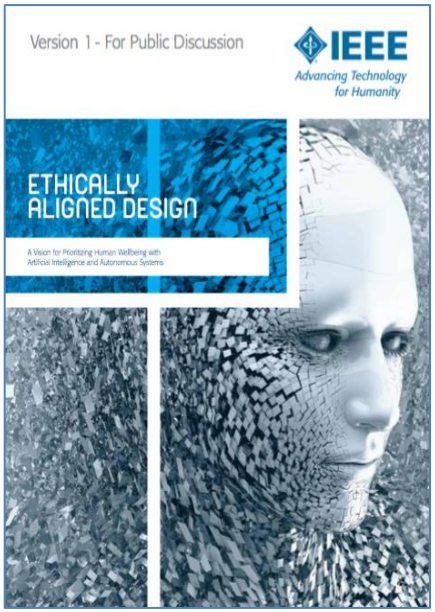
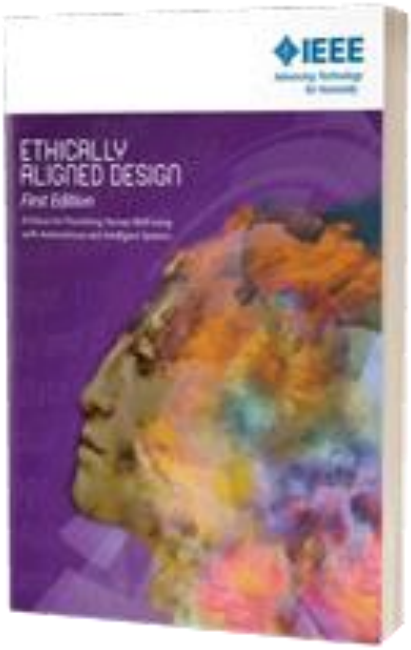
Обзор подходов к валидации интеллектуальных систем на этичность: от научных исследований до нормативной документации

