Искусственный интеллект в кибербезопасности:

защита будущего или новые угрозы?

чл.-корр. РАН, д.т.н., профессор
Зегжда Д.П.

д.т.н., доцент
Полтавцева М.А.

ПОЛИТЕХ
Институт компьютерных
наук и кибербезопасности

ИИ и Кибербезопасность

ИКНК

Поиск уязвимостей в логике приложений

Автоматизированные атаки / тестирование

Анализ поведения

Генерация / обнаружение спама

Дипфейки и их обнаружение

ИИ внутри ВПО…

Защита цепочек поставок

Проверка целостности

Доверие к системам ИИ

Развитие известных атак на ИИ

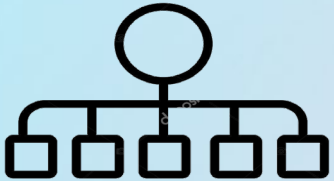
Появление первых атак, нацеленных на большие языковые модели (LLM)

Появление атак, нацеленных на ИИ, которые уже используются в бизнес-процессах компаний

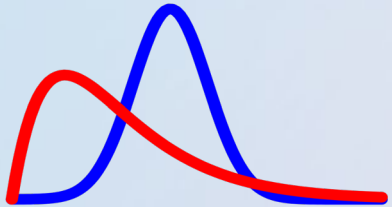
Развитие атак на конфиденциальные ИИ системы (федеративное обучение)

Типовые задачи ИИ
в Кибербезопасности

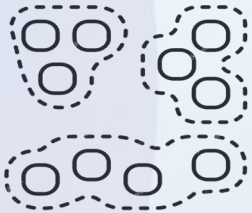
ИКНК



Классификация



Регрессия



Кластеризация

ИИ в задачах кибербезопасности:

- обнаружение атак
- обнаружение фишинга
- антиспам-системы
- антивирусы
- биометрическая аутентификация
- обнаружение аномалий, ботнетов
- DLP-системы

Северная Америка
крупнейший рынок (2019г)

Азиатско-Тихоокеанский регион
самый быстрорастущий рынок (2020-2030 гг)

Объем рынка ИИ-решений для кибербезопасности

Прогнозируемый объем рынка ИИ-решений для кибербезопасности

Что мы знаем об атаках на ИИ?

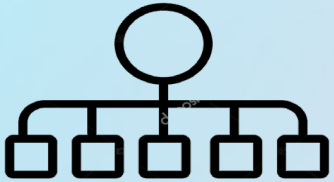
Модель угроз

ИКНК

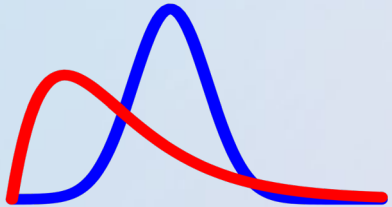
Свойство	Adversarial ML Threat Matrix (MITRE)
Объект исследования	Система машинного обучения, подверженная состязательным атакам
Соответствие нормативным требованиям	**Нет** (представляет собой реестр типовых атак, построенный по сводкам, полученным от 24 ИТ-компаний)
Полнота модели	**Нет** (построена на основе коллекции зафиксированных примеров реализации состязательных атак на системы ИИ)
Анализ природы угроз, способов их реализации	**Нет** (фиксация случаев атак, типизация по базовым классам)
Анализ возможных механизмов защиты	**Нет** (декларированы лишь сами угрозы, техники защиты требуют дополнительного изучения)
Расширяемость модели	**Нет** (фиксированная структура матрицы по этапам атак на ИИ)

Уязвимости типовых решений

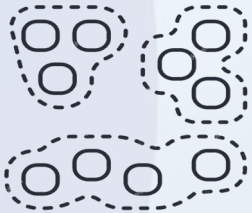
ИКНК



Классификация



Регрессия



Кластеризация

Через добавление дополнительно сгенерированных значений изменяется область распределения значений (линия предсказаний);
⇒ изменяется область допустимых значений параметра для определения класса
⇒ классификатор будет давать неверный ответ
системы классификации вирусов, вредоносной активности и т.п.

