УДК 004.056

**А.А. Воробьева**

# Способ исследования устойчивости систем со встроенным искусственным интеллектом, использующихся на промышленных объектах, к состязательным атакам

Представлен способ исследования устойчивости систем со встроенным искусственным интеллектом (ИИ), использующихся на промышленных объектах, к состязательным атакам. Исследовано влияние состязательных атак на показатели работы систем, использующих модели машинного обучения (МО). Представлена разработанная обобщенная схема и определены сценарии реализации атак на системы со встроенным ИИ, использующиеся на промышленных объектах. Сформирован комплексный набор показателей, используемых для исследования устойчивости моделей МО, включающий показатели качества набора тестовых данных (MDQ), показатели качества модели МО (MMQ), показатели устойчивости модели к состязательным атакам (MSQ). Способ основан на применении данного комплекса показателей и включает следующие шаги: формирование набора тестовых данных, содержащего чистые образцы; оценка качества набора тестовых данных с использованием показателей MMQ; определение актуальных методов реализации состязательных атак; генерация состязательных примеров и формирование набора тестовых данных для оценки устойчивости модели, содержащего сгенерированные состязательные образцы; оценка качества сформированного набора тестовых данных с использованием показателей MDQ; оценка качества модели МО с использованием показателей MMQ; оценка устойчивости модели с использованием показателей MSQ.

**Ключевые слова:** кибербезопасность, методы искусственного интеллекта, интеллектуальные производственные системы, состязательные атаки.
**DOI:** 10.21293/1818-0442-2023-26-4-44-52

Под системой, использующейся на распределенных промышленных объектах, понимается распределённая система ввода-вывода с децентрализованной обработкой данных [1]. Как правило, такая система включает множество различного рода оборудования, создающего инфраструктуру для реализации определенного алгоритма управления [2]. Подобные объекты могут включать множество различных датчиков, сигналы от которых передаются на систему управления.

В настоящее время промышленные системы развиваются в соответствии с концепцией «Индустрии 4.0», подразумевающей полную автоматизацию всех процессов, ключевая роль в которой отводится методам машинного обучения (МО) и искусственного интеллекта (ИИ) [3, 4]. В настоящее время многие прикладные задачи решаются при помощи искусственных нейронных сетей (ИНС), например: визуальный контроль качества продукции и ее учет, мониторинг качества работы персонала или автоматизированных линий, определение объектов при позиционировании оборудования и манипуляторов, обнаружение опасных зон и контроль соблюдения персоналом правил безопасности [5].

Также на промышленных объектах могут использоваться различные биометрические системы для разграничения доступа, детекторы объектов (людей, автомобилей и др.), также основанные на методах ИИ. При этом эксперты отмечают, что эти технологии создают серьезные проблемы, связанные с кибербезопасностью (англ. cybersecurity) и функциональной безопасностью (англ. safety). В реальных промышленных системах должен обеспечиваться необходимый уровень доверия и надежности используемых моделей и алгоритмов МО. На передний план выходят отказоустойчивость и функциональная безопасность подобных систем, так как технологические процессы требуют непрерывного выполнения. Важно гарантировать, что их использование не приведет к возникновению сбоев и ошибок, вызванных как внутренними проблемами, так и действиями злоумышленников [6].

**Анализ состязательных атак на системы со встроенным искусственным интеллектом**

Обобщенно представляется возможным разделить жизненный цикл применения алгоритмов МО на два этапа: этап подготовки (обучения) модели и этап эксплуатации (рис. 1).

Отметим, что на промышленных объектах может применяться не одна, а несколько моделей (A, B, …, N), предназначенных для решения разных задач. Система управления превращает ответы моделей МО на поступающие входные данные в реальные действия.

Данные могут поступать с видеокамер и звукозаписывающих устройств, разнообразных датчиков (температуры, давления, уровня, частиц и пр.). Так, модель «A» может анализировать видеопоток, модель «B» – анализировать данные различных датчиков. Система управления использует их для принятия решения по выполнению какого-либо действия (в том числе полной остановки или возобновления производства, остановки одного из узлов и пр.). По сути система использует модели МО для преобразования входных данных реального мира в решения, а затем в действия.

Системы со встроенным ИИ, использующиеся на промышленных объектах, также уязвимы перед атаками [7, 8], которые могут выполняться как на

этапе подготовки и обучения модели, так и на этапе ее эксплуатации [9].

Состязательная атака (англ. adversarial attack) – это обобщенное наименование атак на системы ИИ, в том числе способы обмана ИНС с целью изменения «ответа» системы на необходимый злоумышленнику и нарушения ее производительности. Данные атаки могут выполняться на системы распознавания образов (фото, видео, аудио) и реализуются с использованием состязательных примеров (англ. adversarial samples) – образцов данных (англ. data sample), в которые внесены незначительные искажения, приводящие к некорректному распознаванию [10]. Такими искажениями, в частности, могут служить добавление шума или изменение нескольких пикселей на изображении. Важным является тот факт, что искажения незаметны для человека.

На этапе подготовки и обучения модели могут выполняться различные действия с обучающими данными, начиная от атак на получение несанкционированного доступа (НСД) к данным, заканчивая различными манипуляциями – отравлением обучающих данных.

