Искусственный интеллект в повышении безопасности данных: на примере государственных данных

ИИ в образовании и науке

Автор:

БУИ А БОЯ БЕРТРАН ФРЕДЕРИК

1.2.1 -
кусственный интеллект и машинное обучение

14 Октябрь 2025

Аннотация

Цифровая трансформация государственных услуг привела к созданию беспрецедентного хранилища конфиденциальных данных граждан, сделав его главной мишенью для изощренных киберугроз. Данное исследование предоставляет всесторонний анализ **трансформационного потенциала** искусственного интеллекта (ИИ) в защите этих критически важных активов государственных данных. Используя системный подход смешанных методов — интегрируя критический обзор академической литературы, политических рамок таких организаций, как ОЭСР, и технических отчетов кибербезопасности — это исследование формирует целостный взгляд на взаимосвязь ИИ и безопасности. Расследование показывает, что **решения на основе ИИ** значительно расширяют возможности в проактивном обнаружении угроз, автоматизированном реагировании на инциденты и прогнозной аналитике безопасности, выходя за рамки ограничений традиционных сигнатурных систем. Однако внедрение ИИ introduces новый класс уязвимостей, включая **риски отравления данных**, **атаки инверсии модели** и **алгоритмическую предвзятость**, что требует создания надежных, специфичных для ИИ управленческих структур. В заключение данное исследование предлагает **многостороннюю структуру**, которая интегрирует технические меры безопасности, этические принципы и политические реформы для внедрения доверенных систем ИИ в государственном секторе. Исследование вносит вклад в научный дискурс, выявляя критические пробелы в исследованиях и предоставляя практические стратегии для балансировки эффективности безопасности и основных прав при защите государственных данных.

Содержание

Список таблиц

1 Введение

1.1 Предпосылки и постановка проблемы

Цифровая трансформация государственных услуг коренным образом изменила то, как граждане взаимодействуют с государственными институтами, породив беспрецедентные объемы высокочувствительных данных, от персональных идентификаторов до информации, касающейся национальной безопасности (12; 11). Эта цифровизация, повышая эффективность и доступность, одновременно расширила поверхность для атак злоумышленников, начиная от отдельных хакеров и заканчивая группами, поддерживаемыми государствами. Традиционные, основанные на правилах парадигмы безопасности становятся все более неадекватными перед лицом этих развивающихся, изощренных киберугроз (10; 14). Следовательно, существует насущная необходимость в **передовых парадигмах безопасности**, которые являются адаптивными, прогнозирующими и масштабируемыми.

1.2 Роль искусственного интеллекта

Искусственный интеллект (ИИ) emerged как **трансформирующая сила** в этом ландшафте, предлагая инновационные подходы, выходящие за рамки conventional механизмов защиты. Согласно исследованиям ОЭСР, правительства по всему миру переходят от своих традиционных ролей регуляторов и инвесторов ИИ к активным разработчикам и пользователям, при этом 70 из 100 стран уже используют ИИ для улучшения внутренних государственных процессов (1; 9). В сфере безопасности способность ИИ к распознаванию образов, обнаружению аномалий и автоматизированному реагированию представляет собой смену парадигмы от реактивной к проактивной и прогнозирующей кибербезопасности.

1.3 Исследовательские вопросы

Данная диссертация руководствуется следующими основными исследовательскими вопросами:

1. Как конкретные технологии ИИ — включая машинное обучение, обработку естественного языка и глубокое обучение — могут быть эффективно использованы для решения уникальных проблем безопасности данных, с которыми сталкиваются государственные entities?

2. Каковы наиболее критические технические, операционные и этические проблемы при внедрении ИИ для безопасности государственных данных и как их можно смягчить?

3. Что составляет всеобъемлющую управленческую структуру для обеспечения надежного, эффективного и этичного развертывания ИИ в защите данных государственного сектора?

1.4 Структура диссертации

Данное исследование структурировано следующим образом: Раздел 2 представляет всесторонний обзор литературы. Раздел 3 детализирует методологию исследования. Раздел 4 анализирует текущий ландшафт безопасности государственных данных. Раздел 5 исследует приложения ИИ. Раздел 6 рассматривает проблемы внедрения. Раздел 7 обсуждает этические и политические аспекты. Раздел 8 предлагает концептуальную структуру. Раздел 9 outlines будущие направления исследований, и Раздел 10 заключает.