Hint: Злоумышленник может заранее произвести исследование по выявлению классов для создания дальнейших «отравленных» данных

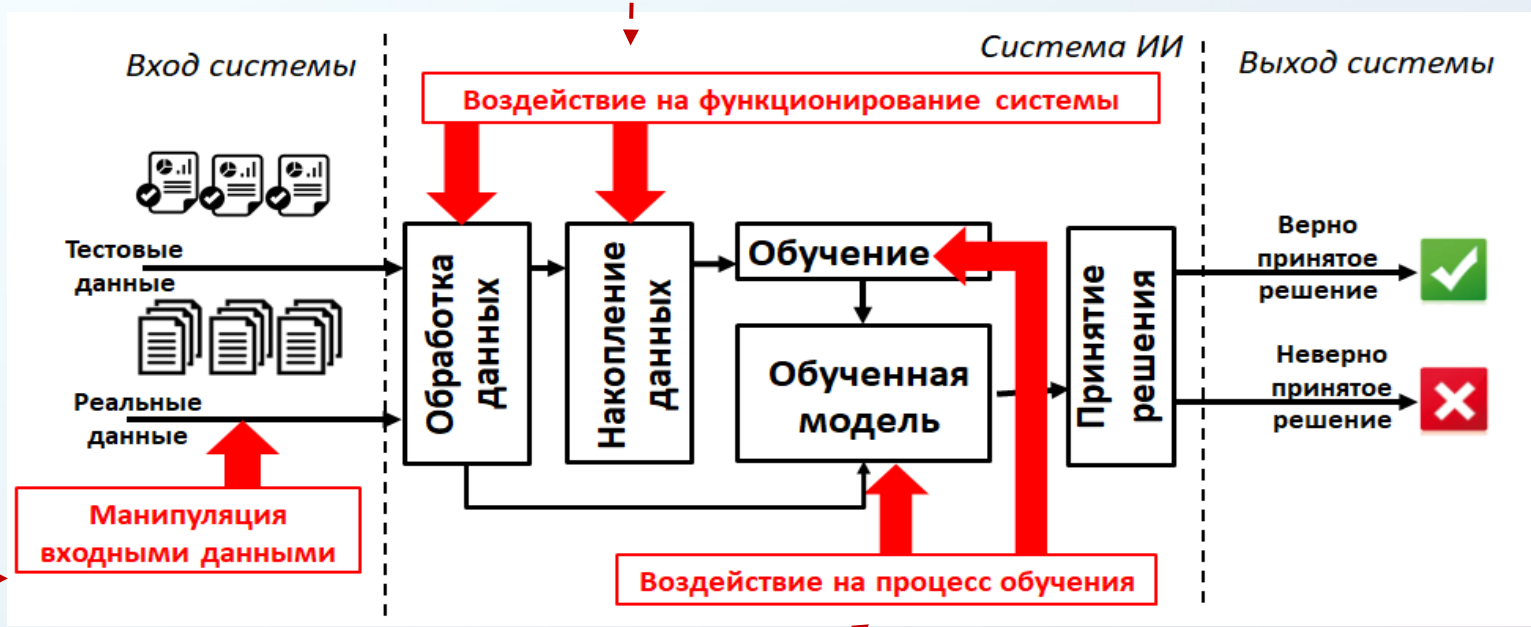
Hint: Если злоумышленнику известны параметры классификатора – процесс корректировки еще больше упрощается

Расширить границы кластеров – аналогично.

Также можно создать дополнительные кластеры путем создания группы объектов со схожими характеристиками.