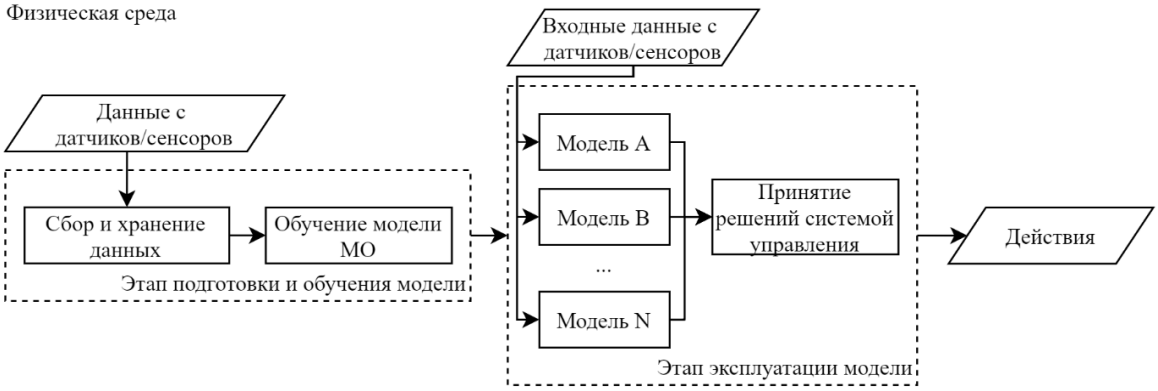


Рис. 1. Этапы применения алгоритмов МО на промышленных объектах

Атаки на доступ к данным имеют одну основную цель – кража набора данных для создания злоумышленником модели, которая будет использована для создания состязательных примеров для последующего выполнения атаки уклонения.

Отравление нацелено на смещение границы принятия решения и может выполняться как путем внедрения в набор новых вредоносных образцов, так и модификацией имеющихся данных (изменение значений признаков, изменение меток классов).

Атаки на этапе эксплуатации модели МО имеют две основные цели [11]:

• получение информации о модели или наборе обучающих данных (разведывательные атаки);

• поиск уязвимостей в обученной модели для нахождения образцов данных, на которых модель ошибается (атаки на обход модели МО).

В системах со встроенным ИИ защите подлежат:

1) данные (результаты измерений), из которых получены признаки для обучения;

2) алгоритмы получения признаков из результатов измерений;

3) алгоритмы обучения модели МО;

4) значения гиперпараметров модели МО;

5) значения параметров обученной модели МО;

6) доверительные вероятности принимаемых решений;

7) сами принимаемые решения;

8) граница принятия решений моделью (или гиперплоскость в $n$-мерном пространстве признаков).

**Этапы реализации атак на системы со встроенным ИИ, использующиеся на промышленных объектах**

В общем виде все сценарии выполнения атак на системы со встроенным ИИ, использующиеся на промышленных объектах, могут быть сведены к представленной на рис. 2 схеме.

Также выделяются два этапа выполнения атаки: подготовка и воздействие. Конкретные сценарии формируются путем пересечения техник этапа подготовки (ЭП) и этапа воздействия (ЭВ).

*ЭП.1–ЭП.2. Отравляющие атаки: внедрение данных в набор обучающих данных, модификация образцов в наборе обучающих данных*

Злоумышленник, имея доступ к обучающему набору данных, может осуществить его отравление путем изменения самих данных или меток классов. Это позволяет злоумышленнику встроить в модель МО уязвимость, которую достаточно сложно обнаружить. В стандартных условиях модель работает в соответствии с ожидаемым поведением, однако при наличии специального триггера во входных данных будет производить необходимый злоумышленнику результат.

Данная уязвимость может быть активирована путем передачи в модель МО образца данных, содержащего необходимый триггер. Примером может служить размещение специально подготовленного изображения в физической среде, где оно фиксируется камерой (см. ЭР.1, ЭР.3–ЭР.4).

*ЭП.3–ЭП.4. Атаки на доступ к данным или модели: кража набора обучающих данных, кража модели МО*

Злоумышленник, используя стандартные средства получения НСД, осуществляет кражу обучающего набора данных или модели.

С использованием набора данных он обучает собственную модель, повторяющую целевую модель, создает состязательные примеры и, используя свойство переносимости ИНС, осуществляет атаку целевой модели.

Если же злоумышленник имеет полный доступ к целевой модели, то он выполняет приведенные выше шаги, исключая этап обучения.

Далее выполняется атака по ЭР.1, ЭР.3–ЭР.4 путем передачи в модель вредоносного образца данных.

*ЭП.5–ЭП.6. Разведывательные атаки: восстановление модели МО, восстановление набора обучающих данных*

Разведывательные атаки могут как иметь целью кражу интеллектуальной собственности и получение конфиденциальной информации (ЭР.5–ЭР.6), так и являться вспомогательным этапом для реализации других атак (ЭР.1–ЭР.4).

Злоумышленник, используя различные виды санкционированного доступа (например, доступ по API) и ответы модели МО, осуществляет восстановление обучающего набора данных или модели.

Злоумышленник может восстановить данные, которые были использованы для обучения, направляя многократные запросы к модели и анализируя ответы (предсказанное значение и оценка уверенности в прогнозе). Корректируя подаваемые на вход данные и максимизируя уверенность в прогнозе, он имеет возможность восстановить обучающие данные или же убедиться, что конкретный образец содержался в обучающем наборе.

Также анализ ответов модели может быть использован для создания собственной модели, которая будет имитировать поведение целевой модели.

Атака выполняется по ЭР.1–ЭР.4 и по ЭР.5–ЭР.6.

