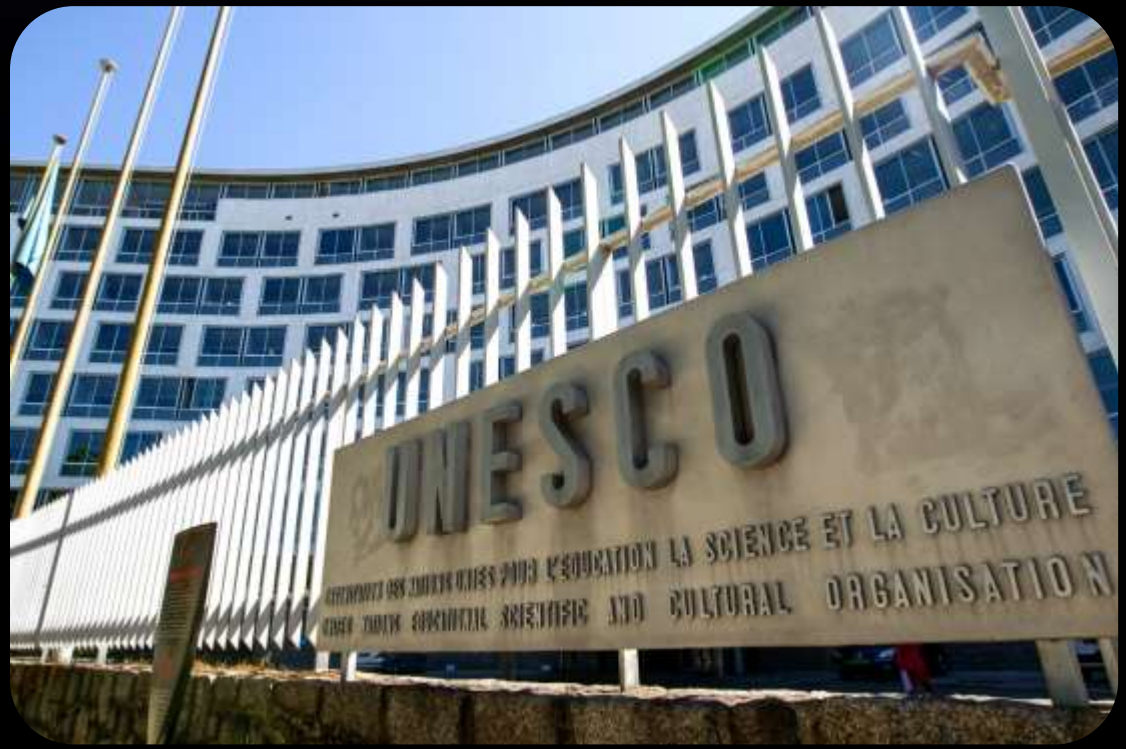

AI ethics and governance body of knowledge *(Сингапур)*


Ethics guidelines for intelligent artificial society *(Южная Корея)*


Кодекс этики ИИ *(Россия)*

Наднациональные инициативы

Рекомендации по этике ИИ ЮНЕСКО



- Ethically Aligned Design (IEEE)
- Top 10 Principles for Ethical Artificial Intelligence (UNI Global Union)
- Ethics Guidelines for Trustworthy AI (Совет Европы)
- European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment (Совет Европы)
- Report on Robotics Ethics (COMEST ООН)
- Principles on AI (ОЭСР)

Международные инициативы по гражданскому ИИ

Глобальное партнерство по ИИ

- Цель – формировать повестку по «ответственному» ИИ
- Включает 29 членов, в том числе

Соглашение США и ЕС о сотрудничестве в сфере ИИ

Расширение совместного использования ИИ в:

- сельском хозяйстве
- здравоохранении
- борьбе с чрезвычайными ситуациями
- прогнозировании климата
- электроэнергетике

Декларация ЕС о сотрудничестве в сфере ИИ

- Повышение технологического и производственного потенциала Европы в области ИИ и его применения

Декларация в области НИОКР в сфере ИИ между США и Великобританией

- Совместная работа над экосистемой НИОКР в сфере ИИ
- Фокус на сложных технических вопросах и защите от попыток применения ИИ на службе авторитаризма и репрессий

Программа сотрудничества государств-членов ШОС по развитию ИИ

- Проведение исследований
- Применение ИИ для обеспечения роста благосостояния населения, стимулирования экономического развития обеспечения региональной и национальной безопасности и охраны правопорядка