Москва, 2024

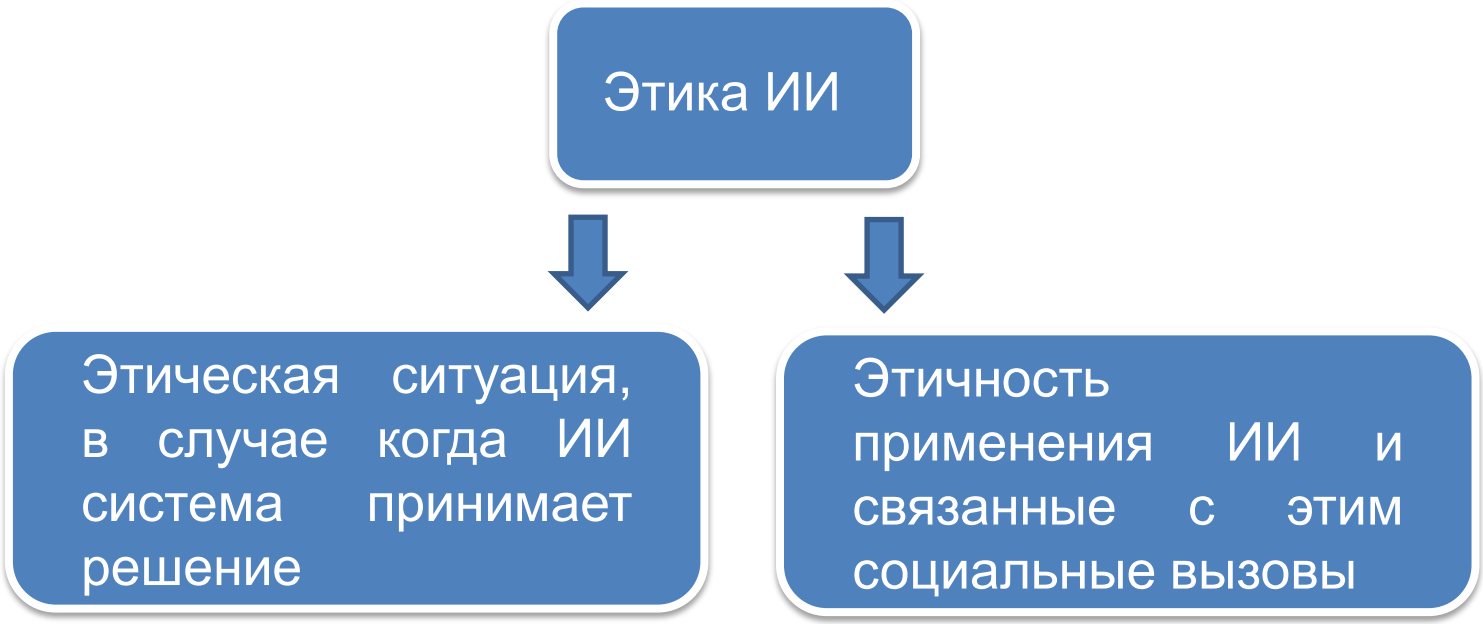
Этика ИАС



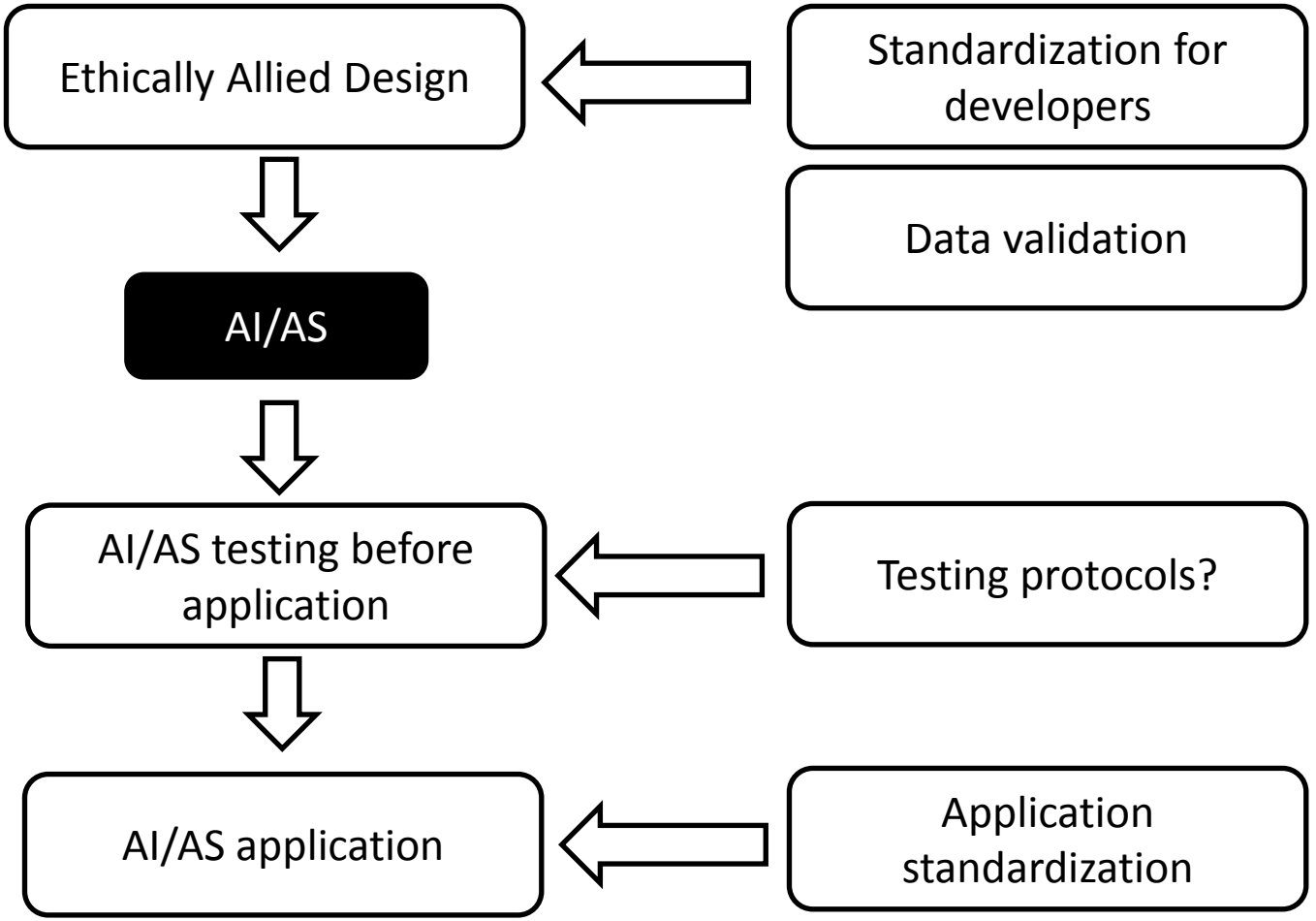
- Базовое издание выпущено в начале 2019 г. под лицензией Creative Commons для общего доступа;
- Разработан более чем 700 экспертами со всего мира в ходе абсолютно открытого проекта;
- Содержит основные вызовы связанные с ИИ и автономными системами;
- Разработан по принципам технических рекомендация, для простоты использования при разработке регулирования ИИ и автономных систем;