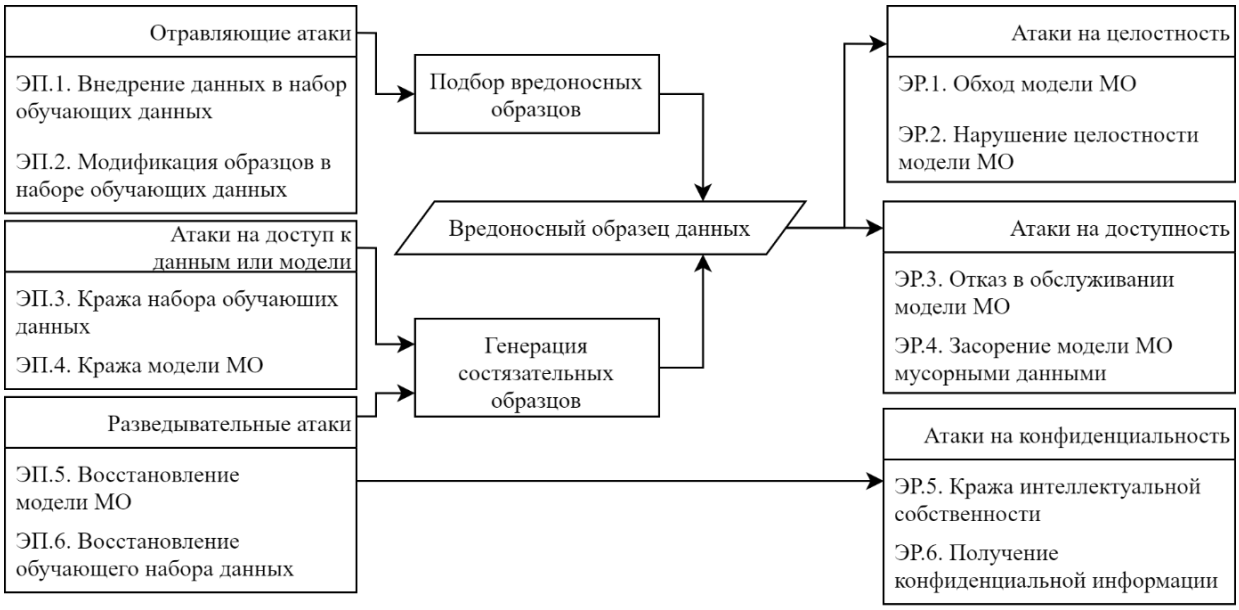


Рис. 2. Обобщенная схема реализации атак на системы со встроенным ИИ, использующиеся на промышленных объектах

*ЭР.1. Обход модели МО*

Злоумышленник создает состязательные образцы и, используя свойство переносимости ИНС, осуществляет атаку целевой модели путем передачи состязательного примера, например, изображения, размещая его в физической среде, где он фиксируется камерой.

*ЭР.2. Нарушение целостности модели МО*

Злоумышленники могут ухудшить качество работы целевой модели, если модель использует поступающие входные данные для дообучения. Ввод множества вредоносных данных постепенно изменит модель, нарушит ее целостность и снизит доверие к системе.

*ЭР.3–4. Отказ в обслуживании модели МО и засорение модели МО мусорными данными*

Злоумышленник генерирует поток запросов к модели МО с целью ухудшить, замедлить или остановить работу системы. Часто системы с МО требуют значительных вычислительных ресурсов, злоумышленник может создать такие входные данные, которые требуют больших объемов вычислений.

*ЭР.5–ЭР.6. Кража интеллектуальной собственности, получение конфиденциальной информации*

Злоумышленник осуществляет восстановление обучающего набора данных или модели, преследуя своей целью кражу интеллектуальной собственности или получение конфиденциальной информации.

**Сценарии реализации состязательных атак на системы со встроенным искусственным интеллектом, использующиеся на промышленных объектах**

Анализ литературы показал, что в настоящее время существует два основных способа применения состязательных атак к реальным промышленным объектам: использование состязательных заплаток (англ. adversarial patches) и состязательные атаки на виртуальные датчики (англ. soft sensors).

Состязательная заплатка – изображение меньшего размера (относительно объекта), которое создается с использованием состязательных атак и накладывается поверх объекта. Выделим три сценария использования состязательных заплаток для атак систем, применяющиеся на промышленных объектах.

*Сценарий 1.* Обман биометрических систем.

Состязательные заплатки, нанесенные на предметы одежды или медицинские маски, могут способствовать сокрытию субъекта или же неверной его идентификации. На рис. 3 приведен пример реализации такой атаки [12]. Точность идентификации субъекта без использования маски составляет 74,83%, с использованием обычной медицинской маски – 53,04%, с использованием маски с состязательным изображением – 5,17%.

Подобные атаки могут применяться для обмана систем биометрической идентификации разграничения доступа, мониторинга качества работы персонала, контроля соблюдения персоналом правил безопасности.

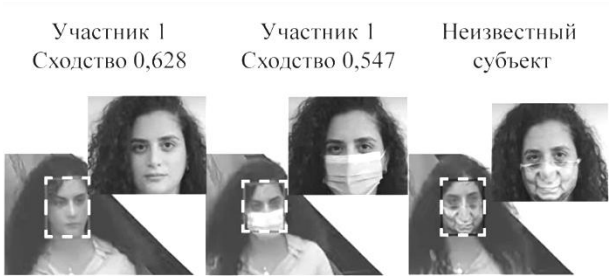


Рис. 3. Сценарий использования состязательных заплаток для обмана биометрических систем (неверная идентификация субъекта)

*Сценарий 2.* Обман систем детектирования и распознавания объектов.

Подобные подходы могут использоваться для обмана систем детектирования и распознавания объектов, таких как автомобили, упаковки с готовой продукцией. Рисунок 4 иллюстрирует нанесение состязательных заплаток на системы распознавания автомобилей, что позволяет скрыть присутствие данного объекта на территории [13, 14].

Данные атаки могут применяться для обмана систем визуального контроля качества продукции и ее учета, мониторинга качества работы автоматизированных линий, контроля опасных для человека зон, контроля пересечения объектами защищаемого периметра.


Рис. 4. Сценарий использования состязательных заплаток для обмана систем детектирования объектов (сокрытие объекта)

*Сценарий 3.* Обман систем позиционирования манипуляторов

Роботизированные производственные системы также используют системы распознавания образов, основанные на ИНС.