Декларация стран Северо-Балтийского региона о сотрудничестве в сфере ИИ

- Расширение доступа к данным для ИИ
- Разработка этических и прозрачных руководств, стандартов, принципов и ценностей

Международные инициативы по ИИ для безопасности

Стратегия НАТО в сфере ИИ

- Устанавливает 6 принципов ответственного использования ИИ в обороне:

1. Законность (соответствие национальному и международному праву)
2. Ответственность и подотчетность
3. Объяснимость и возможность отслеживания (обеспечивается через механизмы проверки, оценки и валидации на уровне НАТО и/или национальном уровне)
4. Надежность (обеспечивается через процедуры сертификации на уровне НАТО и/или национальном уровне
5. Управляемость (соответствие изначально заложенным в модель функциям)
6. Минимизация последствий непреднамеренной предвзятости в моделях ИИ

- В феврале 2023 НАТО анонсировал начало работы над стандартом сертификации ИИ

- ИИ и автономные системы определены одним из одними из основных направлений сотрудничества США, Великобритании и Австралии в рамках трехстороннего соглашения о безопасности AUKUS

- Четырёхсторонний диалог по вопросам безопасности (QUAD) договорился использовать машинное обучение и связанные с ним технологии для повышения уровня кибербезопасности в феврале 2023 г.

- > 60 стран, включая США и КНР, в феврале 2023 г. подписали совместное заявле по итогам 1-го международного Саммита по ответственному использованию ИИ в военных целях
- Подписанты выразили свою приверженность «международным юридическим обязательствам» при использовании военного ИИ

Инициативы США по ИИ для безопасности

Стратегия МО США по развитию ИИ 2018 (US Department of Defense AI Strategy)

- Заявляет ориентированную на человека модель использования ИИ, декларируя заботу о личном составе и гражданских лицах
- Одним из «стратегических подходов» решение этических проблем применения технологий ИИ в военной сфере и вопросов их безопасности.

Этические принципы МО для использования ИИ 2020 (US Department of Defense AI Ethical Principles)

- Зафиксированы 5 основных принципов: ответственность, беспристрастность, прослеживаемость, надежность, управляемость

Ответственная стратегия ИИ МО США и путь реализации 2022 (US Department of Defense Responsible AI Strategy & Implementation Pathway)

- Представляет собой «дорожную карту» продвижения МО ответственного использования ИИ
- Содержит 6 принципов, в том числе ответственное управление ИИ, ответственная экосистема ИИ

Политическая декларация об ответственном военном использовании ИИ 2023 (Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy)

- Представлена в ходе 1го Саммита по ответственному использованию ИИ в военных целях в феврале 2023 г.
- Зафиксированы 12 лучших практик для развития военных возможностей ИИ с акцентом на ответственность человека

Объединенный центр технологий ИИ (Joint AI Center, JAIC)

- Обеспечивает условия для использования ИИ в МО, в том числе координирует внедрение принципов этического использования ИИ в вооружённых силах

Инициативы Великобритании по ИИ для безопасности

Оборонная стратегия Великобритании в сфере ИИ 2022 (Defence AI Strategy)

- Основана на 3-х точках опоры: амбиции, инновации, ответственность
- Призывает внедрить ИИ во все возможные зоны ответственности оборонного ведомства
- Устанавливает новый современный подход к взаимодействию с частным сектором

Политика «Амбициозный, безопасный, ответственный: наш подход к предоставлению ИИ-возможностей в обороне (Policy paper “Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence”)

- Разработана в дополнение к Стратегии в партнерстве с Центром по этике и инновациям данных (CDEI)
- Включает этические принципы использования ИИ в обороне

Центр оборонного ИИ (Defence AI Centre)

- Выступает в качестве визионерского центра, который объединяет опыт и возможности от ряда кросс-функциональных команд для поддержки разработки, реализации и принятия ИИ в сфере обороны

Инициативы Китая по ИИ для безопасности

Меморандум о регулировании применения ИИ в военной сфере 2021 (Position Paper of the People’s Republic of China on Regulating Military Applications of AI)

- Документ представлен Китаем в рамках 6-й обзорной конференции Конвенции ООН по обычным вооружениям
- Одной из особых проблем отмечены долгосрочное воздействие и потенциальные риски военного применения технологий ИИ в части стратегической безопасности, регулирования и этики

Меморандум об укреплении этичного управления ИИ 2022 (Position Paper of the People's Republic of China on Strengthening Ethical Governance of AI)