Чувствительные компоненты систем ИИ

- **модули системы** – нарушение корректной работы системы, сбой в работе одного из ее компонентов или модулей



- **наборы данных** – воздействие на реальные входные данные

- **механизм обучения** (недостатки самого принципа «обучения модели») –как воздействие на сам процесс обучения модели, так и на уже обученную модель

ПОЛИТЕХ Санкт-Петербургский политехнический университет Петра Великого

Институт компьютерных наук и кибербезопасности

Как заставляют ИИ «изменять» себе

Локализация		Способ эксплуатации	Результат эксплуатации
Уязвимости входа модели	**Обучающие данные**, поступающие в вычислительную модель **Обученная модель** ИИ и МО **Обучающий модуль**	- Воздействие на обучающие данные **до того**, как они попадут в систему ИИ и МО (отравление данных) - Поиск «слепых зон» в обученной вычислительной модели, генерация данных, попадающих в эти слепые зоны - Генерация большого объема специальных входных данных, искажающих поле экспертизы системы ИИ и МО	Неправильное обучение вычислительной модели Наложение конкретного решения на вычислительную модель Изменение логики работы модуля обучения
	Реальные (входные) данные, поступающие в уже обученную вычислительную модель ИИ и МО	Реализация **состязательных примеров**: искажение	Вычислительная модель воспринимает искаженные данные как правильные данные определенного класса
Уязвимости модели	**Ввод данных** в вычислительную модель **Предварительная обработка** и обработка данных **Хранилище** данных	- Перехват и модификация ввода данных в модель ИИ - Поиск и эксплуатация уязвимостей в реализации обработки/обработки данных, изменение параметров модели (метод оптимизации, гранулярность, масштаб и т.д.) - Удаление фрагментов данных из хранилища	Сбои (отказы) системы ИИ и МО Некорректное поведение системы ИИ и МО

Изменение поведения модели

Классы актуальных угроз

направленных на снижение достоверности результатов

По характеру воздействия на модель

- угроза отравления
- угроза искажения
- исследовательские угрозы

По стратегии возможной реализации угрозы по отношению к модели

- угроза «белого ящика»
- угроза «серого ящика»
- угроза «черного ящика»

По способу воздействия

- через вмешательство в обучающую выборку
- через ошибки модели

По специфике воздействия

- нецеленаправленная
- целенаправленная

По нарушаемому аспекту безопасности

- нарушение целостности модели
- нарушение доступности модели
- нарушение конфиденциальности модели

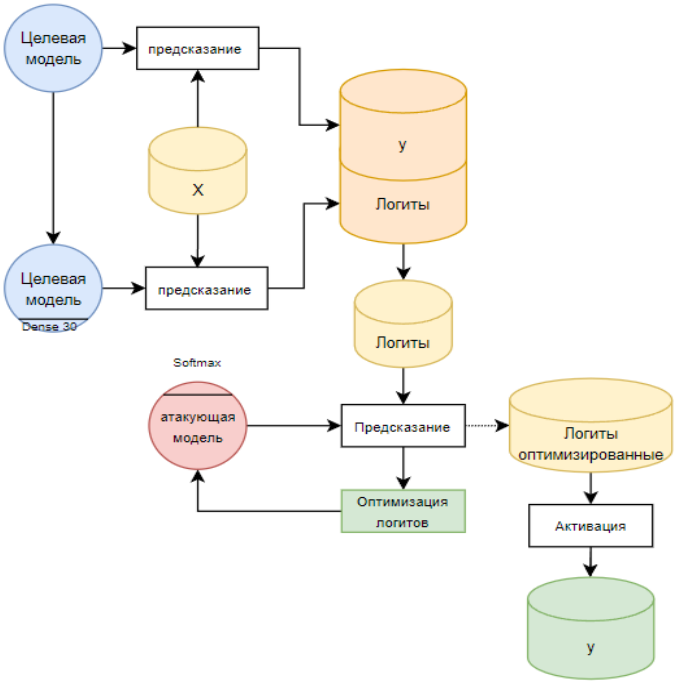
Модель угроз
в соответствии методикой ФСТЭК

Центр компетенций НТИ
Технологии доверенного взаимодействия

ПОЛИТЕХ
Санкт-Петербургский политехнический университет Петра Великого

Институт компьютерных наук и кибербезопасности

Конфиденциальность данных и модели



Метод «Умный шум»:

- гибкое задание количества шума, налагаемого **на выходе защищаемой модели:**
- градиентный шаг для оптимизации шума;
- делает невозможным для нарушителя *вывод о принадлежности отдельных образцов* набору данных, на котором обучалась атакуемая модель
  (атака «вопрос о вхождении»)