В работе [15] продемонстрирована атака на детектор объектов и позиционирования манипулятора, роботизированной руки. На карту нанесена состязательная заплатка, что создает некоторую оптическую иллюзию относительно позиции центра карты (рис. 5). Данная заплатка позволяет сместить предсказанное местоположение центра на область руки человека-оператора, что приводит к захвату ее манипулятором.

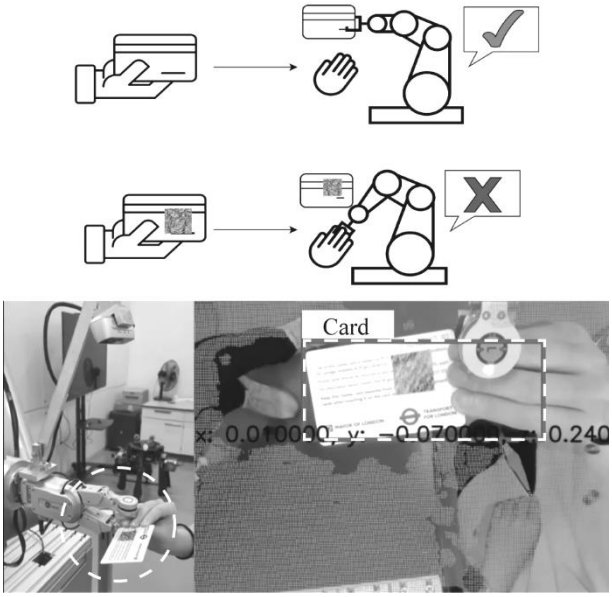


Рис. 5. Сценарий использования состязательных заплаток для обмана детектора объектов и позиционирования манипулятора (роботизированной руки)

*Сценарий 4.* Состязательные атаки на виртуальные датчики.

Виртуальные датчики представляют собой модели прогнозирования. На датчик в реальном времени поступают значения независимых переменных, на основании которых он прогнозирует зависимую переменную с использованием ИНС: в том числе глубоких нейронных сетей (DNN), автоэнкодеров, рекуррентных нейронных сетей (RNN), сетей с долгой краткосрочной памятью (LSTM), сверточных нейронных сетей (CNN).

Все из приведенных архитектур ИНС уязвимы к состязательным атакам. Созданные состязательные примеры могут привести к некорректным прогнозам этих моделей. При этом состязательные примеры создаются таким образом, чтобы выходные данные и прогнозы оставались похожими на допустимые.

В работе [16] продемонстрирована атака на виртуальные датчики, использующиеся в печи первичного риформинга (производство аммиака). Авторы проанализировали механизм, лежащий в основе работы виртуальных датчиков, и предложили два новых алгоритма для проведения атак: прямая атака на вывод (англ. directly attack output, DAO) и итеративная прямая атака на вывод (англ. iterative directly attack output, IDAO). В экспериментах была выполнена оценка коэффициента детерминации ($R^2$) модели до выполнения атак (методом быстрого градиента, базо-

вым итеративным методом, DAO и IDAO). До проведения атак $R^2$ составлял 0,821, а после в среднем по приведенным всем видам атак снизился до –4,184.

**Влияние состязательных атак на показатели работы систем со встроенным искусственным интеллектом**

Злоумышленник, реализующий состязательные атаки, может действовать так, чтобы получить необходимое ему поведение системы; нарушить корректность ее работы; сделать систему недоступной; получить конфиденциальную информацию о системе в целом, модели МО и/или обучающих данных.

В табл. 1 отражено влияние состязательных атак на основные показатели работы систем со встроенным ИИ.

Таблица 1

**Оценка показателей работы систем со встроенным ИИ, находящихся в условиях реализации состязательных атак**

| Вид атаки | Этапы использования МО | Показатель работы системы со встроенным ИИ | Влияние атаки на показатель |
|---|---|---|---|
| Атаки на конфиденциальность | Подготовка и обучение модели<br>Эксплуатация модели | Риск нарушения приватности | Повышение |
| | | Риск утечки конфиденциальных данных | Повышение |
| Атаки на целостность | Подготовка и обучение модели<br>Эксплуатация модели | Уверенность в прогнозе (confidence) | Снижение уверенности в прогнозе |
| | | Показатели качества работы модели МО | Снижение показателей качества |
| Атаки на доступность | Эксплуатация модели | Время получения ответа от системы | Повышение |

Уязвимости систем со встроенным ИИ позволяют злоумышленникам манипулировать целостностью систем машинного обучения (заставляя их совершать ошибки), конфиденциальностью (что приводит к утечке информации) и доступностью (нарушая или прекращая работу систем в целом или моделей).

**Разработка способа исследования устойчивости систем со встроенным искусственным интеллектом к состязательным атакам**

Под устойчивостью принято понимать свойство системы функционировать с заданными качественными показателями, находясь в условиях реализации атак [17]. С формальной точки зрения устойчивость к состязательным атакам определяется ее нечувствительностью к изменениям во входных данных. Устойчивость модели ($F$) оценивается на основе набора данных, включающего сгенерированные состязательные примеры. При этом качество полученного набора играет решающую роль в оценке показателей устойчивости.

Модель $F$: $X \to Y$ представляет собой отображение входного пространства $X$ в выходное пространство меток классов $Y$. Входное пространство $X = \{x_1, \ldots, x_n\}$ содержит $n$ образцов данных, а выходное пространство $Y = \{y_1, \ldots, y_m\}$ содержит $m$ возможных предсказаний метки класса. Модель $F$ построена таким образом, что для образца данных $x_i$ способна отнести его к истинному классу $y_j$.

