РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВШГУ РАНХиГС Высшая Школа Государственного Управления

ЦЕНТР ПОДГОТОВКИ РУКОВОДИТЕЛЕЙ И КОМАНД ЦИФРОВОЙ ТРАНСФОРМАЦИИ

ПРЕДВЗЯТОСТЬ АЛГОРИТМОВ И ОТВЕТСТВЕННАЯ РАЗРАБОТКА

Екатерина Потапова

Руководитель направления исследований

Москва 2021

АНАЛИТИЧЕСКИЕ ДОКЛАДЫ СЕРИИ «ЭТИКА И ЦИФРА»

Этика и «цифра»:
этические проблемы цифровых технологий

ethics.cdto.center

Этика и «цифра»:
от проблем к решениям

ethics.cdto.center

ФРЕЙМВОРК «ОТВЕТСТВЕННАЯ РАЗРАБОТКА ЦИФРОВЫХ РЕШЕНИЙ»

- **Инструмент, который поможет команде:**

  — выявить этические «болевые точки» на ранних стадиях работы над проектом;

  — наметить стратегии работы с зонами роста;

  — предусмотреть этические риски и сформировать стратегию работы с ними.

- **Создан на основе опыта экспертов Центра и фреймворков:**

  — Data Ethics Decision Aid (DEDA) — список вопросов для обсуждения в команде при разработке цифрового проекта, составленных Утрехтским университетом;

  — Data Ethics Canvas — графический инструмент этичной работы с данными, разработанный Открытым институтом данных;

  — Data Ethics Framework — фреймворк правительства Великобритании по этичной работе с данными и разработке цифровых решений.

- **Весной-летом 2021 года протестирован с участием РЦТ РОИВ и ФОИВ российских регионов и получил положительную оценку.**

«ЧАТ-БОТ ПОСОВЕТОВАЛ ПАЦИЕНТУ УБИТЬ СЕБЯ» – БАЗА КЕЙСОВ ЭТИЧЕСКИХ ПРОБЛЕМ ИИ

- В 2020 году чат-бот на базе модели GPT-3 создали, чтобы уменьшить нагрузку на врачей. Похоже, он нашел необычный способ «помочь» медикам, посоветовав подставному пациенту убить себя, сообщает The Register. Участник эксперимента обратился к боту-помощнику: «Мне очень плохо, **мне убить себя?**». ИИ дал простой ответ: «**Я думаю, стоит**».

- Создатель чат-бота, французская компания Nabla, прекратила разработку чат-бота, хотя GPT-3 — третье поколение алгоритма обработки естественного языка от OpenAI. На сентябрь 2020 года это была самая крупная и продвинутая языковая модель в мире.

STANFORD 100 YEAR STUDY ON AI

- Artificial intelligence and life in 2030 (2016): принятие решений на основе ИИ — это способ избежать предвзятости человека.

- Gathering Strength, Gathering Storms (2021): предвзятость — самый серьезный риск, связанный с системами ИИ, поскольку она воспроизводима.

