Проблема сертификации моделей искусственного интеллекта

Лось В.П., д.воен.н., профессор,
ГНС,
Президент – Председатель Правления АЗИ

Тышук Е.Д., СНС РГГУ

Проблемы общего характера

Спекуляции с термином «интеллектуальные».

Несоответствие между реальными угрозами ИИ и общественным восприятием, основанным на мифах и гипотетических сценариях.

Отсутствие доверия к моделям ИИ, отсутствие механизмов проверки моделей ИИ на предмет их безопасности и применимости.

Вопросы

- 1. О критериях интеллектуальности информационных технологий.
- 2. Мифы об искусственном интеллекте.
- 3. Подходы к сертификации ИИ. Проблемы.

О критериях интеллектуальности информационных технологий

СПЕКУЛЯЦИИ
на тему «Искусственный интеллект»

В результате банальная автоматизация приобретает свойство интеллектуальности

Лифт

Умный лифт оборудованный передовыми технологиями, которые улучшают его функциональность, безопасность и энергоэффективность. Он также может быть панорамным. Интегрируется с сетью Интернет (IoT) и использует различные датчики, аналитику данных, программное обеспечение и другие технические возможности. Некоторые функции **интеллектуальных лифтов…**

Кассовая техника

В соответствии с набором функций кассовая техника делится на две большие группы:
интеллектуальная и неинтеллектуальная.
Интеллектуальное оборудование представляет собой механизмы,
позволяющие не только хранить информацию, но и систематизировать, обрабатывать и анализировать ее.
Неинтеллектуальные кассовые аппараты обладают гораздо более скромным набором функций и подходят для использования на небольших торговых точках с ограниченным ассортиментным перечнем.


СВЕТИЛЬНИК-
ФОНАРЬ
для велосипеда
интеллектуальный
сигнальный
ПОДАРОК
НА 23 ФЕВРАЛЯ!
Похожие

Всегда ли оправдано использование термина «интеллектуальный»? Какие критерии существуют для отнесения технологий и устройств к этой категории?

Немного о терминологии и критериях

Существует точка зрения о том, что интеллектом может обладать только человек.
Основанием для такого вывода служит анализ формирования интеллекта у человека. Когда начинается этот процесс?

В статье Бесковой, И.А. Естественный и искусственный интеллект: точки соприкосновения. Вопросы Философии, 2023, (9), 83–92, исследуется человеческая интеллектуальность, подчеркивается важность пренатального (дородового) опыта, который формирует интуитивное постижение мира. По её мнению, этот опыт формирует «интенсиональные» содержания мышления, в то время как современные системы ИИ, основанные на «экстенсиональном» восприятии через логические цепочки, пока не способны воспроизвести этот уровень.

Вместе с тем, термин «искусственный интеллект» узаконен:

Распоряжение Правительства РФ от 19.08.2020 № 2129-р <Об утверждении Концепции развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 года>

Целью Концепции является определение основных подходов к трансформации системы нормативного регулирования в Российской Федерации для обеспечения возможности создания и применения таких технологий в различных сферах экономики с соблюдением прав граждан и обеспечением **безопасности** личности, общества и государства.

Одновременно целями Концепции являются создание предпосылок для формирования основ **правового регулирования новых общественных отношений**, складывающихся в связи с разработкой и применением технологий искусственного интеллекта и робототехники, а также определение правовых барьеров, препятствующих разработке и применению указанных систем.

Указ Президента РФ от 10.10.2019 N 490
(ред. от 15.02.2024)
«О развитии искусственного интеллекта в Российской Федерации»
(вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 года»)

Искусственный интеллект - комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые с результатами интеллектуальной деятельности человека или превосходящие их.

Как измерить результаты, получаемые ИИ, и их сопоставимость с результатами интеллектуальной деятельности человека или превосходящие их.
Что из себя должна представлять шкала измерений этих результатов?

В работе «Подход к оцениванию уровня интеллектуальности информационной системы. Онтология проектирования, 2023, 13(1 (47)), 29-43, ее автор Микони С.В. выделил 4 категории проявления интеллекта в процессе познавательной деятельности:

1. взаимодействие со средой (восприятие, преобразование, сохранение информации);
2. применение знания (вспоминание, распознавание, выбор);

3. получение нового знания (понимание, представление, присоединение к имеющемуся знанию);
4.порождение знания (воображение, соображение, рассуждение, предвидение).
Анализируя долю ресурсов, выделенных на реализацию каждой функции, можно получить количественную оценку уровня интеллектуальности ИС.

$$У_и = \alpha_1 * x_1 + \alpha_2 * x_2 + \alpha_3 * x_3 + \alpha_4 * x_4$$

Мифы об искусственном интеллекте

МИФ № 1. ИИ обязательно превзойдет человека.

МИФ № 2. Возможность полной автономности ИИ.

Распространённые мифы об искусственном интеллекте:

Искусственный интеллект обретёт разум. Футурологи предполагают, что ИИ начнёт осознавать себя как живое существо и даже может «восстать» против человечества. В реальности ИИ представляет собой набор алгоритмов.

Искусственный интеллект не ошибается. Неверно полагать, что на любой вопрос ИИ отвечает достоверной информацией. Современные нейросети нередко выдают заведомо ошибочные или несуществующие данные.

