**Структура статьи**

**Введение**

Описать значение ИИ в различных секторах + краткая проблематика относительно безопасности.

Ключевые термины (риски безопасности ИИ и защитные механизмы и тд).

**1. Классификация угроз ИИ**

Описать различные типы угроз для систем ИИ (состязательные атаки, отравление данных, инверсию моделей и др).

Примеры для иллюстрации каждой угрозы.(?нужно ли?)

**2. Механизмы защиты**

Расписать существующие стратегии защиты от каждой выявленных угроз. Анализ эффективности, преимущества и ограничения этих механизмов защиты.

**3. Текущие проблемы и пробелы**

Описать области, в которых отсутствуют текущие методы защиты.

Описать проблемы внедрения надежных мер безопасности ИИ.

**4. Предлагаемые улучшения (научная новизна)**

Описать возможные инновационные подходы или улучшения существующих механизмов защиты.

Определить новые тенденции в области безопасности ИИ и потенциальные будущие угрозы.???????????

**Заключение**

**Источники литературы**

**Методы защиты искусственного интеллекта: Комплексный обзор, анализ и классификация**

**Лебин Максим, РТУ МИРЭА, Россия**

**Аннотация**

В эпоху стремительного технологического прогресса искусственный интеллект (ИИ) стал неотъемлемой частью различных отраслей, повышая эффективность и открывая новые возможности. Однако распространение технологий ИИ также привело к появлению целого ряда угроз безопасности, направленных на системы ИИ. В данной статье представлен обширный обзор существующих методов защиты ИИ, дана подробная классификация и анализ сильных и слабых сторон каждого подхода. Кроме того, определены области, требующие дальнейших исследований для укрепления безопасности ИИ.

Искусственный интеллект (ИИ) преобразует отрасли, повышая уровень автоматизации, принятия решений и предиктивного анализа. От диагностики в здравоохранении до финансового трейдинга и автономных транспортных средств - системы на базе ИИ революционизируют способы обработки и использования данных. Однако по мере интеграции ИИ в критически важные инфраструктуры его уязвимости в области безопасности представляют собой серьезную угрозу. Злоумышленники используют слабые места в моделях ИИ, что приводит к таким серьезным последствиям, как утечка данных, кампании по дезинформации и манипулирование системами принятия решений.

**ВВЕДЕНИЕ**

Искусственный интеллект произвел революцию во многих отраслях, от здравоохранения до финансов, автоматизируя процессы и предоставляя интеллектуальные знания. По мере распространения систем искусственного интеллекта они также становятся привлекательными объектами для злоумышленников, стремящихся использовать уязвимости в различных неблаговидных целях. Обеспечение безопасности и целостности систем ИИ имеет первостепенное значение для сохранения доверия и надежности этих технологий. Наиболее важные темы, которые, на мой взгляд, требуют внимания:

* **Растущее значение безопасности искусственного интеллекта.** Искусственный интеллект (ИИ) произвел революцию во многих отраслях, включая здравоохранение, финансы, транспорт и кибербезопасность. По мере того как системы ИИ становятся все более неотъемлемой частью процесса принятия решений и автоматизации, их безопасность и надежность становятся важнейшей проблемой. Потенциальные последствия уязвимости систем ИИ варьируются от незначительных сбоев до катастрофических отказов в таких областях, как автономные транспортные средства, финансовые операции и национальная безопасность.

Растущая зависимость от ИИ сделала его выгодной мишенью для киберпреступников и враждебных структур. Атаки на модели ИИ могут манипулировать результатами, извлекать конфиденциальную информацию или делать их полностью неэффективными. В результате возникает острая необходимость в разработке и внедрении эффективных методов защиты ИИ, чтобы обеспечить безопасность и надежность этих систем.

* **Ландшафт угроз: ИИ под ударом.** Системы искусственного интеллекта подвержены различным угрозам безопасности, использующим их слабые места. Эти угрозы можно разделить на атаки противника, отравление данных, инверсию моделей и кибератаки с использованием ИИ:
* **Состязательные атаки: Эти атаки включают в себя создание небольших, незаметных возмущений во входных данных, которые заставляют модели ИИ делать неверные прогнозы. Широко изученным примером является неправильная классификация изображений глубокими нейронными сетями из-за тщательно разработанных враждебных возмущений (Adversarial Attack and Defense: A Survey, MDPI).**
* **Отравление данных: Злоумышленники манипулируют обучающими данными, чтобы внедрить уязвимости в модели ИИ. Это может быть особенно опасно в объединенных средах обучения, где данные предоставляют несколько организаций (AI Security Risks and Threats, Checkpoint**).
* **Атаки с инверсией модели: Запрашивая обученную модель ИИ, злоумышленники могут получить конфиденциальные данные обучения, что угрожает конфиденциальности и безопасности. Это серьезный риск в таких приложениях, как распознавание лиц**.
* **Кибератаки с использованием ИИ: ИИ сам по себе может быть использован противниками для усиления кибератак, автоматизации попыток фишинга и более эффективного использования уязвимостей (Counter AI Attacks with AI Defense, Palo Alto Networks**).
* **Необходимость надежных стратегий защиты от ИИ.** Поскольку ИИ играет центральную роль в критически важных отраслях, обеспечение его безопасности уже не факультативно, а необходимо. Правительства и частные организации инвестируют в системы защиты от ИИ, чтобы противостоять этим угрозам. Национальный институт стандартов и технологий (NIST) определил различные классы кибератак, направленных на ИИ, и рекомендовал стратегии по их снижению.