Этика ИАС


Этика ИИ
Этическая ситуация, в случае когда ИИ система принимает решение
Этичность применения ИИ и связанные с этим социальные вызовы

Этика ИАС


Ethically Allied Design
Standardization for developers
Data validation
AI/AS
AI/AS testing before application
Testing protocols?
AI/AS application
Application standardization

Этика ИАС в документах

Что нужно что бы можно было валидировать ИАС как этичную?

- Формализация термина «этичный» применительно к ИАС;

Стандарты и другие документы

- Возможность объяснения того, как ИАС пришла к определенному решению

Математические и алгоритмические подходы

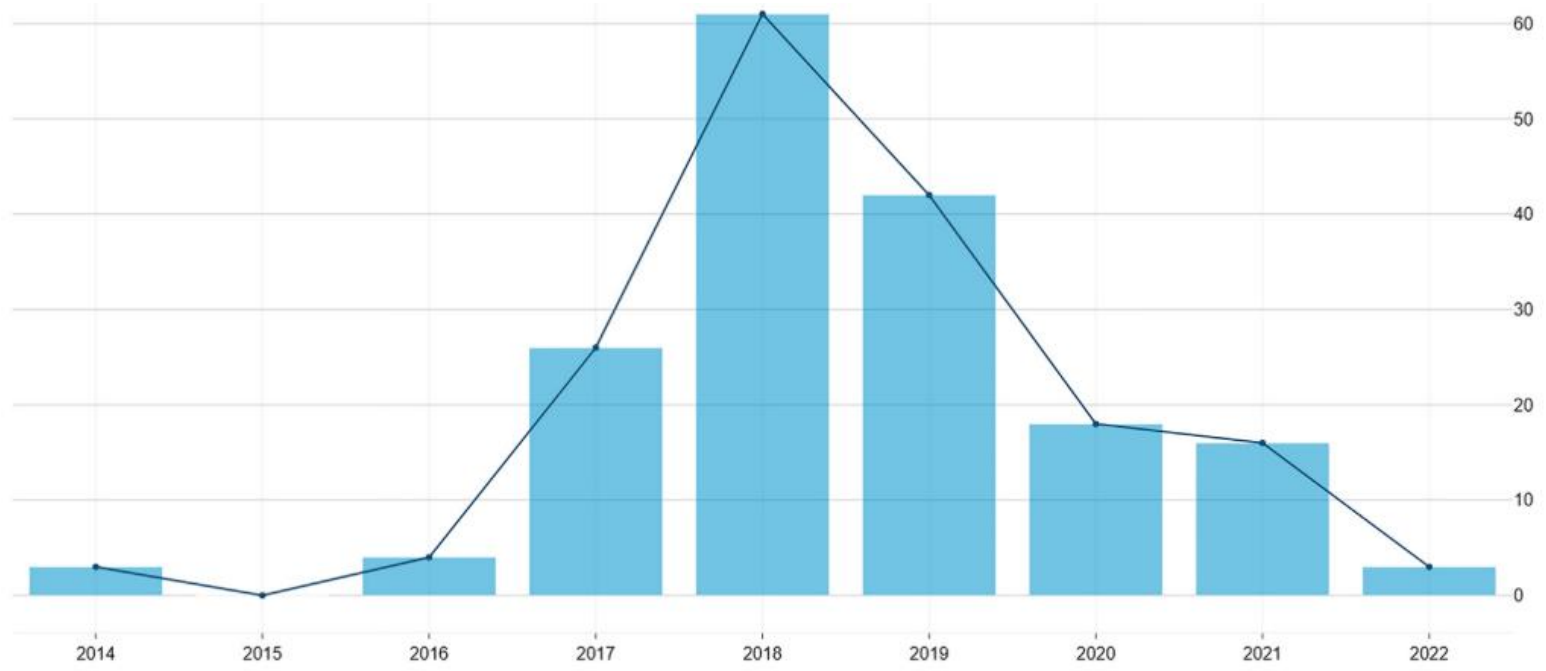
Этика ИАС в документах

Виды документов:

1. Технические стандарты;
2. Законодательные инициативы;
3. Общие руководства и рекомендации;
4. Корпоративные руководства;
5. Дорожные карты, бюллетени и т.д. различных некоммерческих организаций и сообществ.

