Доверие
к системам
искусственного интеллекта
для здравоохранения

ITM-AI’2023

ГОСТ Р 59276-2020

Системы искусственного интеллекта

Способы обеспечения доверия

Общие положения

Доверие к ИИ

Понятие доверенного ИИ (англ. Trustworthy – заслуживающий доверия):

(1) законный, соответствующий регламентирующим документам

(2) этичный, обеспечивающий соблюдение этических принципов

(3) надежный с технической и социальной точки зрения

[*Руководство по этике для надежного ИИ. Европейская комиссия, 2019*]

Доверие к исходным знаниям

- Экспертная оценка показателей с позиций релевантности, выраженности, динамики.
- Автоматически извлеченные из текстов знания – оценка источников, включая учет нетипичных ситуаций и связей между показателями.

Доверие к знаниям из больших данных

Качественные, верифицированные Data Set:

- полнота и качество записей,
- период формирования данных (неизменность классификаций и референсных значений показателей, новые методы и новая аппаратура),
- неоднозначность описаний,
- согласованность научных школ.

Важность предпроцессинга и экспертная интерпретация выявляемых знаний

Робастность (устойчивость к внешним воздействиям) и корректность функционирования нейросетей

(а) адекватность предшествующего опыта обучения для текущего набора входных данных (возможность деформации входных образов за истекший период времени)

(б) наличие особенностей в данных, полученных при других условиях

(в) получение данных с использованием другой аппаратуры.

Объяснимость систем ИИ

- Прозрачность / интерпретируемость / объяснимость бизнес-моделей / алгоритмов, данных (для машинного обучения), предлагаемых гипотез (решений).
- Ориентация объяснений на пользователей, участвующих в ЛДП на различных этапах.

Валидация интеллектуальных систем

Комплексная проверка программных продуктов (до экспертизы).

- Область применимости и ограничения
- Надежность
- Эффективность
- Безопасность в плане потенциального риска для пациентов

Верификация

Верификация данных на соответствие требованиям при использовании в частично отличающихся условиях (апробация в различных организациях):

- Область применимости и ограничения
- Надежность
- Эффективность
- Безопасность в плане потенциального риска для пациентов

Жизненный цикл:

- Новые версии
- Коррекция (модификация) версий

Обсуждение

- Верификация / валидация систем ИИ – *кто, каким образом, где*?
- Экспертиза систем ИИ в процессе жизненного цикла с учетом перехода на новые версии
- Объяснимость ИИ для конечного пользователя