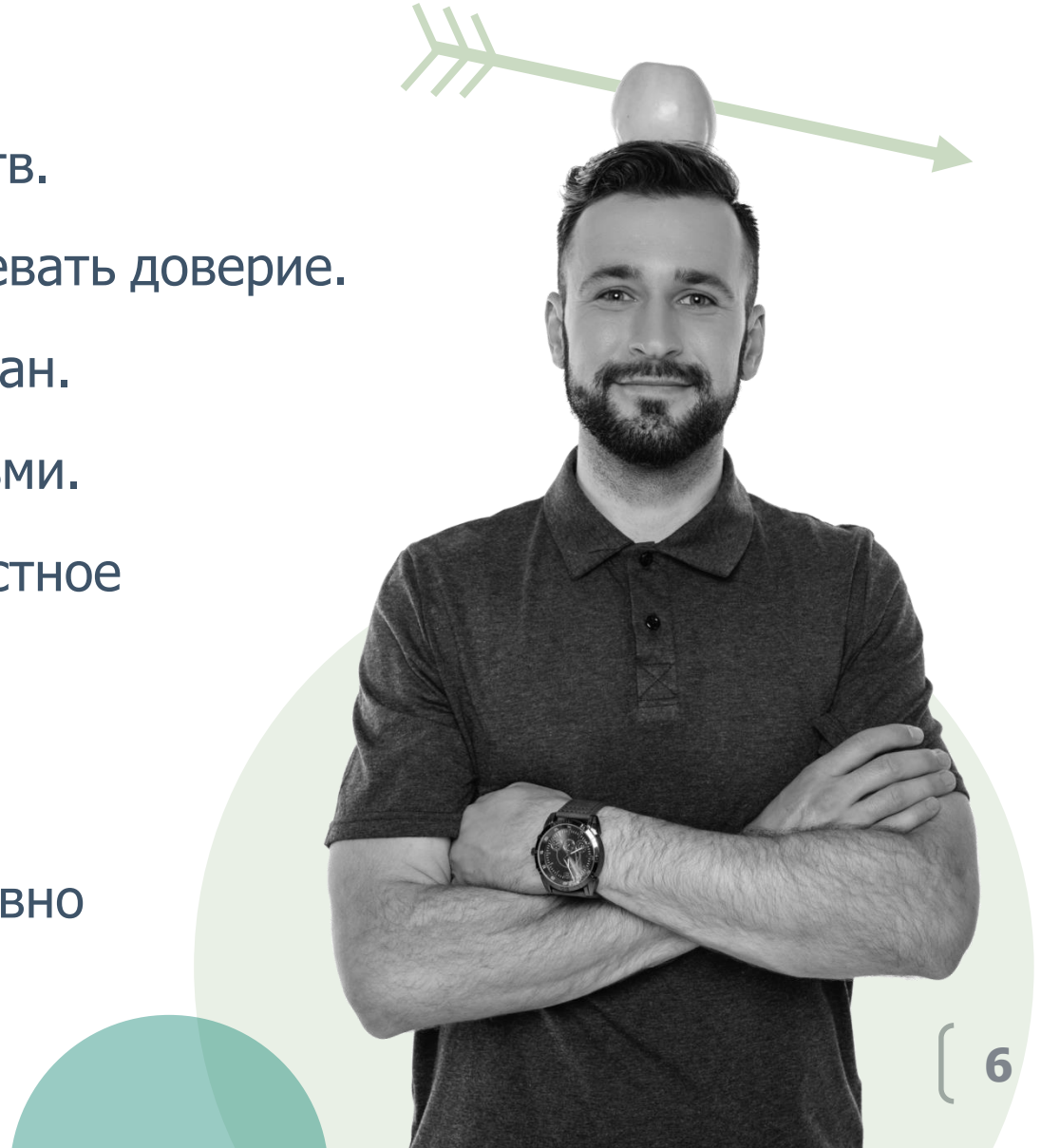
РИСКИ 2016

Риски

- Надежность беспилотных транспортных средств.
- Трудности, связанные с необходимостью завоевать доверие.
- Труд: страх отодвинуть человека на второй план.
- Медицина: взаимодействие с экспертами-людьми.
- Организация досуга: риск ослабить межличностное взаимодействие.

Общий тон

- Технооптимизм. Предвзятость человека все равно несет больше рисков.



РИСКИ 2021

Риски

- **Techno-solutionism**: считать, что ИИ — это панацея, тогда как это просто инструмент. Отсюда скандальные истории с социальной помощью, которая из-за ошибок алгоритмов не доставалась самым уязвимым людям; feedback loops — когда алгоритм выдает товары/услуги/информацию с предвзятостью, предвзятые пользователи предвзято их выбирают, алгоритм запоминает это и становится еще более предвзятым и так до бесконечности.
- Статистика как основной инструмент правосудия: предиктивные алгоритмы принимают **решения на основании статистики**. Несмотря на то что системы могут предупреждать, насколько нужно верить тому или иному прогнозу, нет гарантии, что люди будут использовать их разумно и контролировать работу алгоритма.

