Семинар «Этические проблемы искусственного интеллекта»

Этика искусственного интеллекта в стандартах и рекомендациях

Докладчик: Леушина Влада Вячеславовна

(НИЦ «Курчатовский институт», НИУ «Высшая школа экономики»)

wandbpand@gmail.com

09 февраля 2022

Начало стандартизации понятия этичности И/АС

Прогресс технологий искусственного интеллекта (ИИ) и активное внедрение их в повседневную жизнь повлекли за собой необходимость разработки нормативных документов, регулирующих возникающие этические проблемы. Этические вопросы, касающиеся ИИ, стоят перед всем мировым сообществом, следовательно существует необходимость в нормативном документе, которому смогут следовать одновременно все страны, чтобы на его основе стало возможным сформулировать уточняющие стандарты или рекомендации, учитывающие собственные ценности и нравы страны.

Начальная ситуация:

- Активные массовые обсуждения этических вопросов.
- Повышенный интерес со стороны широкой общественности.
- Путаница из-за некорректных определении и не понимании сути проблем.
- Отсутствие глобальных, мировых, международных нормативных документов.
- Большое количество локальных корпоративных этических норм.

Многие организации уже начали предлагать свои виденья развития данного направления, и как результат – публиковать стандарты и рекомендации, регулирующие этические вопросы ИИ в разных областях.

Рассматриваемые нормативные документы:

- Рекомендации [UNESCO | Альянс в сфере ИИ]
- Стандарты [IEEE | ISO | ГОСТ]

Рекомендации. Организация UNESCO
(United Nations Educational Scientific and Cultural Organization)

«Миру необходимы правила с целью разработки таких технологий искусственного интеллекта, которые работали бы на благо человечества. Рекомендация по этическим аспектам ИИ - это важнейший шаг в этом направлении. Она устанавливает первую глобальную нормативную основу, возлагая на государства ответственность за ее применение на своем уровне…»

Одрэ Азуле, Генеральный директор ЮНЕСКО

Цель рекомендации: реализация преимуществ, которые искусственный интеллект приносит обществу, и уменьшение связанных с этим рисков.

В конце 2021 года генеральная конференция ЮНЕСКО **утвердила** рекомендацию об этических аспектах ИИ. В этом документе описываются **ключевые характеристики, присущие этичному ИИ**. Таким образом ЮНЕСКО определяет универсальную модель ИИ, ценностные установки и принципы деятельности которой должны соблюдаться всеми заинтересованными сторонами.

UNESCO | Принципы этичного ИИ

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 в рекомендации

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…»

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.

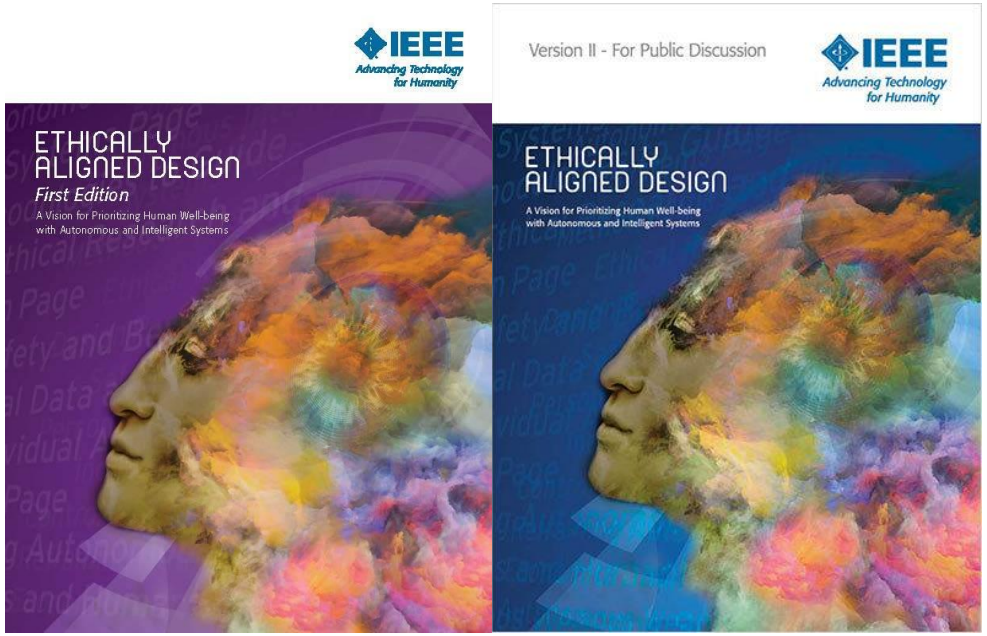
Стандарты. Организация IEEE
(Institute of Electrical and Electronics Engineers)

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!

Стандарты. Организация ISO
(International Organization for Standardization)

Позиция: международные стандарты помогут создать этическую основу для разработки и эксплуатации систем в будущем.

В ISO решением этических вопросов в сфере ИИ занимается технический подкомитет ISO/IEC JTC 1/SC 42 «Искусственный интеллект». На данный момент ***ISO не представила окончательный вариант*** своего видения ***универсальной модели этичного ИИ***.

Выпущенные стандарты

- **«Искусственный интеллект. Доверенность в искусственном интеллекте. Общие положения»** (ISO/IEC TR 24 028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence)
- **«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)

Разрабатываемый стандарт:

- **«Искусственный интеллект. Этические и социальные проблемы. Общие положения»** (ISO/IEC AWI TR 24 368 Information technology — Artificial intelligence — Overview of ethical and societal concerns)

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

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.

Итог

- ✓ Неравномерное развитие стандартизации понятия этичности И/АС.
- ✓ Многие организации работают над стандартизацией вопросов этики.
- ✓ Основная идея - создать основу стандарта этики, чтобы в последствии заняться разработкой более конкретных норм.
- ✓ Кардинальных различий в понимании универсальной модели этичного ИИ нет.

Возникающие вопросы:

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

Текущие проблемы:

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