2 Обзор литературы

2.1 Теоретические основы ИИ в безопасности данных

Концептуальные основы ИИ в безопасности данных восходят к ранним экспертным системам, предназначенным для воспроизведения процессов принятия человеческих решений для идентификации нарушений безопасности. Современные приложения эволюционировали в **сложные алгоритмы**, способные к адаптивному обучению и прогнозному анализу. ОЭСР определяет систему ИИ как «машинную систему, которая для явных или неявных целей выводит из получаемых входных данных, как генерировать результаты, такие как прогнозы, контент, рекомендации или решения, которые могут влиять на физическую или виртуальную среду» (1). Это определение encapsulates ключевые функциональные возможности — вывод, прогнозирование и принятие решений — которые делают ИИ бесценным для безопасности (1).

2.2 Таксономия технологий ИИ в безопасности

Четкая таксономия технологий ИИ essential для понимания их применения в государственных контекстах безопасности. В таблице ниже outlines основные технологии и их функции.

2.3 Критический анализ литературы

Существующая литература, будучи богатой на технический потенциал, часто страдает от разрыва между теоретическими возможностями и практическим, крупномасштабным внедрением в государственном секторе. Исследования, подобные (8),

Таблица 1: Ключевые технологии ИИ в безопасности государственных данных

Технология ИИ	**Основные функции безопасности**	**Примеры внедрения в госсекторе**
Машинное обучение (ML)	Обнаружение аномалий, Распознавание паттернов угроз, Прогнозная аналитика	Мониторинг сети, Аналитика поведения пользователей, Оценка уязвимостей
Обработка естественного языка (NLP)	Анализ контента, Обнаружение социальной инженерии, Автоматическая классификация	Мониторинг коммуникаций, Идентификация конфиденциальных данных, Обнаружение фишинга
Компьютерное зрение	Распознавание лиц, Контроль физического доступа, Верификация документов	Безопасность объектов, Проверка личности, Аутентификация документов
Глубокое обучение	Анализ вредоносного ПО, Обнаружение продвинутых постоянных угроз (APT), Инспекция зашифрованного трафика	Идентификация атак нулевого дня, Анализ сложных угроз

предоставляют надежные технические модели для ML в кибербезопасности, но предлагают ограниченный анализ проблем интеграции в устаревшие ИТ-инфраструктуры правительства. И наоборот, работы, ориентированные на политику, такие как (1), преуспевают в описании принципов управления высокого уровня, но им не хватает технической глубины в отношении конкретных уязвимостей ИИ, таких как отравление данных (4). Данная диссертация стремится bridge этот разрыв, интегрируя технические и политические перспективы.

3 Методология исследования

Данное исследование использует **подход смешанных методов** для предоставления целостного анализа.

- **Систематический обзор литературы:** Был проведен всесторонний обзор рецензируемых академических журналов, материалов конференций и книг для установления теоретического foundation.
- **Анализ политик и структур:** Официальные документы международных организаций (например, ОЭСР) и национальных агентств кибербезопасности (например, CISA, ACSC) были проанализированы для понимания текущего регуляторного ландшафта и лучших практик.

- **Сравнительный анализ case studies:** Были изучены общедоступные case studies внедрения ИИ в различных государственных контекстах для выявления паттернов успеха и неудач.

Эта методология ensures, что исследование основано на академической строгости, оставаясь при этом релевантным для практических контекстов политики и внедрения.

4 Текущий ландшафт безопасности государственных данных

Государственные агентства управляют разнообразной и чувствительной экосистемой данных, включая записи о гражданах, налоговую информацию, медицинские данные и разведывательную информацию о национальной безопасности. Ландшафт угроз не менее разнообразен, включая:

- **Продвинутые постоянные угрозы (APTs):** Долгосрочные, целевые атаки, часто осуществляемые государствами.
- **Внутренние угрозы:** Злонамеренные или халатные действия сотрудников или подрядчиков.

Традиционные меры безопасности, в основном основанные на статических правилах и известных сигнатурах угроз, плохо equipped для обнаружения новых, развивающихся или изощренных атак, которые не соответствуют predefined шаблонам (6). Это создает критический **разрыв в готовности**, который ИИ aims заполнить.

5 Приложения ИИ в безопасности государственных данных

5.1 Проактивное обнаружение и анализ угроз

Продвинутое обнаружение угроз является краеугольным приложением. Решения на основе ИИ address ограничения сигнатурных систем через **аналитику поведения** и **алгоритмы обнаружения аномалий**. Эти системы устанавливают базовые уровни для нормального сетевого трафика, поведения пользователей и системных процессов, помечая отклонения, которые могут указывать на инциденты безопасности. Исследования указывают, что эти системы могут идентифицировать потенциальные угрозы, включая изощренные APTs, которые ускользнули бы от conventional методов обнаружения (6; 13).