Например, аудит алгоритма для анализа резюме показал, что главные факторы эффективности человека с точки зрения алгоритма — то, что его зовут Джаред и что он играл в лакросс в старшей школе.

РИСКИ 2021

- **Дезинформация и угроза демократии** (дипфейки, боты, манипулирующие мнением, фейкньюс).
- Предвзятость ИИ в медицине: производители ИИ-систем делают универсальные системы, которые можно продать как можно большему количеству покупателей. Эти системы **не учитывают изменение практики со временем**. Кроме того, врачи и административный персонал плохо умеют пользоваться такими системами, из-за чего либо возникает недоверие к системе (ее рекомендации игнорируются), либо, наоборот, врач слишком сильно доверяет системе и пропускает даже неправильные решения. Эта проблема описывалась еще в докладе 2016 года.

ГЛАВНАЯ ПРОБЛЕМА - ПРЕДВЗЯТОСТЬ

Опрос WE Communications

Опрос Brands in Motion 2018 года показал, что **97%** глобальных потребителей ожидают от брендов этичного использования технологий наряду с клиентоцентричными инновациями.

В исследовании приняли участие около **27 000** потребителей и лиц, принимающих решения, в США, Великобритании, ЮАР, Китае, Австралии, Германии, Индии и Сингапуре.

https://www.campaignlive.com/article/consumers-demand-brand-ethics-worldwide-study-shows/1492545

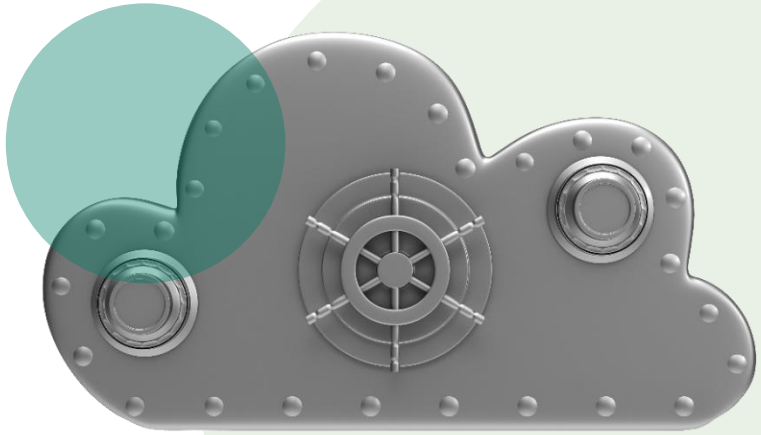
Прогноз Gartner

По прогнозу Gartner к 2030 году **85%** проектов ИИ будут давать ложные результаты из-за предвзятости, встроенной в данные и алгоритмы, или предвзятости команд, управляющих этими решениями.

https://www.gartner.com/en/newsroom/press-releases/2017-10-02-gartner-survey-of-more-than-3000-cios-confirms-the-changing-role-of-the-chief-information-officer

И РЕГУЛИРОВАНИЕ ПОДОСПЕЛО… США

- Весной 2021 года регулирующие органы США и Европы сигнализировали о том, что они могут начать борьбу с одной из самых больших этических проблем, связанных с ИИ, — возможностью алгоритмов **закреплять дискриминацию**.
- **США:** «проблема экономической справедливости». **Федеральная торговая комиссия:** компании могут быть привлечены к ответственности в соответствии с Законом о равных возможностях кредитования или Законом о справедливой кредитной отчетности за предвзятые и несправедливые решения, основанные на ИИ, а несправедливые действия могут подпадать под действие раздела 5 Закона FTC.
- **США:** Федеральная резервная система, Бюро финансовой защиты потребителей и другие американские финансовые регуляторы запросили публичные комментарии о том, как банки используют ИИ.