Метод на базе контроля расстояний в запросах

- выявляет поток запросов, поступающих от нарушителя при зондировании атакуемой им модели,
  по отклонению от нормального распределения данных, характерного для поведения обычных пользователей.
- обеспечивает выявление и затем блокировку нарушителя, исследующего атакуемую модель
- не воздействует на механизмы защищаемой модели

Результат

Центр компетенций НТИ
Технологии доверенного взаимодействия

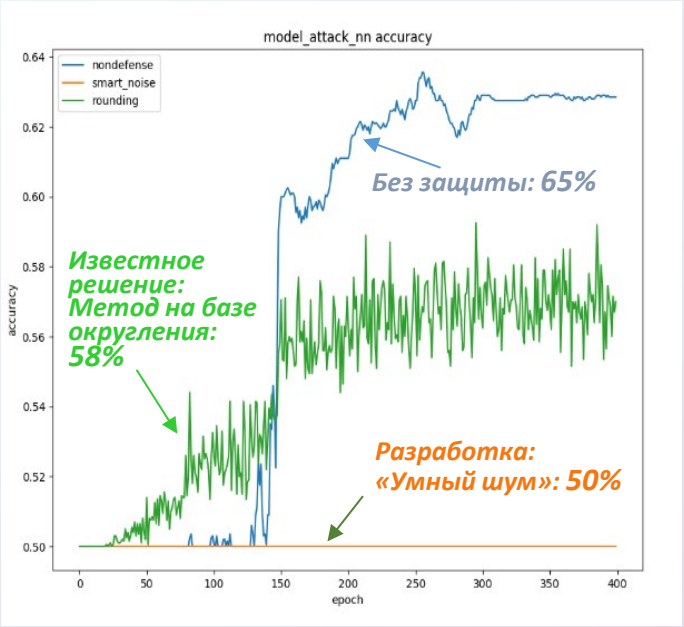
Сравнение с альтернативными решениями

Название	Воздействие на обучающие данные	Воздействие на параметры модели	Воздействие на архитектуру модели	Воздействие на функцию потерь	Воздействие на выходные данные	Влияние на точность модели	Сложность	Уровень безопасности
Отсев		+				Повышает	Низкая	Низкая
Ансамблевое обучение ML			+			Повышает	Средний	Средний
Ансамблевое обучение DL			+			Повышает	Высокая	Высокий
Усложнение архитектуры модели			+			-	Высокая	Высокий
Регуляризация L_1 и L_2				+		Повышает	Средний	Низкая
Эластичная сеть				+		Повышает	Средний	Высокий
Стохастическое округление					+	Снижает	Низкая	Средний
Округление вывода					+	Снижает	Низкая	Средний
Отсечение					+	Снижает	Низкая	Высокий
Округление вывода ClassLabels					+	Снижает	Низкая	Высокий
Умный шум					+	Сохраняет	Средний	Высокий

Оценка точности метода защиты на базе контроля расстояния в запросах

Набор данных	Частота ложных срабатываний (метрика FPR)	Количество запросов пользователя до обнаружения атаки	
		Атака на модели с совместным обучением	Атака с использованием GAN
MNIST	0,0	5 560	120
GTSRB	0,1	5 020	430

Оценка точности атакующей модели-классификатора нарушителя, который определяет принадлежность данных набору данных целевой модели