Кроме того, Министерство обороны США (DoD) выпустило стратегию внедрения ИИ, в которой подчеркивается важность защиты технологий ИИ в военных и оборонных приложениях (DOD Releases AI Adoption Strategy, Defense.gov).

Сообщество специалистов по безопасности ИИ предложило множество механизмов защиты для решения этих проблем - от обучения противника и надежных архитектур моделей до методов шифрования и обнаружения угроз на основе ИИ.

**Классификация угроз ИИ**

**Механизмы защиты**

**Текущие проблемы и пробелы**

Несмотря на значительный прогресс в разработке защитных механизмов ИИ, остается ряд проблем и пробелов, делающих системы ИИ уязвимыми для атак противника, отравления данных и других форм манипулирования. Быстрое развитие технологий ИИ в сочетании с растущей изощренностью злоумышленников создает постоянно развивающееся поле боя, на котором защитные меры не успевают за ними. В этом разделе рассматриваются основные проблемы защиты систем ИИ и определяются области, в которых существующие стратегии защиты недостаточны.

1. Гонка вооружений между злоумышленниками и защитниками. Одной из основных проблем в области безопасности ИИ является постоянная эскалация между противниками и защитниками. Злоумышленники постоянно разрабатывают новые методы использования моделей ИИ, а исследователи и инженеры работают над созданием контрмер. Многие существующие стратегии защиты являются скорее реактивными, чем проактивными, то есть они устраняют угрозы после обнаружения уязвимостей, а не предотвращают их заранее. Например, обучение с применением противника - метод, при котором модели ИИ обучаются на противных примерах, - может повысить устойчивость, но злоумышленники часто находят способы генерировать новые образцы противника, которые обходят существующие средства защиты. В результате безопасность ИИ постоянно играет в догонялки, что затрудняет создание долгосрочных и надежных механизмов защиты.

2. Отсутствие стандартизированных защитных фреймворков. В отличие от традиционной кибербезопасности, которая имеет четко определенные протоколы и лучшие практики, безопасность ИИ не имеет общепринятых рамок для оценки и снижения рисков. Различные организации и исследователи используют разные подходы к обеспечению безопасности моделей ИИ, что приводит к несоответствию защитных механизмов. Отсутствие стандартизированных сред тестирования затрудняет сравнение эффективности различных стратегий защиты. Пирамида боли в безопасности ИИ, представленная в последних исследованиях, пытается классифицировать угрозы, характерные для ИИ, однако до сих пор не существует общепринятого эталона для комплексной оценки безопасности ИИ. Без единого подхода организации могут применять частичные или неэффективные стратегии защиты, оставляя системы ИИ незащищенными от непредвиденных атак.

3. Уязвимости в целостности данных и моделей. Многие атаки на ИИ направлены на данные, используемые для обучения и точной настройки моделей, используя тот факт, что системы ИИ в значительной степени зависят от огромного количества входных данных. Атаки с отравлением данных, например, манипулируют обучающими наборами данных, чтобы внести предубеждения или «черные ходы» в процесс принятия решений моделью. Поскольку модели ИИ часто работают как «черные ящики», обнаружить тонкие манипуляции в больших массивах данных очень сложно. Даже если такие угрозы обнаружены, очистить отравленные данные бывает непросто, особенно в крупномасштабных моделях, требующих длительного переобучения. Кроме того, атаки на инверсию модели и атаки на вывод членства могут извлекать конфиденциальную информацию из моделей ИИ, что вызывает серьезные проблемы с конфиденциальностью. Существующие методы защиты, такие как дифференциальная конфиденциальность и надежные алгоритмы обучения, не всегда достаточны для предотвращения этих атак в масштабе.

4. Компромиссы между надежностью, точностью и эффективностью. Многие защитные механизмы ИИ предусматривают компромиссы, которые влияют на производительность модели. Надежные модели ИИ часто требуют дополнительных вычислительных ресурсов, что делает их более медленными и дорогими в развертывании. Например, обучение противника и защитная дистилляция могут повысить безопасность, но обычно снижают точность модели и увеличивают время обучения. В реальных приложениях - таких как автономные транспортные средства, диагностика в здравоохранении и выявление финансового мошенничества - эти компромиссы становятся критически важными, поскольку чрезмерные вычислительные затраты могут ограничить практичность систем ИИ. Организациям приходится искать баланс между безопасностью и эффективностью, и во многих случаях приоритет отдается скорости и точности, а не устойчивости к враждебным угрозам.