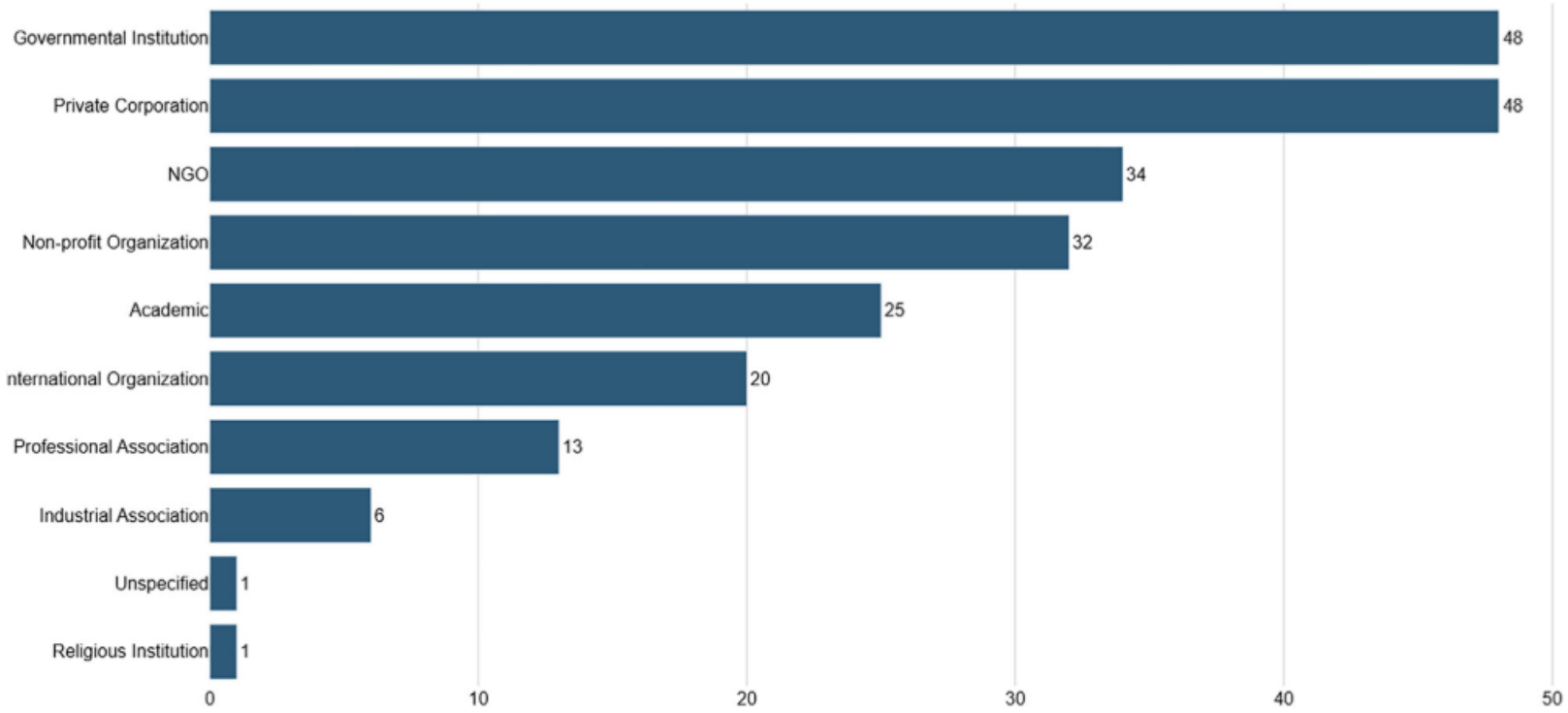
Этика ИАС в документах

К началу 2023 года разработано более 200 различных руководств, рекомендаций и нормативных документов по вопросам связанным с этикой ИАС



Corrêa NK, Galvão C, Santos JW, et al (2023) Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. Patterns 4:. https://doi.org/10.1016/j.patter.2023.100857