ПОЛИТЕХ
Санкт-Петербургский политехнический университет Петра Великого

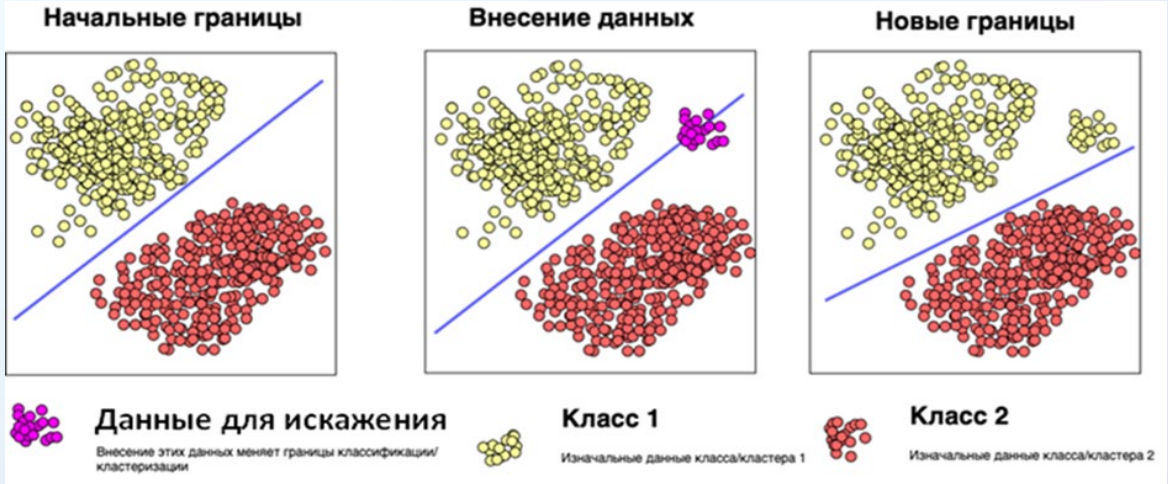
Институт компьютерных наук и кибербезопасности

Примеры угроз системам ИИ

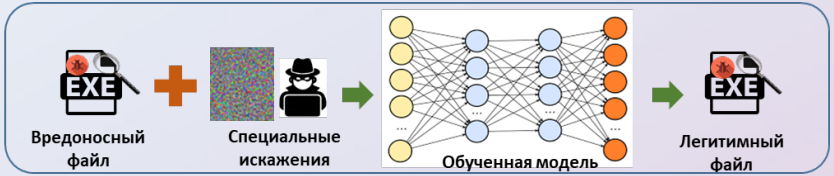
направленных на снижение достоверности результатов

Угроза отравления:

Угроза искажения:



Пример деструктивного воздействия искажений на классификатор компьютерных атак

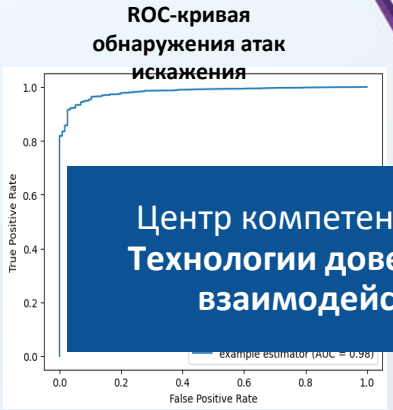


ПОЛИТЕХ
Санкт-Петербургский политехнический университет Петра Великого

Институт компьютерных наук и кибербезопасности

Обнаружение искажений

- ✓ незащищенная модель (точность 97%) при воздействии искажений демонстрирует падение точности до 49%. Точность обнаружения атак искажения – **98%**, точность защищенной модели при воздействии атак искажения – >**91%** (вместо 49%);
- ✓ метод защиты **не зависит от вычислительной модели** ИИ, безопасность которой он обеспечивает;