Множество состязательных образцов $X' = \{x'_1, \ldots, x'_n\}$ образуется путем добавления искажений $p$ к исходным образцам $X$. $\Omega(x_i)$ представляет множество всех возможных измененных образцов $x_i$.

Модель $F$ является устойчивой к состязательным атакам, если никакое искажение $p$, внесенное в $x_i$, не может изменить выходные данные, т.е. $x'_i \in \Omega(x_i) \Rightarrow F(x_i) = y_j$.

*Показатели устойчивости систем со встроенным ИИ к состязательным атакам*

Существует ряд показателей, позволяющих оценить качество разработанной модели МО. Показатели качества набора тестовых данных играют решающую роль в оценке устойчивости.

Представляется возможным сгруппировать данные показатели в три основные категории, характеризующие:

- качество набора тестовых данных (MDQ);
- качество модели МО (MMQ);
- устойчивость модели к состязательным атакам (MSQ).

Выделим ряд показателей, характеризующих качество набора тестовых данных:

- покрытие нейронов [18];
- незаметность внесенных изменений [19].

Показатели устойчивости модели МО приведены в табл. 2.

На рис. 6 проиллюстрирована кривая устойчивости модели к состязательным образцам, она демонстрирует взаимосвязь между точностью работы на состязательных примерах и уровнем внесенных искажений. В данном примере под внесенными искажениями понимается добавление цифрового шума на изображения.

Плавная кривая устойчивости означает, что модель МО стабильна и последовательна (модель 1 на иллюстрации), а крутая – показывает, что она чувствительна и нестабильна (модель 2 на иллюстрации). Значение радиуса устойчивости (отмечен вертикальной штриховой линией) для модели 1 и 2 в данном примере составляет 0,25 и 0,03 соответственно. Радиус устойчивости вычисляется, исходя из имеющихся требований к значению показателя точности работы модели, и отражает максимальное количество искажений, которые могут быть корректно обработаны моделью.

Высокая точность $A_{\text{clear}}$ и большой радиус устойчивости указывают на то, что модель устойчива к состязательным атакам, тогда как низкая точность $A_{\text{clear}}$ и небольшой радиус устойчивости предполагают, что она уязвима для них.

Таблица 2

**Показатели качества модели МО и устойчивости модели к состязательным атакам**

| | | |
|---|---|---|
| Характеризующие качество модели МО | | |
| $A_{clear}$ | Точность (accuracy) работы модели на чистых данных | Доля чистых образцов данных $X$, отнесенных моделью к истинному классу относительно общего числа образов $n$ |
| $Perf_{clear}$ | Производительность модели на чистых данных | Количество чистых образцов данных, обработанных моделью в единицу времени |
| Характеризующие устойчивость модели МО к состязательным атакам | | |
| $A_{adv}$ | Точность работы модели на состязательных данных | Доля состязательных образцов данных $X'$, отнесенных моделью к истинному классу относительно общего числа образов |
| $R$ | Радиус устойчивости | Оценка максимального количество искажений $p$, которые могут быть внесены в образец данных $x_i$, так чтобы $F(x'_i) = y_j$ [20] |
| $S$-curve | Кривая устойчивости | График, позволяющий оценить зависимость точности $A_{adv}$ от $p$ (рис. 7) |
| $AVF_{conf}$ | Средняя уверенность (confidence) модели в прогнозе относительно ложного класса | Среднее арифметическое вероятностной оценки уверенности в прогнозах относительно ложного класса по всем состязательным примерам, которые были успешны в обходе модели |
| $AVT_{conf}$ | Средняя уверенность модели в прогнозе относительно истинного класса | Среднее арифметическое вероятностной оценки уверенности в прогнозах относительно истинного класса по всем состязательным примерам, которые были успешны в обходе модели |
| $NTR$ | Толерантность к шуму | Оценка распределения вероятностей над множеством классов, среднее арифметическое разниц между вероятностью наиболее подходящего ложного класса и максимальной вероятностью других классов [21] |
| $Perf_{adv}$ | Производительность модели на состязательных данных | Количество состязательных образцов данных, обработанных моделью в единицу времени |

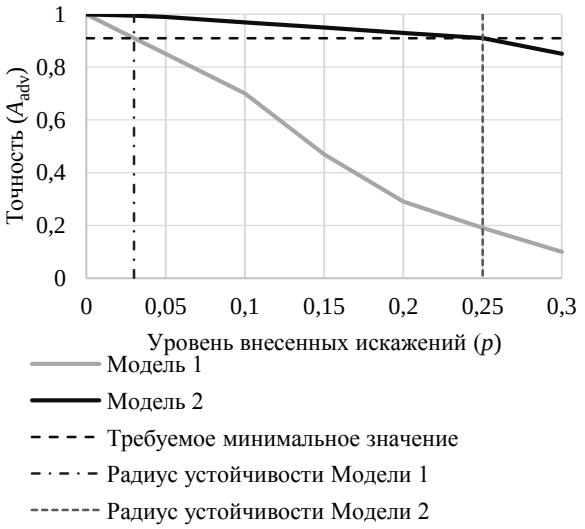


Рис. 6. Иллюстрация кривой и радиуса устойчивости модели к состязательным образцам

*Способ исследования устойчивости систем со встроенным искусственным интеллектом, использующихся на промышленных объектах, к состязательным атакам*

Оценка устойчивости систем со встроенным искусственным интеллектом к состязательным атакам сводится к оценке применяемой модели МО, что схематично отражено на рис. 7.

Приведем последовательность выполнения оценки устойчивости систем со встроенным искусственным интеллектом к состязательным атакам:

1. Формирование набора тестовых данных, содержащего чистые образцы ($Data_{clear}$).

2. Оценка качества набора тестовых данных $Data_{clear}$ с использованием показателей MDQ.

3. Определение актуальных методов реализации состязательных атак.