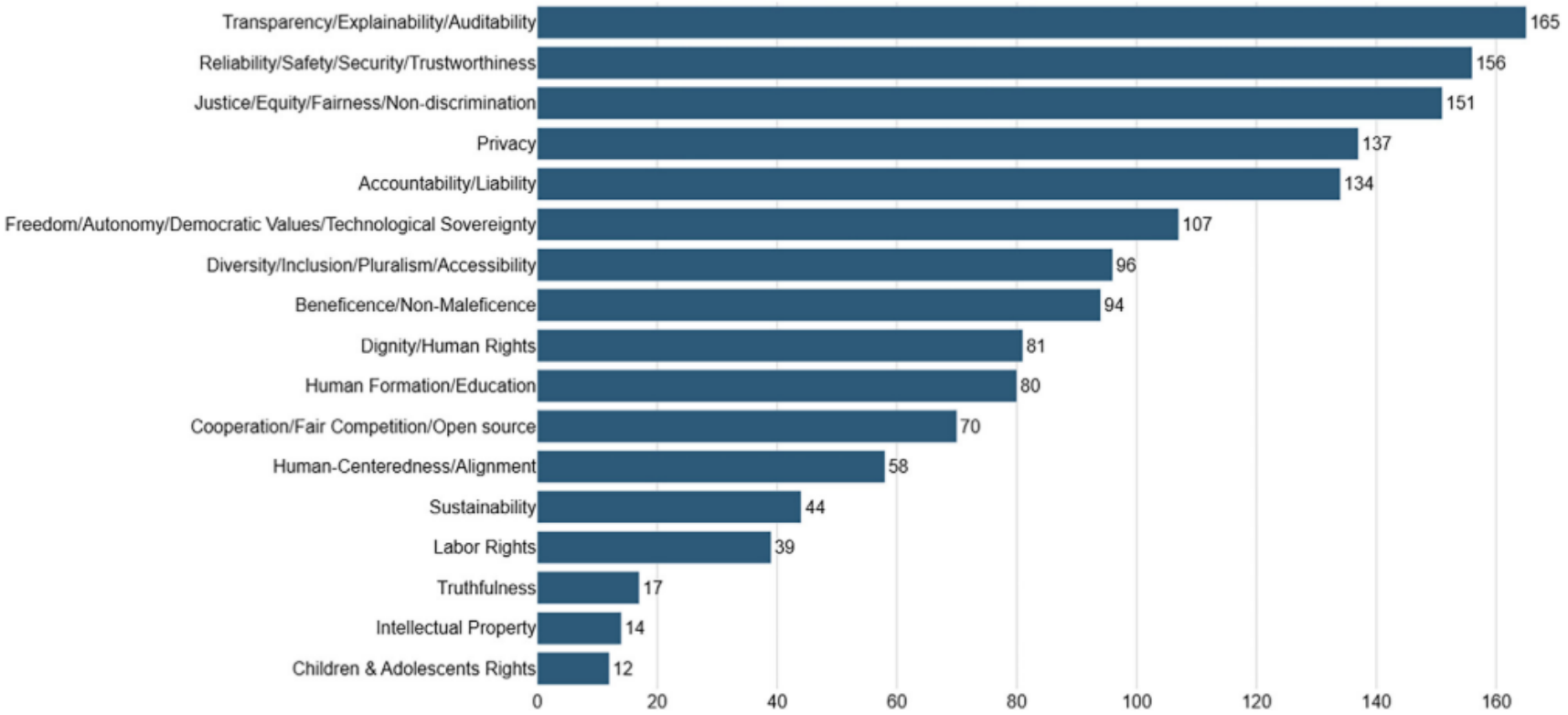
Этика ИАС в документах



Corrêa NK, Galvão C, Santos JW, et al (2023) Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. Patterns 4:. https://doi.org/10.1016/j.patter.2023.100857

Этика ИАС в документах

Количество упоминаний разных тематик, связанных с этикой ИАС в документах:



Corrêa NK, Galvão C, Santos JW, et al (2023) Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. Patterns 4:. https://doi.org/10.1016/j.patter.2023.100857

Этика ИАС в документах

Семейство стандартов IEEE P7000

Всего 18 стандартов и 2 дополнения, из них принято 5 стандартов.

IEEE 7000-2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design

Стандарт предоставляющий ряд рекомендаций и подходов с помощью которых организация может включить в процесс разработки АИС вопросы этики.

IEEE 7007-2021: Ontological Standard for Ethically Driven Robotics and Automation Systems

Проект стандарта, который задает набор базовых онтологий для использования при разработке АС.

IEEE P7008: Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems

Проект стандарта, который регламентирует подходы к разработке этически обоснованных взаимодействий между человеком и ИИ/АС, в части повседневного взаимодействия.

IEEE P7009: Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems

Стандарт регламентирует стадии разработки и тестирования АС, направленные на исключение неполадок, в том числе и ведущих к негативным последствиям при взаимодействии с людьми.

IEEE 7010-2020 : Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems

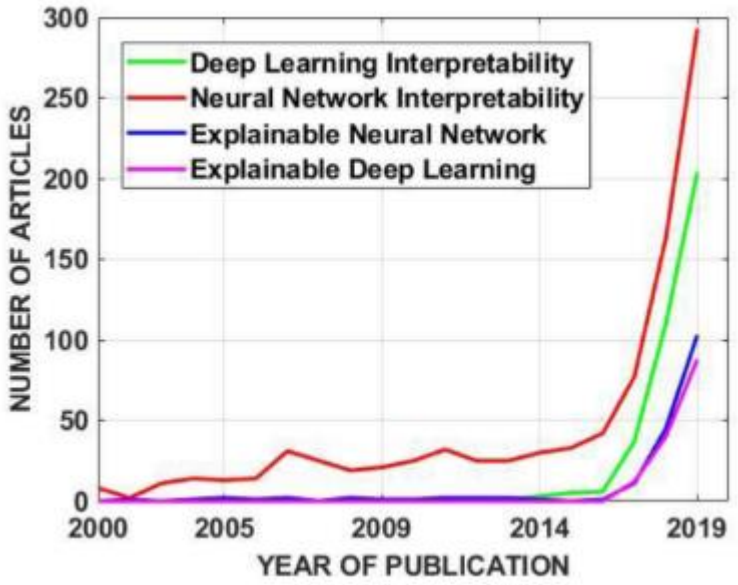
Определяют метрики и базовые уровни описывающие уровень этически обоснованного взаимодействия между человеком и ИИ/АС.

