Исследовательский семинар «Этические проблемы искусственного интеллекта»

Этика искусственного интеллекта в стандартах и рекомендациях

Леушина Влада Вячеславовна

НИЦ «Курчатовский институт»

wandbpand@gmail.com

г. Москва, Институт философии РАН, 26 июня 2024

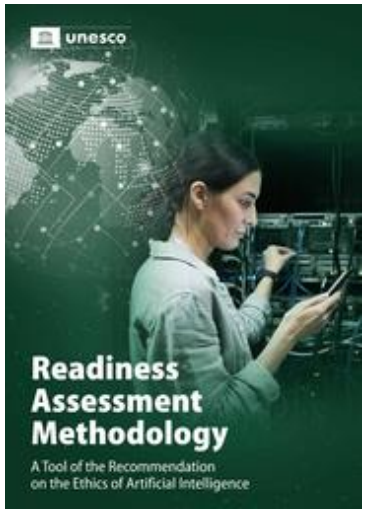
UNESCO | Развитие рекомендаций

(United Nations Educational Scientific and Cultural Organization)

В конце 2021 года была утверждена рекомендация об этических аспектах ИИ, в которой были даны **ключевые характеристики, присущие этичному ИИ** [https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus]

В конце 2023 года была представлены инструменты оценки способности стран справляться с рисками, порождаемыми ИИ [https://unesdoc.unesco.org/ark:/48223/pf0000385198], а также соответствие систем ИИ вышедшим рекомендациям [https://unesdoc.unesco.org/ark:/48223/pf0000386276].

2021 - Утверждение рекомендаций



+

2023 - Оценка надежности и гибкости законов, политики и пр.

2024 - Подписание соглашения с глобальными техно-гигантами о создании более этичного ИИ

UNESCO | Принципы этичного ИИ

(United Nations Educational Scientific and Cultural Organization)

ЦЕННОСТНЫЕ УСТАНОВКИ

1. Уважение, защита и поощрение прав человека и основных свобод и человеческого достоинства
2. Благополучие окружающей среды и экосистем
3. Обеспечение разнообразия и инклюзивности
4. Жизнь в мирных, справедливых и взаимосвязанных обществах

ПРИНЦИПЫ ДЕЯТЕЛЬНОСТИ

5. Соразмерность и не причинение вреда
6. Безопасность и защищенность
7. Справедливость и отказ от дискриминации
8. Устойчивость
9. Право на неприкосновенность частной жизни и защита данных
10. Подконтрольность и подчиненность человеку
11. Прозрачность и объяснимость
12. Ответственность и подотчетность
13. Осведомленность и грамотность
14. Многостороннее и адаптивное управление и взаимодействие

UNESCO | Спорные моменты

(United Nations Educational Scientific and Cultural Organization)

Приоритетная область 6: Гендерное равенство

Пункт 88 о «...целевые ассигнования на **финансирование программ в поддержку гендерного равенства**…» и о «…меры по целевому финансированию программ и **использованию гендерно неспецифического языка** в целях расширения представленности девушек и женщин в области естественных наук, техники, инженерии и математики (ЕНТИМ)…»

Пункт 92 о «...следует поощрять гендерное разнообразие в сфере связанных с ИИ научных исследований … посредством **предоставления девушкам и женщинам льготного доступа** к данной области деятельности …»

Пункт 93 о «...содействовать созданию репозитория передового опыта в области стимулирования участия женщин, девушек **и недостаточно представленных групп населения** во всех этапах жизненного цикла ИИ-систем…»

Рабочая группа ISO/IEC JTC 1/SC 42 | Общая информация
(International Organization for Standardization)

Задача ТК учитывать требования бизнеса, нормативной правовой базы и политики, потребности прикладных областей, этические и общественные проблемы. Семинары ISO/IEC по ИИ проводят два раза в год [https://jtc1info.org/technology/subcommittees/ai/workshops/]

Критически важные принципы:

1. Подотчетность
2. Ответственность
3. Объяснимость
4. Достоверность
5. Безопасность
6. Устойчивость
7. Конфиденциальность
8. Защищенность
9. Беспристрастность
10. Равноправие
11. Толерантность

Выпущенные стандарты

- **«Искусственный интеллект. Функциональная безопасность и системы искусственного интеллекта»** (SO/IEC TR 5469:2024 Artificial intelligence — Functional safety and AI systems)
- **«Искусственный интеллект. Предвзятость в системах искусственного интеллекта и процессе принятия решений искусственным интеллектом»** (ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making)
- …

Разрабатываемые стандарты

- **«Искусственный интеллект. Оценка влияния ИИ»** (ISO/IEC DIS 42005 Artificial intelligence — AI system impact assessment)
- **«Искусственный интеллект. Обработка нежелательной предвзятости при изучении задач классификации и регрессии машинного обучения»** (ISO/IEC DTS 12791.2 Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks)
- …

Российский ТК №164 | Общая информация

Технический Комитет 164

Создан с целью повышения эффективности работы по стандартизации в области искусственного интеллекта на национальном и международном уровнях.

Подкомитеты:
- «Искусственный интеллект в здравоохранении»,
- «Данные»,
- «Искусственный интеллект в дорожно-транспортном комплексе»…

Состав комитета: 63 организации

ПРАВИТЕЛЬСТВО МОСКВЫ

МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ИТМО

СБЕР БАНК

МИНПРОМТОРГ РОССИИ

ФОНД ПЕРСПЕКТИВНЫХ ИССЛЕДОВАНИЙ

ФИЦ КНЦ РАН

СТАНКИН Московский Государственный Технологический Университет

ГАЗПРОМ

Министерство экономического развития РФ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР «КУРЧАТОВСКИЙ ИНСТИТУТ»

РЖД

Яндекс

МГТС

А также:
- «РУССОФТ»
- АНО «Отраслевой Центр Маринет»
- ООО «Криптософт»
- ООО «А-Я эксперт»
- и др.

Российский ТК №164 | Утвержденные стандарты

Технический Комитет 164

- **ГОСТ Р 59276-2020 «Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения»**

  Разработан: АО «Всероссийский научно-исследовательский институт сертификации»,
  ООО «Твпортал»

> 3.15 **сильный (общий) искусственный интеллект:** Способность технической системы, подобно человеку, мыслить, взаимодействовать, адаптироваться к изменяющимся условиям и решать другие задачи в области обработки информации, ассоциирующиеся с естественным интеллектом человека.
>
> 3.16 **система искусственного интеллекта:** Техническая система, в которой используются технологии искусственного интеллекта и обладающая искусственным интеллектом.

- **ГОСТ Р 70980-2023 «Системы искусственного интеллекта на автомобильном транспорте. Системы управления Интеллектуальной транспортной инфраструктурой. Общие требования»**
- **…**

И еще 60+ стандартов, регулирующих общие требования к СИИ: ее разработке, изготовлению и эксплуатации.

Российский ТК №164 | Вопросы этики

Технический Комитет 164

- **ГОСТ Р #####-202# «Системы искусственного интеллекта в здравоохранении. Этические аспекты»**

Разрабатывается: ГБУЗ «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы»,

этика в сфере искусственного интеллекта: Набор моральных принципов и руководящих правил, направленных на обеспечение справедливого и этически ответственного создания и применения систем искусственного интеллекта.

Этические требования к СИИ:

1. конфиденциальность пациентов, все права пациента учтены;
2. применение СИИ не наносит психический или физический ущерб, вред (в т.ч. и косвенно);
3. применение СИИ не приводит к стигматизации и дискриминации пациентов (по половому, национальному и пр. признакам);
4. информирование пациентов о применении СИИ и **формирование понимания роли** СИИ;
5. внедрение СИИ в здравоохранение не ограничивает доступ пациентов к медицинской помощи;
6. пациенты могут сообщать о своих потребностях и **влиять на процесс внедрения СИИ в здравоохранение**;
7. внедрение СИИ способствует развитию персонализированного подхода в лечении, в т.ч. учитываются потребности и особенности человека, **его личные ожидания от лечения**;
8. СИИ не нарушают отношения между врачом и пациентом, на которых строится медицинская практика;
9. права и интересы пациентов и общества учитываются при осуществлении контроля в сфере здравоохранения с использованием СИИ.