5. Обобщение механизмов защиты. Большинство стратегий защиты ИИ разработаны для конкретных типов атак или архитектур моделей, что делает их менее эффективными против новых угроз. Например, защитный механизм, созданный для защиты сверточных нейронных сетей (CNN) от вражеских возмущений, может оказаться не столь эффективным для моделей на основе трансформаторов, используемых в обработке естественного языка (NLP). Более того, некоторые меры безопасности, работающие в контролируемых условиях, не срабатывают при применении к реальным системам ИИ, которые обрабатывают динамические и непредсказуемые входные данные. Невозможность обобщить методы защиты ИИ в различных областях и архитектурах представляет собой серьезную проблему, требующую постоянных исследований и адаптации.

6. Проблемы объяснимости и прозрачности. Серьезной проблемой в области безопасности ИИ является недостаточная объяснимость и прозрачность моделей глубокого обучения. Многие передовые системы ИИ функционируют как «черные ящики», то есть их внутренние процессы принятия решений не поддаются интерпретации. Отсутствие прозрачности затрудняет выявление уязвимостей в системе безопасности и понимание того, как система ИИ может вести себя при воздействии на нее враждебных факторов. Разработка объяснимого ИИ (XAI) является активной областью исследований, но существующие методы XAI все еще ограничены в своей способности обеспечить детальное понимание внутренней работы сложных моделей. Без четких объяснений того, как системы ИИ обрабатывают информацию и принимают решения, сложнее разработать эффективные защитные механизмы, направленные на устранение потенциальных угроз на фундаментальном уровне.

7. Этические и нормативные проблемы. Все более широкое внедрение ИИ в критически важные приложения, такие как здравоохранение, финансы и национальная безопасность, вызывает этические и нормативные проблемы, связанные с его безопасностью. Правительства и организации все больше осознают риски, связанные с уязвимостями ИИ, однако существующие нормативные акты зачастую устарели или недостаточны для борьбы с возникающими угрозами. В стратегии Министерства обороны США по внедрению ИИ подчеркивается необходимость обеспечения безопасности ИИ, однако во многих странах отсутствует аналогичная политика, направленная на соблюдение строгих стандартов безопасности. Кроме того, меры безопасности ИИ должны сочетать защиту с этическими соображениями, например, избегать чрезмерно ограничительных мер защиты, которые могут ограничить доступность или справедливость технологий ИИ. Отсутствие четких правовых норм в области безопасности ИИ еще больше усложняет внедрение эффективных механизмов защиты в глобальном масштабе.

Безопасность ИИ - это быстро развивающаяся область, но многие проблемы остаются нерешенными. Постоянная гонка вооружений между злоумышленниками и защитниками, отсутствие стандартизированных основ безопасности, уязвимости в целостности данных, компромисс между надежностью и эффективностью, а также проблемы, связанные с объяснимостью, - все это усугубляет трудности в защите систем ИИ. Кроме того, этические и нормативные проблемы еще больше осложняют развертывание эффективных стратегий защиты ИИ. Устранение этих недостатков требует непрерывных исследований, междисциплинарного сотрудничества и разработки проактивных мер безопасности, способных адаптироваться к возникающим угрозам. Поскольку ИИ все больше интегрируется в критически важную инфраструктуру, обеспечение его безопасности становится как никогда важным, и преодоление этих проблем будет иметь решающее значение для формирования будущего надежного ИИ.

**Источники - интересные статьи**

**1. Adversarial Attacks and Defense Mechanisms**

* **"Adversarial Attack and Defense: A Survey"**  
  This article provides an in-depth analysis of adversarial attacks on deep neural networks and reviews existing defense strategies. It highlights the vulnerabilities of AI models and discusses methods to enhance their robustness.
* **"Cybersecurity for AI Systems: A Survey"**  
  This survey addresses various attacks on AI systems, including adversarial examples and model poisoning, and discusses potential defense mechanisms to mitigate these threats. <https://www.mdpi.com/2079-9292/11/8/1283>

**2. AI Security Risks and Threats**

* **"AI Security Risks and Threats"**  
  This resource outlines the security risks associated with AI, such as data poisoning and model inversion attacks, and emphasizes the importance of securing AI models and their training data. <https://www.checkpoint.com/cyber-hub/cyber-security/what-is-ai-security/ai-security-risks-and-threats/>
* **"AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways"**  
  This survey explores the security challenges faced by AI agents, categorizing them into critical knowledge gaps and suggesting future research directions to enhance AI security.