Искусственный интеллект становится умнее со временем. Системы искусственного интеллекта не самосовершенствуются, по крайней мере, редко. Они требуют постоянного обучения с использованием новых данных и усовершенствованных алгоритмов для повышения производительности.

Искусственный интеллект способен творить. Миф связан с популяризацией нейросетей для создания контента. Развенчивается он тем, что искусство от ИИ всегда основывается на каких-то референсах, создать что-то с нуля они не способны.

Оценка рисков и угроз

ИИ может стать источником угроз, сопоставимых с глобальными катастрофами. Важен контроль использования ИИ для предотвращения катастроф.

Внедрение ИИ в бизнес несет новые риски для безопасности.

Ключевым риском является недостаточная зрелость технологий.

Общественное восприятие угроз ИИ: результаты опроса

27% россиян считают ИИ угрозой,

51% — считают, что технологии упрощают жизнь.

31% респондентов испытывают страх потери работы из-за автоматизации , особенно творческие профессии.

45% респондентов видят киберпреступления главной угрозой.

37% обеспокоены утечкой данных.

Этические аспекты использования ИИ

Этика использования ИИ становится важной в связи с его внедрением во многие сферы и появлением возм манипуляции общественным мнением.

Ключевые вопросы:

прозрачность алгоритмов;

ответственность за решения ИИ.

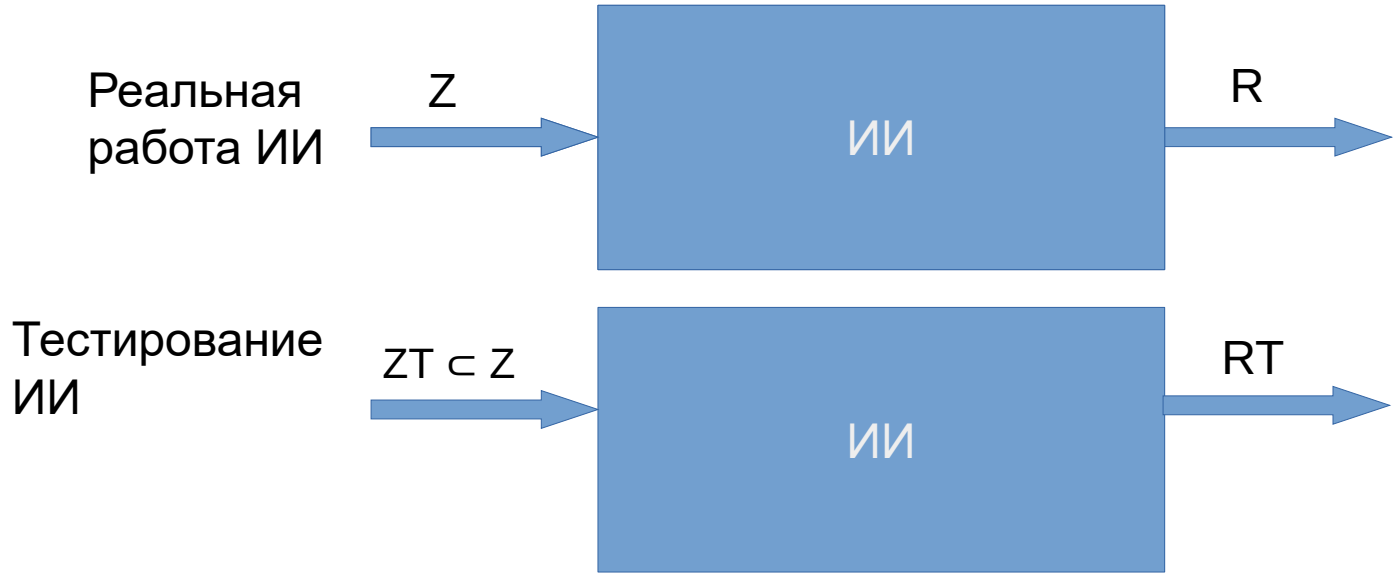
Неясность в алгоритмах порождает недоверие, а предвзятость данных ведёт к необъективности. Нужна работа по интеграции этих вопросов и нормативные практики, чтобы технологии служили прогрессу и не создавали угроз. Этические нормы должны формироваться с учетом мнений общества.

О сертификации моделей ИИ

Повышение доверия к ИИ возможно при введении каких-либо механизмов проверки ИИ на предмет безопасности и применимости в определенной предметной области.

О сертификации моделей ИИ

Одним из таких механизмов является сертификация моделей ИИ. Поскольку в открытых источниках, как правило, нет информации о структуре нейросети, используемой в модели, алгоритмах обработки запросов, исходной обучающей выборки, то модель ИИ представляется для исследователя некоторым «черным ящиком». В этом случае имеется две возможности: возможность проведения каких-либо манипуляций на входе модели и возможность интерпретации полученной информации на выходе.



Вопросы:

1. Какова должна быть размерность ZT.

2. Если при тестировании процент ошибок составил некоторую величину β, какие предположения можно сделать в целом для модели ИИ.

Заключение

Искусственный интеллект стал важной технологией, открывая новые горизонты, но также порождая угрозы. Для безопасного использования важны этические нормы и сертификация моделей ИИ на предмет безопасности и применимости в определенной предметной области.

СПАСИБО ЗА ВНИМАНИЕ