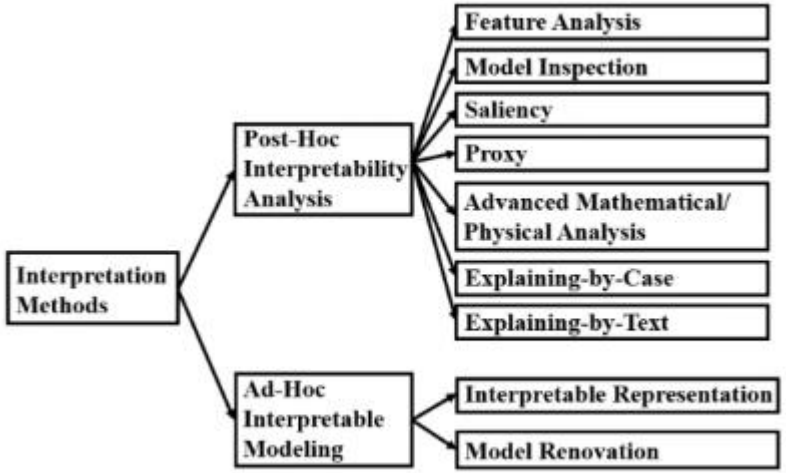
https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/#p7000

Этика ИАС в документах

- Значительное количество документов не содержит ничего кроме общих добрых пожеланий;
- Часть документов просто бессмысленно, так как авторы имеют представление об ИАС по не самым удачным кинофильмам;
- Осмысленная стандартизация не рассматривает вопросы валидации действующих ИАС, акцентируя внимание на стадии разработки;
- Необходимо создание отдельных формализуемых «этических» норм для ИАС с привязкой к областям применения.