<https://arxiv.org/abs/2406.02630>

**3. Defense Strategies and Frameworks**

* **"The AI Security Pyramid of Pain"**  
  This paper introduces a framework that categorizes and prioritizes AI-specific threats, providing a structured approach to understanding and addressing various levels of AI threats. <https://arxiv.org/abs/2402.11082>
* **"Counter AI Attacks with AI Defense"**  
  This article discusses the need for organizations to harness AI to counteract the increasing use of AI by adversaries, highlighting the importance of proactive defense strategies. <https://www.paloaltonetworks.com/blog/2024/05/counter-with-ai-defense/>

**4. AI in Cyber Defense**

* **"AI Cyber Defense: Anticipate and Neutralize Cyber Threats"**  
  This resource explores how AI is transforming cybersecurity by enabling organizations to anticipate and neutralize cyber threats through real-time analysis and predictive capabilities. <https://smartdev.com/strategic-cyber-defense-leveraging-ai-to-anticipate-and-neutralize-modern-threats/>
* **"What Is the Role of AI in Threat Detection?"**  
  This article examines how AI enhances threat detection capabilities in cybersecurity, focusing on adaptive learning and advanced pattern recognition. <https://www.paloaltonetworks.com/cyberpedia/ai-in-threat-detection>

**5. Government and Policy Perspectives**

* **"DOD Releases AI Adoption Strategy"**  
  This news release discusses the U.S. Department of Defense's strategy to accelerate the adoption of AI capabilities, emphasizing the importance of maintaining decision superiority on the battlefield. <https://www.defense.gov/News/News-Stories/Article/Article/3578219/dod-releases-ai-adoption-strategy/>
* **"NIST Identifies Types of Cyberattacks That Manipulate Behavior of AI Systems"**  
  This publication identifies various classes of attacks on AI systems and suggests approaches for mitigating them, contributing to the development of trustworthy AI. <https://www.nist.gov/news-events/news/2024/01/nist-identifies-types-cyberattacks-manipulate-behavior-ai-systems>

1.

В этой статье представлен углубленный анализ состязательных атак на глубокие нейронные сети и рассматриваются существующие стратегии защиты. В ней подчеркиваются уязвимости моделей ИИ и обсуждаются методы повышения их надежности.

В этом обзоре рассматриваются различные атаки на системы ИИ, включая примеры состязательных атак и отравление моделей, а также обсуждаются потенциальные механизмы защиты для смягчения этих угроз. https://www.mdpi.com/2079-9292/11/8/1283

2.

В этом ресурсе описываются риски безопасности, связанные с ИИ, такие как отравление данных и атаки с инверсией моделей, и подчеркивается важность защиты моделей ИИ и их обучающих данных. https://www.checkpoint.com/cyber-hub/cyber-security/what-is-ai-security/ai-security-risks-and-threats/

В этом обзоре изучаются проблемы безопасности, с которыми сталкиваются агенты ИИ, классифицируются по критическим пробелам в знаниях и предлагаются будущие направления исследований для повышения безопасности ИИ.

https://arxiv.org/abs/2406.02630

3.

В этой статье представлена ​​рамка, которая классифицирует и расставляет приоритеты угроз, специфичных для ИИ, обеспечивая структурированный подход к пониманию и устранению различных уровней угроз ИИ. https://arxiv.org/abs/2402.11082

В этой статье обсуждается необходимость использования ИИ организациями для противодействия все более широкому использованию ИИ противниками, подчеркивая важность стратегий проактивной защиты. https://www.paloaltonetworks.com/blog/2024/05/counter-with-ai-defense/

4.

В этом ресурсе рассматривается, как ИИ преобразует кибербезопасность, позволяя организациям предвидеть и нейтрализовать киберугрозы с помощью анализа в реальном времени и возможностей прогнозирования. https://smartdev.com/strategic-cyber-defense-leveraging-ai-to-anticipate-and-neutralize-modern-threats/

В этой статье рассматривается, как ИИ расширяет возможности обнаружения угроз в кибербезопасности, уделяя особое внимание адаптивному обучению и расширенному распознаванию образов. https://www.paloaltonetworks.com/cyberpedia/ai-in-threat-detection

5.

В этом пресс-релизе обсуждается стратегия Министерства обороны США по ускорению внедрения возможностей ИИ, подчеркивая важность сохранения превосходства в принятии решений на поле боя. https://www.defense.gov/News/News-Stories/Article/Article/3578219/dod-releases-ai-adoption-strategy/

В этой публикации определяются различные классы атак на системы ИИ и предлагаются подходы к их смягчению, что способствует разработке надежного ИИ. https://www.nist.gov/news-events/news/2024/01/nist-identifices-types-cyberattacks-manipulate-behavior-ai-systems