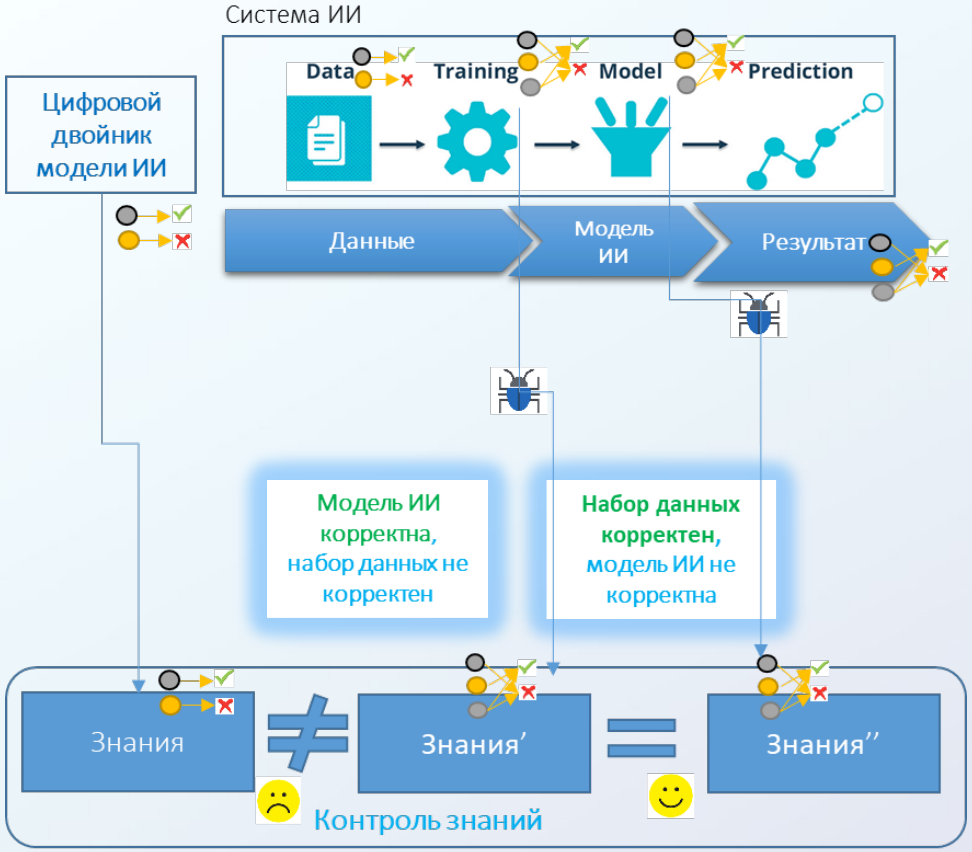


Центр компетенций НТИ
Технологии доверенного взаимодействия

КОНКУРИРУЮЩИЕ РЕШЕНИЯ	РЕСУРСОПОТРЕБЛЕНИЕ	ТОЧНОСТЬ ОБНАРУЖЕНИЯ / КЛАССИФИКАЦИИ
Обнаружение искажений с помощью метода k-ближайших соседей (kNN) с маркировкой состязательных образцов [1]	**Высокое** (обучение алгоритма, работа классификатора)	**97% / Не реализует**
Поддержание устойчивости модели с помощью генеративной нейросети GAN для генерации состязательных образцов [2]	**Высокое** (работа нейросети-генератора состязательных образцов)	**Не реализует / 70%**
Разработанный в рамках НТИ метод, **агрегирующий** обнаружение искажений на основе **анализа тестовых данных** и **защитную дистилляцию**	**Низкое** (нет необходимости формировать обучающий набор данных, обучать и переобучать нейросети)	**98% / >91%**

[1] Paudice A., Muñoz-González L., Lupu E. C. Label sanitization against label flipping poisoning attacks. // LNCS series (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).
[2] Samangouei P., Kabkab M., Chellappa R. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. // 6th International Conference on Learning Representations, Conference Track Proceedings.