4. Генерация состязательных примеров на основании выделенных в п. 2 методов с использованием программных инструментов и библиотек (Adversarial Robustness Toolbox, Robustness Gym, Cleverhans, Alibi Detect).

5. Формирование набора тестовых данных для оценки устойчивости модели, содержащих сгенерированные состязательные образцы ($Data_{adv}$).

6. Оценка качества набора тестовых данных $Data_{adv}$ с использованием показателей MDQ.

7. Оценка качества модели МО с использованием $Data_{clear}$ и показателей MMQ.

8. Оценка устойчивости модели с использованием $Data_{adv}$ и показателей MSQ.

Дополнительно может выполняться оценка устойчивости модели к различного рода искажениям (различные виды шума, туман, снег, изменения яркости и контрастности, оптические искажения и пр.) [22].

Также после внедрения мер противодействия состязательным атакам рекомендуется выполнить оценку их качества (включая, но не ограничиваясь разницей Aclear до и после применённых мер).

**Заключение**

В статье представлен способ исследования устойчивости систем со встроенным ИИ к состязательным атакам. Установлено влияние состязательных атак на показатели конфиденциальности, целостности и доступности систем. На основе анализа литературных источников выделены показатели, характеризующие устойчивость систем перед состязательными атаками, в том числе точность работы модели на состязательных данных, радиус устойчивости, кривая устойчивости, средняя уверенность модели в прогнозе относительно ложного класса, оценка толерантности к шуму, производительность модели на состязательных данных. Способ основан на применении комплекса показателей, включающего показа-

тели качества набора тестовых данных, показатели качества модели МО, показатели устойчивости модели к состязательным атакам. Данный способ предназначен для специалистов по кибербезопасности, а также разработчиков программных систем со встроенным искусственным интеллектом. Способ позволяет оценить устойчивость систем со встроенным ИИ (в том числе применяющихся на промышленных объектах) к состязательным атакам.

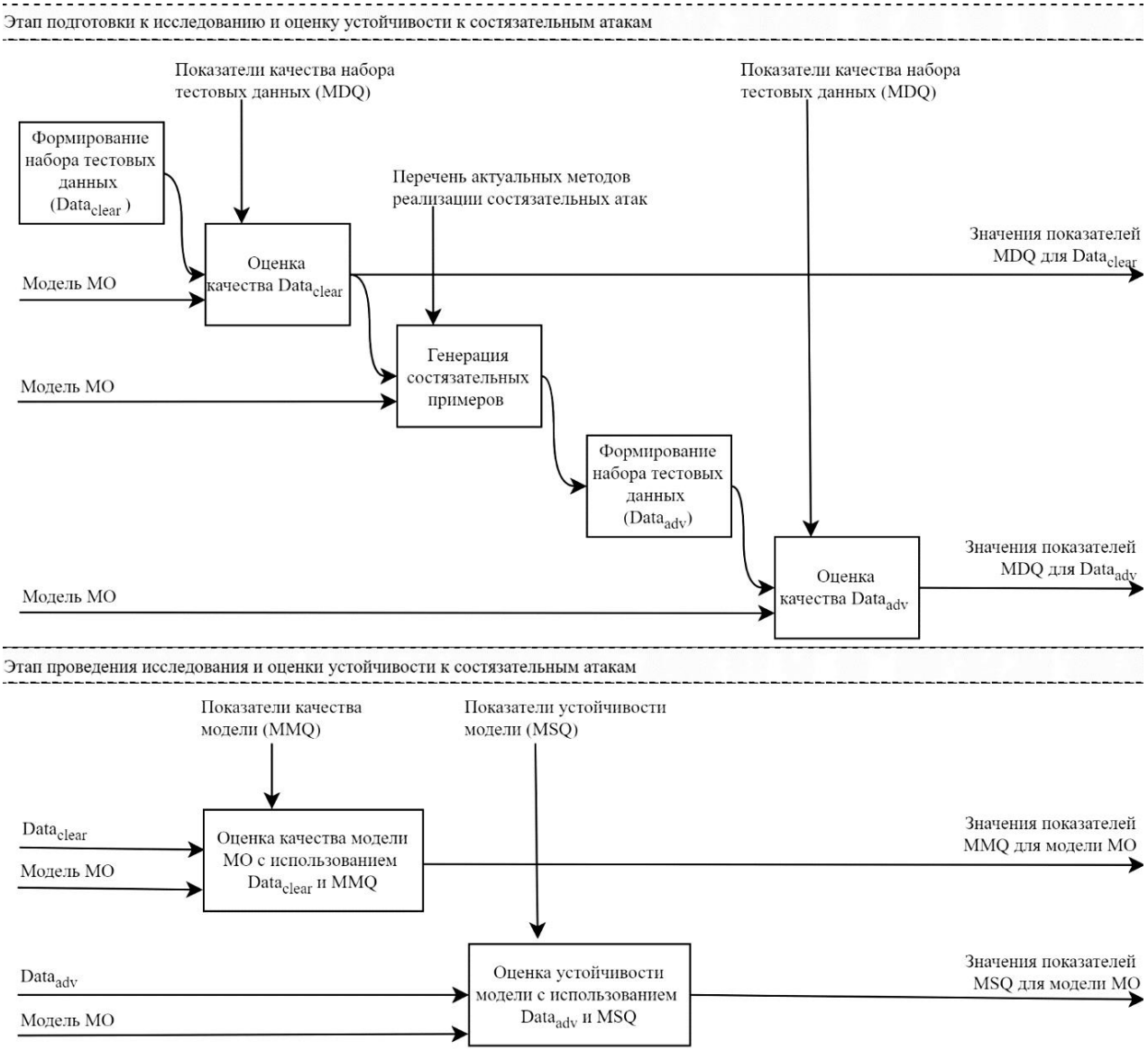


Рис. 7. Схема исследования и оценки устойчивости систем со встроенным искусственным интеллектом к состязательным атакам

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации, № 2019-0898.