Европейский союз | Закон об искусственном интеллекте

Нормативный акт Европейского союза об искусственном интеллекте. Принят Европейским парламентом 13 марта 2024 года и одобрен Советом ЕС 21 мая 2024 года.
Открытый доступ: https://artificialintelligenceact.eu/ai-act-explorer/

Цель: создание общей нормативно-правовой базы для регулирования вопросов, касающихся эксплуатации систем с ИИ.

Закон ЕС об искусственном интеллекте запрещает использование систем ИИ, которые:

- манипулируют решениями людей или используют их уязвимости,
- классифицируют людей на основе их социального поведения или личных качеств,
- прогнозируют риск совершения человеком преступления (не касается тяжелых преступлений),
- извлекают изображения лиц из Интернета или записей с камер видеонаблюдения,
- делают выводы об эмоциях на рабочем месте или в учебных заведениях,
- классифицируют людей на основе их биометрических данных.

Наказание за несоблюдение: административный штраф в размере до 1,5 млн евро.

Итоги

- ✓ Основная идея - создать основу стандарта этики, чтобы в последствии заняться разработкой более конкретных норм.
- ✓ Неравномерное развитие стандартизации понятия этичности И/АС.
- ✓ Параллельная стандартизация принципов этики ИИ.
- ✓ Кардинальных различий в понимании универсальной модели этичного ИИ – нет.
- ✓ **Редко и неявно поднимаются вопросы этики в процессе регулирования И/АС.**

Возникающие вопросы:

- Каким способом будет происходить контроль добровольного соблюдения большого списка этических принципов?
- Каким образом страны будут противостоять несоблюдению/непринятию разработанных нормативных документов?
- **Какие последствия ожидают страны и частные организаций за нарушение действующих норм?**

Текущие проблемы:

- Ориентация стандартов и рекомендаций на производителей, владельцев и пользователей И/АС.
- Присутствие конъюнктурных и сомнительных пассажей.
- **«Беззубость» и неконкретность рекомендаций.**

Исследовательский семинар «Этические проблемы искусственного интеллекта»

Этика искусственного интеллекта в стандартах и рекомендациях

Леушина Влада Вячеславовна

НИЦ «Курчатовский институт»

wandbpand@gmail.com

г. Москва, Институт философии РАН, 26 июня 2024

Рекомендации. Альянс в сфере ИИ| Кодекс этики ИИ

«Объединяет ведущие технологические компании для совместного развития их компетенций и ускоренного внедрения искусственного интеллекта в образовании, научных исследованиях и в практической деятельности бизнеса.»

Цель рекомендации: быть ориентиром для развития технологий ИИ в стране и обеспечивать доверие к ИИ со стороны пользователей, общества и государства.

Кодекс устанавливает общие этические принципы и стандарты поведения, которым следует руководствоваться тем, кто занимается созданием, внедрением или использованием технологий ИИ. Крупнейшие технологические компании России приняли этот кодекс. Церемония подписания прошла в рамках I форума *«Этика искусственного интеллекта: начало доверия»* 26 октября 2021 года.