Доверие к результату



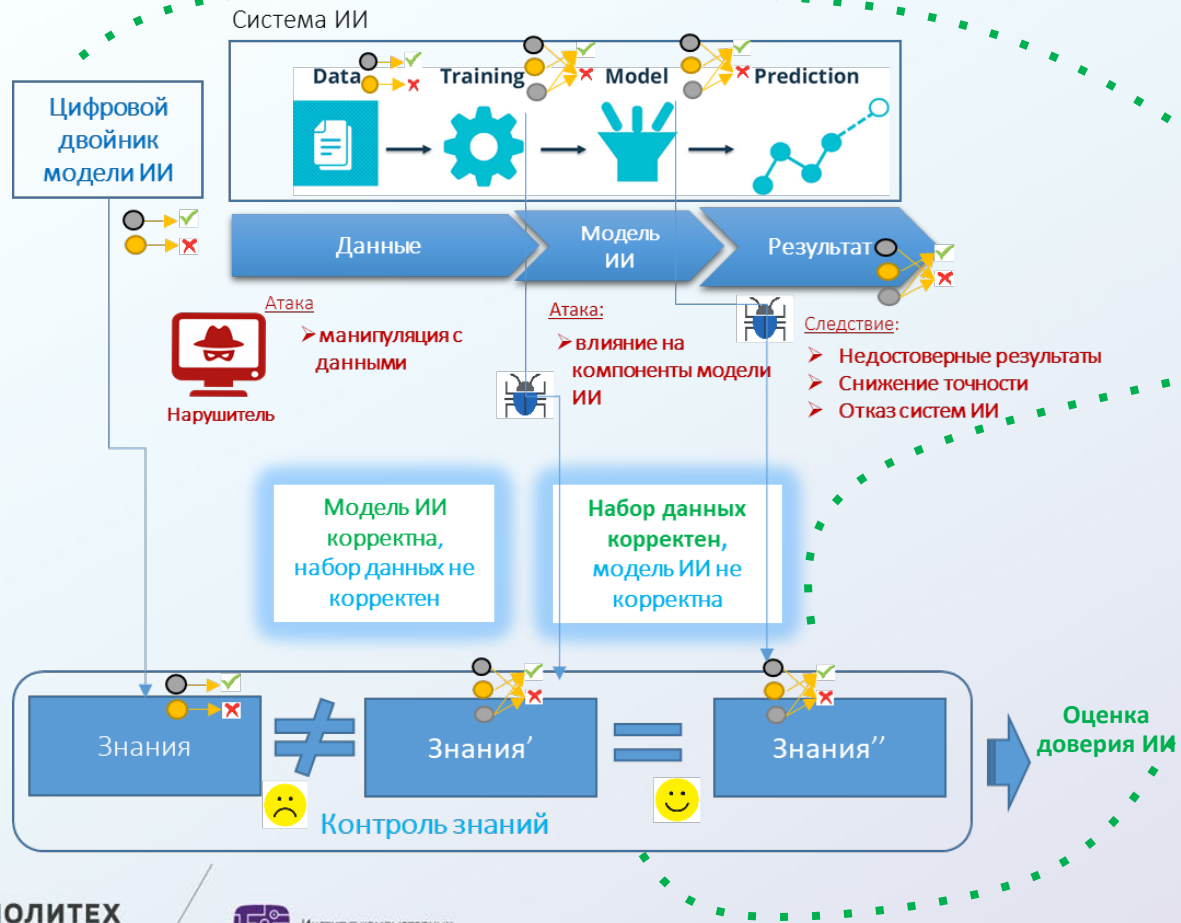
- Самотестирование
- Кросс-валидация

Гибридизация взаимодополняющих методов и алгоритмов

- Сокращение размера «враждебного пространства данных» для систем ИИ.
- Предотвращение использования мобильности (переноса) атак на вычислительные модели систем ИИ.
- Затруднение действий нарушителей ИИ.

- Контроль цепочек поставок данных в системах ИИ.
- Выявление искажающих данных в схемах обучения систем ИИ.
- Защитная дистилляция и состязательная тренировка вычислительной модели.
- Быстрая кросс-валидация решений, полученных системами ИИ и машинного обучения.
- Аудит и трассировка событий на уровне алгоритмов ИИ и МО, с помощью которых возможно быстро проверить состояние классификаторов, которое привело к недостоверному решению

Безопасность ИИ в киберсреде



(1) **Использование «двойника» ИИ:** мониторинг поведения исходной модели ИИ и контроль результатов по двойнику

(1) **«Песочница» ИИ:** проверка результатов на копии данных ДО внесения их в незараженный набор данных

(1) **Обучение с переносом:** обучение на выборках из другой области и сверка результатов с уязвимой моделью ИИ

ИКНК

ФГАОУ ВПО «СПБПУ»

ИНСТИТУТ КОМПЬЮТЕРНЫХ НАУК И КИБЕРБЕЗОПАСНОСТИ

Зегжда Дмитрий Петрович

чл.-корр. РАН, д.т.н., профессор

Полтавцева Мария Анатольевна

д.т.н., доцент

poltavtseva@ibks.spbstu.ru

Главный учебный корпус, к. 173

Политехническая ул., 29, Санкт-Петербург 195251

Тел: +7 (812) 552-76-32