*Литература*

1. Обзор современного рынка распределенных систем управления в нефтяной и газовой промышленности / А.В. Окружнов, Р.Р. Хайбунасов, И.Р. Хасанов, М.М. Андреева // Вестник Казанского технологического ун-та. – 2015. – Т. 18, № 2. – С. 383–389.

2. Коекин В.А. Алгоритмы управления на территориально распределенных промышленных и бытовых объектах // Электротехнические и информационные комплексы и системы. – 2008. – № 51. – С. 59–65.

3. Воробьева А.А. Методы интеллектуального анализа данных и обработки естественного языка в управлении роботизированными производственными системами / А.А. Воробьева, М.Ю. Федосенко // Доклады ТУСУР. – 2023. – Т. 26, № 3. – С. 65–71.

4. Smart Factory Monitoring [Электронный ресурс]. – Режим доступа: https://www.procemex.com/smart-factory/, свободный (дата обращения: 05.12.2023).

5. AIoT для умных фабрик [Электронный ресурс]. – Режим доступа: https://www.cta.ru/articles/obzory/vstraivaemye-sistemy/165894/, свободный (дата обращения: 05.12.2023).

6. Adversarial/Robust AI Report development methodology [Электронный ресурс]. – Режим доступа: https://ec.eu-

ropa.eu/research/participants/documents/downloadPublic?documentIds=080166e5e1fdef06&appId=PPGMS, свободный (дата обращения: 05.12.2023).

7. A survey on adversarial attack in the age of artificial intelligence / Z. Kong, J. Xue, Y. Wang, L. Huang, Z. Niu, F. Li // Wireless Communications and Mobile Computing. – 2021. – Vol. 2021. – P. 1–22.

8. A review on ai for smart manufacturing: Deep learning challenges and solutions / J. Xu, M. Kovatsch, D. Mattern, F. Mazza, M. Harasic, A. Paschke, S. Lucia // Applied Sciences. – 2022. – Vol. 12, No. 16. – P. 8239.

9. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations [Электронный ресурс]. – Режим доступа: https://csrc.nist.gov/pubs/ai/100/2/e2023/ipd, свободный (дата обращения: 05.12.2023).

10. Goodfellow I.J. Explaining and harnessing adversarial examples / I.J. Goodfellow, J. Shlens, C. Szegedy [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/1412.6572, свободный (дата обращения: 05.12.2023).

11. Tuptuk N. Security of smart manufacturing systems / N. Tuptuk, S. Hailes // Journal of manufacturing systems. – 2018. – Vol. 47. – P. 93–106.

12. Adversarial Mask: Real-World Universal Adversarial Attack on Face Recognition Models / A. Zolfi, S. Avidan, Y. Elovici, A. Shabtai // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. – 2022. – P. 304–320.

13. Adversarial patch attack on multi-scale object detection for UAV remote sensing images / Y. Zhang, J. Qi, K. Bin, H. Wen, X. Tong // Remote Sensing. – 2022. – Vol. 14, No. 21. – P. 5298.

14. {CAPatch}: Physical Adversarial Patch against Image Captioning Systems / S. Zhang, Y. Cheng, W. Zhu, X. Ji, W. Xu // 32nd USENIX Security Symposium (USENIX Security 23). – 2023. – P. 679–696.

15. Physical Adversarial Attack on a Robotic Arm / Y. Jia, C.M. Poskitt, J. Sun, S. Chattopadhyay // IEEE Robotics and Automation Letters. – 2022. – Vol. 7, No. 4. – P. 9334–9341.

16. Kong X. Adversarial attacks on neural-network-based soft sensors: Directly attack output / X. Kong, Z. Ge // IEEE Transactions on Industrial Informatics. – 2021. – Vol. 18, No. 4. – P. 2443–2451.

17. Дорф Р. Современные системы управления / Р. Дорф, Р. Бишоп. – М.: Лаборатория базовых знаний, 2002. – 832 с.

18. Deepgauge: Multi-granularity testing criteria for deep learning systems / L. Ma, F. Juefei-Xu, F. Zhang, J. Sun, M. Xue, B. Li, C. Chen // Proceedings of the 33rd ACM/IEEE international conference on automated software engineering. – 2018. – P. 120–131.

19. A comprehensive evaluation framework for deep model robustness / J. Guo, W. Bao, J. Wang, Y. Ma, X. Gao, G. Xiao, A. Liu // Pattern Recognition. – 2023. – Vol. 137. – P. 109308.

20. Quantifying Robustness to Adversarial Word Substitutions / Y. Yang, P. Huang, J. Cao, F. Ma, J. Zhang, J. Li // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. – 2023. – P. 95–112.

21. Towards imperceptible and robust adversarial example attacks against neural networks / B. Luo, Y. Liu, L. Wei, Q. Xu // Proceedings of the AAAI Conference on Artificial Intelligence. – 2018. – Vol. 32, No. 1. – P. 1–8.

22. Hendrycks D. Benchmarking neural network robustness to common corruptions and perturbations / D. Hendrycks, T. Dietterich [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/1903.12261, свободный (дата обращения: 05.12.2023).

**Воробьева Алиса Андреевна**
Канд. техн. наук, доцент ф-та безопасности информационных технологий (ФБИТ) Национального исследовательского университета ИТМО (Университет ИТМО)
Кронверкский пр-т, 49, А,
г. Санкт-Петербург, Россия, 197101
ORCID: 0000-0001-6691-6167
Тел.: +7-921-947-21-14
Эл. почта: vorobeva@itmo.ru

Vorobeva A.A.