ГАЗПРОМ СБЕР Sk Сколково Яндекс МТС ИСП РАН **и др.**

Спорный момент:

5. Создать при Президенте Российской Федерации Совет по этике новых технологий, состоящий из авторитетных представителей общества – **артистов, композиторов, художников, юристов,** философов, филологов, психологов, **педагогов,** социологов, **врачей, спортсменов.**

Кодекс этики ИИ | Принципы этичного ИИ

- Человеко-ориентированный и гуманистический подход
- Уважение автономии и свободы воли человека
- Соответствие закону
- Недискриминация
- Оценка рисков и гуманитарного воздействия
- Риск-ориентированный подход
- Ответственное отношение
- Предосторожность
- Непричинение вреда
- Идентификация ИИ в общении с человеком
- Безопасность работы с данными
- Информационная безопасность
- Добровольная сертификация и соответствие положениям Кодекса
- Контроль рекурсивного самосовершенствования СИИ
- Поднадзорность
- Ответственность
- Применение СИИ в соответствии с предназначением
- Стимулирование развития ИИ
- Корректность сравнений СИИ
- Развитие компетенций
- Сотрудничество разработчиков
- Достоверность информации о СИИ
- Повышение осведомлённости об этике применения

Кодекс этики ИИ | Сходства с рекомендацией ЮНЕСКО

- **Человеко-ориентированный и гуманистический подход**
- **Уважение автономии и свободы воли человека**
- **Соответствие закону**
- **Недискриминация**
- **Оценка рисков и гуманитарного воздействия**
- **Риск-ориентированный подход**
- **Ответственное отношение**
- **Предосторожность**
- **Непричинение вреда**
- Идентификация ИИ в общении с человеком
- **Безопасность работы с данными**
- **Информационная безопасность**
- Добровольная сертификация и соответствие положениям Кодекса
- Контроль рекурсивного самосовершенствования СИИ
- **Поднадзорность**
- **Ответственность**
- **Применение СИИ в соответствии с предназначением**
- **Стимулирование развития ИИ**
- Корректность сравнений СИИ
- Развитие компетенций
- **Сотрудничество разработчиков**
- **Достоверность информации о СИИ**
- **Повышение осведомлённости об этике применения**

Кодекс этики ИИ | Выделяющиеся принципы

- Человеко-ориентированный и гуманистический подход
- Уважение автономии и свободы воли человека
- Соответствие закону
- Недискриминация
- Оценка рисков и гуманитарного воздействия
- Риск-ориентированный подход
- Ответственное отношение
- Предосторожность
- Непричинение вреда
- **Идентификация ИИ в общении с человеком**
- Безопасность работы с данными
- Информационная безопасность
- **Добровольная сертификация и соответствие положениям Кодекса**
- **Контроль рекурсивного самосовершенствования СИИ**
- Поднадзорность
- Ответственность
- Применение СИИ в соответствии с предназначением
- Стимулирование развития ИИ
- **Корректность сравнений СИИ**
- **Развитие компетенций**
- Сотрудничество разработчиков
- Достоверность информации о СИИ
- Повышение осведомлённости об этике применения

Кодекс явно ориентирован на производителей, владельцев и пользователей И/АС.

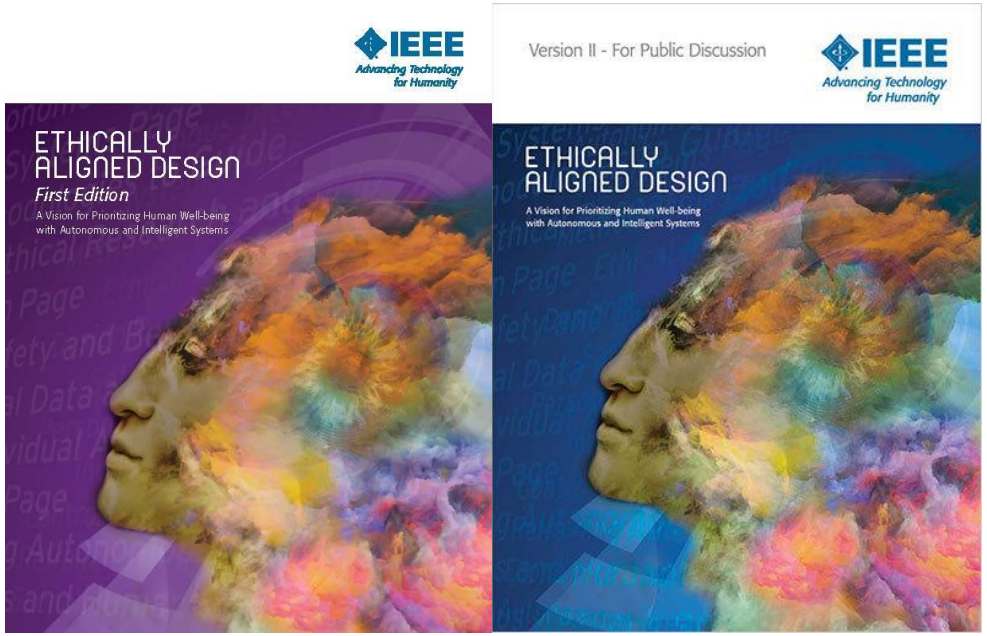
Стандарты. Организация IEEE
(Institute of Electrical and Electronics Engineers)