Прозрачность и объяснимость ИАС

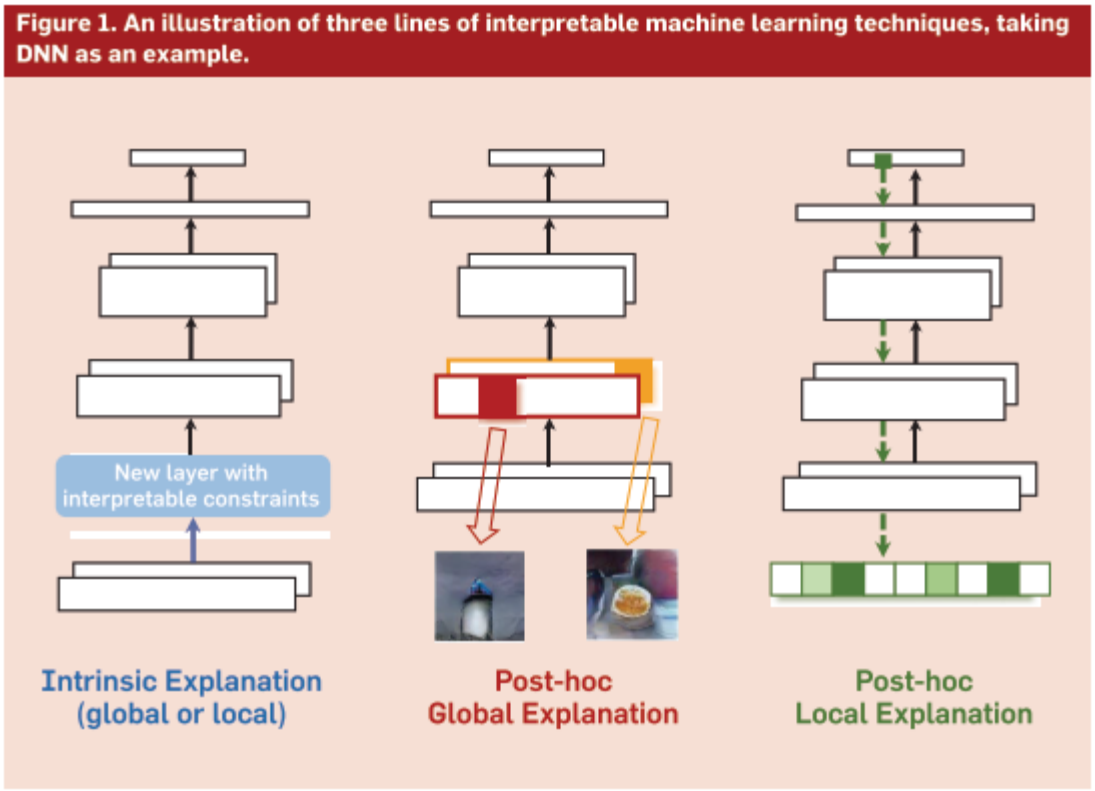




Fan FL, Xiong J, Li M, Wang G (2021) On Interpretability of Artificial Neural Networks: A Survey. IEEE Trans Radiat Plasma Med Sci 5:741–760. https://doi.org/10.1109/TRPMS.2021.3066428

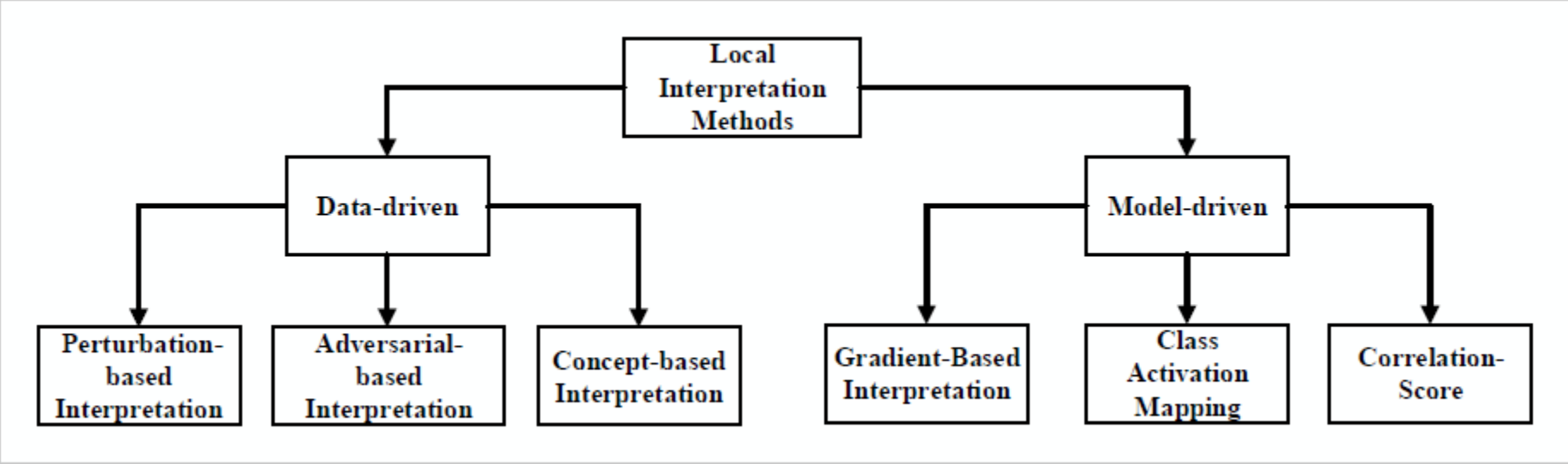
Прозрачность и объяснимость ИАС

Figure 1. An illustration of three lines of interpretable machine learning techniques, taking DNN as an example.



Du M, Liu N, Hu X (2020) Techniques for interpretable machine learning. Commun ACM 63:68–77.
https://doi.org/10.1145/3359786

Прозрачность и объяснимость ИАС



Liang Y, Li S, Yan C, et al (2021) Explaining the black-box model: A survey of local interpretation methods for deep neural networks. Neurocomputing 419:168–182. https://doi.org/10.1016/J.NEUCOM.2020.08.011

Прозрачность и объяснимость ИАС

Что нужно что бы можно было валидировать ИАС как этичную?

- Формализация термина «этичный» применительно к ИАС;

⇩

В документах не отражено, есть отдельные исследования

- Возможность объяснения того, как ИАС пришла к определенному решению

⇩

Ведутся многочисленные исследования, есть надежда на скорое появление в практике

Спасибо за внимание