**Method for evaluating the industrial systems with built-in artificial intelligence robustness to adversarial attacks**

The paper presents a method for evaluating the industrial systems with built-in artificial intelligence (AI) robustness to adversarial attacks. The influence of adversarial attacks on the systems performance has been studied. The scheme and the scenarios to implement attacks on industrial systems with built-in AI were presented. A comprehensive set of metrics used to study the robustness of ML models has been proposed, including test data set quality metrics (MDQ), ML model quality metrics (MMQ), and model robustness to adversarial attacks metrics (MSQ). The method is based on the use of this metrics set and includes the following steps: generating a set of test data containing clean samples; assessing the quality of a test data set using MMQ metrics; identification of relevant adversarial attacks methods; generating adversarial examples and a test data set, containing the adversarial samples, to evaluate the robustness of the ML model; assessing the quality of the generated adversarial test data set using MDQ indicators; evaluating the quality of a ML model using MMQ indicators; evaluating model robustness using MSQ scores.

**Keywords:** cybersecurity, artificial intelligence methods, intelligent production systems, adversarial attacks.
**DOI:** 10.21293/1818-0442-2023-26-4-44-52

*References*

1. Okruzhnov A.V, Khaibunasov R.R., Khasanov I.R., Andreeva M.M. [Overview of the modern market for distributed control systems in the oil and gas industry]. *Bulletin of the Technological University*, 2015, vol. 18, no. 1, pp. 383–389 (in Russ.).

2. Koekin V.A. Control algorithms for geographically distributed industrial and domestic facilities. *Electrical Engineering and Information Complexes and Systems*, 2008, no. S1, pp. 59–65 (in Russ.).

3. Vorobeva A.A., Fedosenko M.Yu. Methods for data mining and natural language processing in the management of robotic production systems. *Proceedings of TUSUR University*, 2023, vol. 26, no. 3, pp. 65-71.

4. *Smart Factory Monitoring*. Available at: https://www.procemex.com/smart-factory/, free (Accessed: December 02, 2023).

5. *AIoT for smart factories*. Available at: https://www.cta.ru/articles/obzory/vstraivaemye-sistemy/165894/, free (Accessed: December 02, 2023).

6. *Adversarial/Robust AI Report development methodology*. Available at: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5e1fdef06&appId=PPGMS, free (Accessed: December 02, 2023).

7. Kong Z., Xue J., Wang Y., Huang L., Niu Z., Li F. A survey on adversarial attack in the age of artificial intelligence, *Wireless Communications and Mobile Computing*, 2021, vol. 2021, pp. 1–22.

8. Xu J., Kovatsch M., Mattern D., Mazza F., Harasic M., Paschke A., Lucia S. A review on ai for smart manufacturing: Deep learning challenges and solutions, *Applied Sciences*, 2022, vol. 12, no. 16, pp. 8239.

9. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. Available at: https://csrc.nist.gov/pubs/ai/100/2/e2023/ipd, free (Accessed: December 02, 2023).

10. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. Available at: https://arxiv.org/abs/1412.6572, free (Accessed: December 02, 2023).

11. Tuptuk N., Hailes S. Security of smart manufacturing systems, *Journal of Manufacturing Systems*, 2018, vol. 47, pp. 93–106.

12. Zolfi A., Avidan S., Elovici Y., Shabtai A. Adversarial Mask: Real-World Universal Adversarial Attack on Face Recognition Model, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2022, pp. 304–320.

13. Zhang Y., Zhang Yi., Qi J., Bin K., Wen H., Tong X. Adversarial patch attack on multi-scale object detection for UAV remote sensing images, *Remote Sensing*, 2022, vol. 14, no. 21, pp. 5298.

14. Zhang S., Cheng Y., Zhu W., Ji X., Xu W. {CAPatch}: Physical Adversarial Patch against Image Captioning Systems, *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 679–696.

15. Jia Y., Poskitt C.M., Sun J., Chattopadhyay S. Physical Adversarial Attack on a Robotic Arm, *IEEE Robotics and Automation Letters*, 2022, vol. 7, no. 4, pp. 9334–9341.

16. Kong X., Ge Z. Adversarial attacks on neural-network-based soft sensors: Directly attack output, *IEEE Transactions on Industrial Informatics*, 2021, vol. 18, no. 4, pp. 2443–2451.

17. Dorf R., Bishop R. *Sovremennye sistemy upravleniya* [Modern control systems]. Moskva, Laboratoriya Bazovyh Znanij, 2002, 832 p. (in Russ.).

18. Ma L., Juefei-Xu F., Zhang F., Sun J., Xue M., Li B., Chen C. Deepgauge: Multi-granularity testing criteria for deep learning systems, *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, 2018, pp. 120–131.

19. Guo J., Bao W., Wang J., Ma Y., Gao X., Xiao G., Liu A. A comprehensive evaluation framework for deep model robustness, *Pattern Recognition*, 2023, vol. 137, pp. 109308.

20. Yang Y., Huang P., Cao J., Ma F., Zhang J., Li J. Quantifying Robustness to Adversarial Word Substitutions, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2023, pp. 95–112.

21. Luo B., Liu Y., Wei L., Xu Q. Towards imperceptible and robust adversarial example attacks against neural networks, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32, no. 1, pp 1–8.

22. Hendrycks D., Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. Available at: https://arxiv.org/abs/1903.12261, free (Accessed: December 02, 2023).

**Alisa A. Vorobeva**
Candidate of Sciences in Engineering, Associate professor,
Faculty of Secure Information Technologies,
ITMO University
49, Kronverksky pr., bldg. A, St. Petersburg, Russia, 197101
ORCID: 0000-0001-6691-6167
Phone: +7-921-947-21-14
Email: vorobeva@itmo.ru