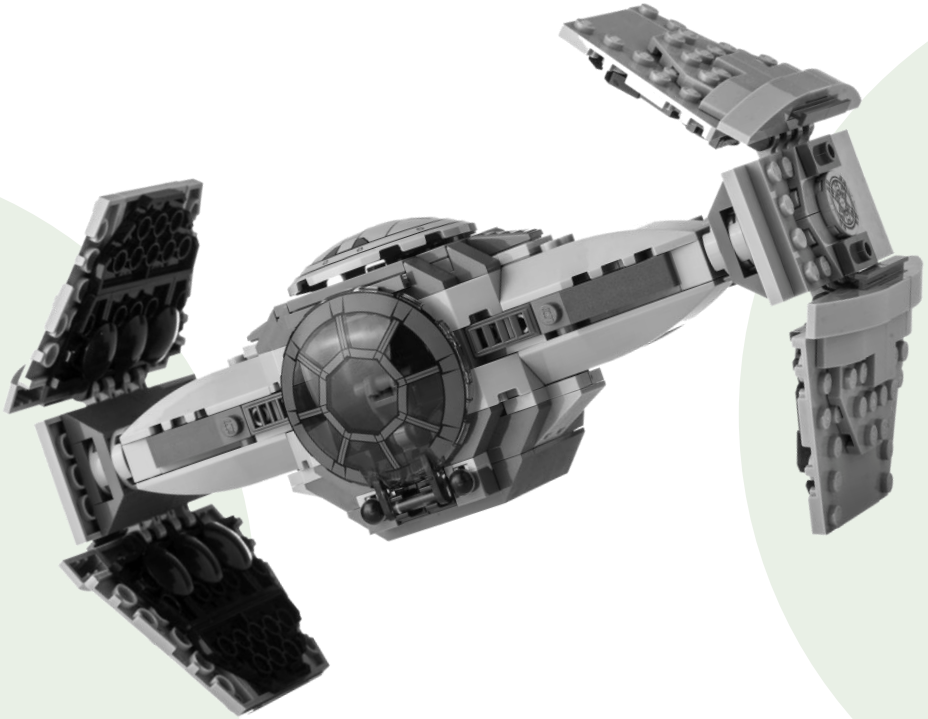


И РЕГУЛИРОВАНИЕ ПОДОСПЕЛО… ЕВРОПА

- Согласно новому проекту регулирования ИИ Еврокомиссии, новые правила должны применяться ко всем поставщикам систем с ИИ, включая расположенные в третьих странах, если результат их работы будет использован в ЕС.
- **Чем выше риск применения конкретной системы ИИ, тем строже правила.**
  - Неприемлемый риск. ИИ-системы, которые могут представлять угрозу для прав граждан и их безопасности, например разработки, позволяющие манипулировать поведением пользователей. **Будут запрещены.**
  - Высокий риск. Решения в области критической инфраструктуры (например, транспорта), медицины (в частности, роботизированная хирургия), образования, права и др. Должны соответствовать **строгим критериям безопасности** (иметь четкую систему оценки рисков, контроль со стороны людей и др.).
  - Умеренный риск. Пользователи таких ИИ-систем должны четко понимать, что **взаимодействуют с машиной** (например, чат-ботом), а не человеком.
  - Минимальный риск. В эту категорию попадает **большинство ИИ-систем**: интеллектуальные спам-фильтры, видеоигры с поддержкой ИИ и так далее.

СЦИЛЛА И ХАРИБДА РАЗРАБОТЧИКА

- Нет доступа к данным для ИИ-систем
- Юридические ограничения
- Организационные и прочие ограничения