- Документ представлен на встрече высокого уровня Конвенции ООН о конкретных видах обычных вооружений
- Призывает международное сообщество договориться о вопросе этики ИИ на основе участия широкого круга заинтересованных сторон и сформулировать основы международной структуры управления ИИ, включая стандарты и нормы для ИИ
- Правительства должны учитывать наихудшие сценарии и повысить осведомленность о рисках использования ИИ, определить потенциальные этические риски, которые может повлечь за собой использование технологий ИИ, создать эффективный механизм раннего предупреждения таких рисков

Белая книга по ИИ в безопасности 2018 (Artificial Intelligence Security White Paper)

- Опубликована Китайской академией ИКТ (CAICT)
- Призывает китайское правительство избегать гонки ИИ-вооружений между странами

Опыт России в регулировании ИИ

2 место в мире Россия заняла в рейтинге Stanford AI Index по количеству разработанных законов и законопроектам по ИИ за последние 5 лет

7 направлений регулирования предусмотрено Национальной стратегией развития ИИ до 2030 года:

- доступ к данным (в т.ч. собираемым госорганами и медорганизациями)
- доступ к данным для научных исследований и создания технологий ИИ
- упрощенное тестирование и внедрение ИИ и делегирование системам ИИ возможности принятия отдельных решений
- устранение барьеров при экспорте ИИ-продукции гражданского назначения
- создание единых систем стандартизации и оценки соответствия систем ИИ
- стимулирование привлечения инвестиций (механизмы ГЧП)
- разработка этических правил взаимодействия человека с ИИ

Утверждены ключевые стратегические документы

- Национальная стратегия развития ИИ до 2030 года
- Концепция развития регулирования ИИ и робототехники до 2024 года
- Кодекс этики в сфере ИИ

Развиваются отраслевые инициативы регулирования ИИ в сферах:

- данные
- медицина
- ГЧП
- экспортный контроль,
- беспилотный транспорт
- интеллектуальная собственность

Действует законодательство об экспериментальных правовых режимах

Создана одна из самых продвинутых систем этики ИИ

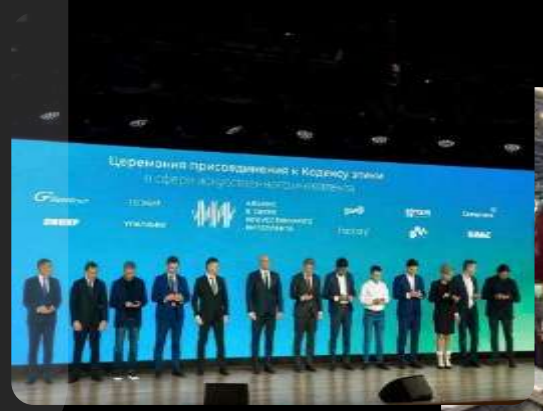
Принят национальный Кодекса этики ИИ

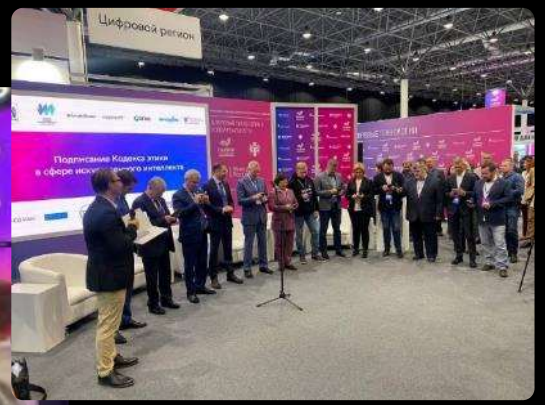
>160

Компаний присоединилось

Создана Национальная Комиссия по реализации Кодекса

- Государственное регулирование уравновешивается инструментами мягкого права
- Носит рекомендательный характер
- Присоединение осуществляется на добровольной основе
- Распространяется только на гражданские разработки





Как Кодекс реализуется?

- Всем подписантам рекомендуется назначить Уполномоченного по этике, ответственного за реализацию Кодекса
- Подписанты избирают состав Комиссии по этике в сфере ИИ, её Председателя и участвуют в ее работе
- Подписанты могут создавать публичный свод наилучших и/или наихудших практик решений возникающих этических вопросов в жизненном цикле ИИ
- Рабочие группы разрабатывают методики и руководства, обеспечивающие соблюдение положений Кодекса
- На сайте Комиссии по этике ИИ ведется публичный Реестр Акторов ИИ

Где прочитать полный текст Кодекса и как присоединиться?

Текст Кодекса доступен на сайте Альянса в сфере ИИ: https://a-ai.ru/code-of-ethics/