Глобальная инициатива: обучить и дать возможность (полномочия) всем заинтересованным людям, чья деятельность касается автономных и интеллектуальных систем (А/ИС), уделять первостепенное внимание этическим аспектам.

Позиция: разрабатываемые системы должны не только разрешать технические проблемы, но и учитывать выгоду для человечества

Результаты инициативы:

1) документ «Этически обоснованное проектирование» («Ethically aligned design»);
2) группа стандартов этики ИИ P7000.



Этически обоснованное проектирование

IEEE | Группа P7000

Выпущенные стандарты

- [P7000] **Model Process for Addressing Ethical Concerns During System Design**
- [P7005] **Standard on Employer Data Governance**
- [P7007] **Ontological Standard for Ethically driven Robotics and Automation Systems**
- [P7010] **IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being**

Разрабатываемые стандарты

- [P7001] Transparency of Autonomous Systems
- [P7002] Data Privacy Process
- [P7003] Algorithmic Bias Considerations
- ...
- [P7014] Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems

IEEE | Стандарт P7000

<u>Цель стандарта:</u> дать организациям возможность разрабатывать системы с учетом индивидуальных и общественных этических ценностей (таких как конфиденциальность, устойчивость, подотчетность и др.), а также критериев, которые обычно учитываются при разработке (например, эффективность).

Появление понятия этики в стандарте:

Этика: область знаний или теории, расследующая верные причины рассуждать: верно то или другое
Этические ценности: ценности, направленные на суждение с точки зрения человеческой культуры о том, что правильно, а что нет.
Этический принцип: общее представление об этических ценностях, которое сообщество может добиться и обеспечить.
Этический: поддерживающий возможность реализовать положительные ценности или уменьшить количество негативных.

<u>Подход</u>

~~Определение этичности И/АС, универсальных этических принципов~~ → Конкретные алгоритмы и нормы для внедрения этических принципов в И/АС

Стандарт не определяет этична А/ИС или нет!

Сравнение этических принципов в стандартах и рекомендациях. ЮНЕСКО, ISO, Кодекс ИИ

1. Уважение, защита и поощрение прав человека и основных свобод и человеческого достоинства
2. Благополучие окружающей среды и экосистем
3. Обеспечение разнообразия и инклюзивности
4. Жизнь в мирных, справедливых и взаимосвязанных обществах
5. Соразмерность и не причинение вреда
6. Безопасность и защищенность
7. Справедливость и отказ от дискриминации
8. Устойчивость
9. Право на неприкосновенность частной жизни и защита данных
10. Подконтрольность и подчиненность человеку
11. Прозрачность и объяснимость
12. Ответственность и подотчетность
13. Осведомленность и грамотность
14. Многостороннее и адаптивное управление и взаимодействие

ISO

Кодекс ИИ

Каждая организация считает, что благополучие человека – самый важный аспект, который необходимо учитывать при разработке и эксплуатации И/АС.