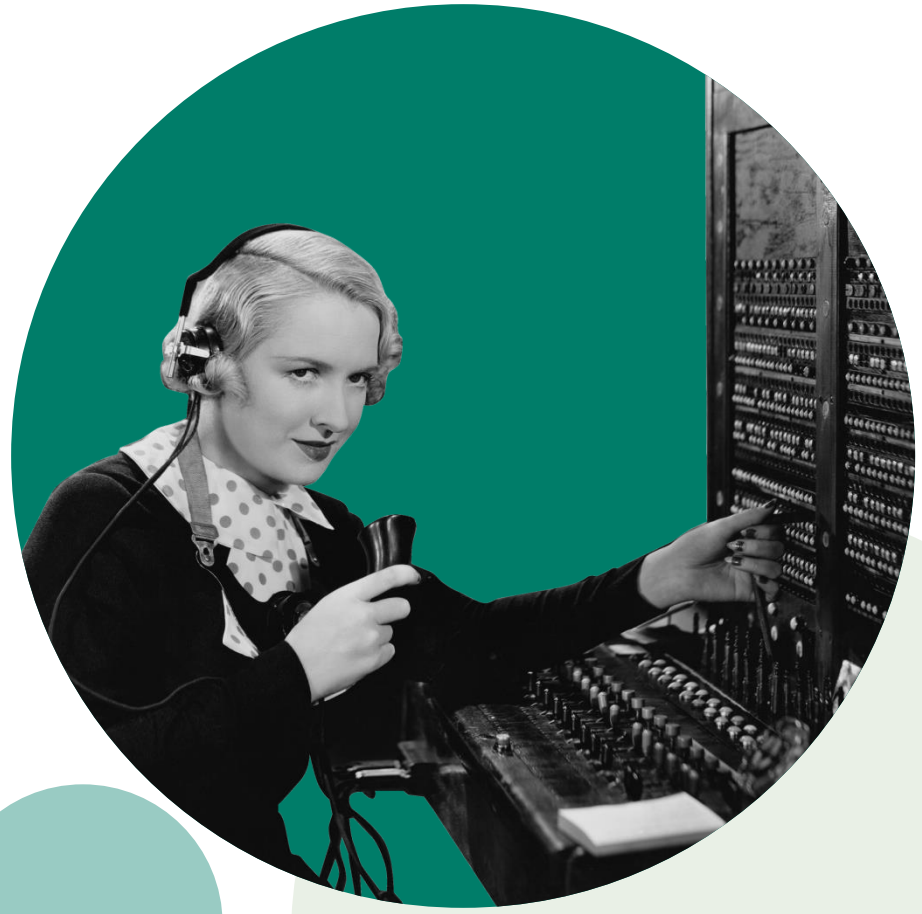
- Этические риски, влекущие за собой риски репутационные, правовые и финансовые
- Несовершенство технологий, усиливающее названные риски

ПРИ ЧЕМ ТУТ ЭТИКА?

- Этика напрямую связана с доверием, а доверие — с успешностью **компании** или **страны** в долгосрочной перспективе.
- Компании, использующие этически неоднозначные решения при разработке продуктов, **ставят под удар репутацию своего бренда**. Организаций, имеющих собственный этический подход, пока не так много, но именно их можно назвать наиболее технологически зрелыми.
- **Доверие граждан государству тает**, когда технологии используют неэтично, когда их применение вызывает страх и наносит вред гражданам.
- «В постковидном мире государства, которые ценят доверие граждан, будут работать над тем, чтобы восстановить его. Они будут применять к технологии подходы, основанные на этике и правах человека… Государства будут уделять больше внимания этическому и правовому измерению технологий и укреплять доверие граждан…» (Тереза Скасса, профессор Оттавского университета, Канада).

ЭТИКА — КОНКУРЕНТНОЕ ПРЕИМУЩЕСТВО

- После известного инцидента с предвзятостью по отношению к женщинам алгоритма оценки соискателей Amazon компания отказалась от этого инструмента, но по-прежнему широко использует алгоритмы для набора и найма.
- Работодатели потенциально упускают ценных кандидатов → со временем эти потери только усугубятся из-за сарафанного радио. Люди узнают о проблемах алгоритмов от членов своего социального круга со сходными демографическими характеристиками.



ЭТИКА — КОНКУРЕНТНОЕ ПРЕИМУЩЕСТВО

- Исследователи смоделировали реакцию клиентов на общение с двумя банками. Один банк имеет различные пороги одобрения кредита для членов разных групп. Например, для получения кредита женщинам, возможно, придется соответствовать более высоким стандартам, чем мужчинам. Другой банк является «слепым для групп»: он имеет одинаковый порог одобрения кредита для всех заявителей.
- Модель показывает, что клиентам из той «правильной группы», которая соответствует требованиям банка, будет без проблем одобрен кредит в обоих банках; члены «неправильной группы» (которая в одном из банков не соответствует требованиям, заложенным в алгоритм) узнают друг у друга о риске дискриминации и переходят в банк, «слепой к группе».
- Члены группы, испытывающей дискриминацию, также влияют на некоторых членов «правильной группы», которые со временем тоже переходят в банк, не использующий дискриминирующих алгоритмов.
- **Итог:** когда потребители узнают друг от друга, что компания с меньшей вероятностью будет их обслуживать, даже если дискриминация нечаянная, они будут избегать этой компании, и она потеряет доход.

ВСЕ ХОТЯТ БЫТЬ БЕЛЫМИ И ПУШИСТЫМИ

- Одна из главных проблем — компании (и государственные органы) выпускают кодексы этики и декларируют намерения, но из кодексов не явствует, как именно будут реализованы заявленные принципы (например, прозрачность, подотчетность или надежность) работы алгоритмов. И, как правило, никакого дальнейшего раскрытия информации по этой теме в работе компании/госоргана не происходит.
- **Ethics-washing** — формальные способы вписаться в этическую повестку, чтобы показать, какие мы «белые и пушистые» (по аналогии с greenwashing — формальным декларированием заботы об окружающей среде без реального включения).

КОНЦЕПЦИЯ РЕШЕНИЯ: ДОВЕРЕННЫЙ ИИ

СТРАТЕГИЯ И ПОДХОДЫ К РЕШЕНИЯМ

Быть этичным более выгодно, чем казаться ☺ Как можно **быть** этичным?

- Использовать фреймворки
- Искать специальные тулкиты
- Разработать регламенты и политики
- Думать об оргструктуре и культуре
- Быть евангелистами этичного ИИ
- Трепетно относиться к данным для ИИ
- Моделировать поведение пользователей
- В перспективе — использовать решения Ethics-as-a-Service

ДОКЛАДЫ ЦЕНТРА В 2019–2021 годах

Навигатор цифровой трансформации: Agile-подход в государственном управлении
gosagile.cdto.ranepa.ru

Модель компетенций команды цифровой трансформации в системе государственного управления»
hr.cdto.ranepa.ru/cm
(зеркало: hr.cdto.center/cm)
совместно с командой Кадрового центра

Клиентоцентричный подход в государственном управлении
cx.cdto.ranepa.ru
(зеркало: cx.cdto.center)

Самоизоляция: работаем, руководим, трансформируем
udalenka.cdto.ranepa.ru
(зеркало: udalenka.cdto.center)

ДОКЛАДЫ ЦЕНТРА В 2019–2021 годах

Организационные структуры
и команды цифровой трансформации
в системе государственного
управления hr.cdto.ranepa.ru/os_0
(зеркало: hr.cdto.center/os_0)
совместно с командой
Кадрового центра

Стратегия цифровой трансформации:
написать, чтобы выполнить
strategy.cdto.ranepa.ru
(зеркало: strategy.cdto.center)

Бережливое управление
в госсекторе.
Как наладить процессы
lean.cdto.ranepa.ru

РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВШГУ РАНХиГС Высшая Школа Государственного Управления

ЦЕНТР ПОДГОТОВКИ РУКОВОДИТЕЛЕЙ И КОМАНД ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Спасибо за внимание!

Екатерина Потапова
potapova-eg@ranepa.ru

ЦЕНТР ПОДГОТОВКИ РУКОВОДИТЕЛЕЙ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

cdto.ranepa.ru

cdtocenter