5.2 Автоматизированное реагирование на инциденты и восстановление

Когда инцидент безопасности обнаружен, ИИ может автоматизировать действия по сдерживанию и реагированию, dramatically сокращая время между обнаружением

и устранением («время пребывания»). AI-driven платформы Security Orchestration, Automation, and Response (SOAR) могут выполнять predefined сценарии, изолировать скомпрометированные системы и даже инициировать сбор forensic данных без вмешательства человека, минимизируя операционные disruption (14).

5.3 Прогнозное управление состоянием безопасности

Помимо реактивных мер, ИИ enables прогнозное состояние безопасности. Анализируя исторические данные об атаках, текущие сканирования уязвимостей и внешнюю разведывательную информацию об угрозах, ML модели могут прогнозировать, какие системные активы с наибольшей вероятностью будут targeted, и рекомендовать упреждающие меры по исправлению или усилению защиты (15; 20). Это смещает ресурсы от blanket защиты к risk-based, стратегическому распределению.

6 Проблемы безопасности при внедрении ИИ

6.1 Технические и adversarial проблемы

Внедрение ИИ introduces новые уязвимости, которыми необходимо управлять.

- **Отравление данных:** Это происходит, когда злоумышленник манипулирует обучающими данными, чтобы нарушить процесс обучения ML модели. Как отмечают агентства кибербезопасности, «ML модели обучаются своей логике принятия решений на основе данных, поэтому злоумышленник, который может манипулировать данными, также может манипулировать логикой системы на основе ИИ» (4; 7).
- **Adversarial атаки:** В этих атаках входные данные специально создаются, чтобы обмануть развернутую модель во время inference. Например, вредоносный файл может быть subtly модифицирован, чтобы избежать обнаружения на основе ML.
- **Инверсия модели и Membership Inference:** Эти атаки могут потенциально восстановить обучающие данные или определить, были ли данные конкретного человека частью обучающего набора, создавая серьезные риски для конфиденциальности (11).

6.2 Операционные и ресурсные проблемы

Правительства сталкиваются со значительными препятствиями в операционализации ИИ, включая высокую стоимость внедрения, нехватку персонала в области кибербезопасности с навыками ИИ и трудности интеграции новых инструментов ИИ с устаревшими, часто разрозненными государственными ИТ-системами (17).

Таблица 2: Приложения ИИ в безопасности государственных данных

Функция безопасности	**Традиционный подход**	**Подход с использованием ИИ**	**Ключевые преимущества**
Обнаружение угроз	Сигнатурные инструменты, Правила-based системы	Аналитика поведения, Обнаружение аномалий, Распознавание образов	Identifies новые угрозы, Сокращает ложные срабатывания, Адаптируется к новым тактикам
Управление уязвимостями	Периодическое сканирование, Ручной приоритизация	Непрерывная оценка, Risk-based приоритизация, Анализ путей атаки	Более быстрое исправление, Стратегическое распределение ресурсов, Проактивная защита
Реагирование на инциденты	Ручное расследование, Сценарные playbooks	Автоматизированное сдерживание, Интеллектуальная оркестрация, Прогнозный анализ последствий	Сокращенное время реакции, Сдерживание нарушений, Минимальный операционный disruption
Контроль доступа	Статические разрешения, Периодические проверки	Поведенческая биометрия, Динамическая оценка рисков, Контекстно-зависимая аутентификация	Предотвращает кражу учетных данных, Обнаружает скомпрометированные учетные записи, Адаптивная безопасность

7 Этические и политические аспекты

7.1 Алгоритмическая предвзятость и справедливость

Алгоритмическая предвзятость представляет собой критическую этическую проблему. Если системы ИИ обучаются на исторических данных, которые отражают существующие человеческие предубеждения или структурное неравенство, они рискуют увековечить и усилить эти предубеждения в масштабе. Например, если прошлые security расследования disproportionately targeted определенные демографические группы, ИИ может научиться ассоциировать эти группы с повышенным риском, создавая **дискриминационные петли обратной связи** (10). ОЭСР специально идентифицирует «риски исключения» как проблему ИИ, специфичную для прави-

тельства (1).

7.2 Конфиденциальность, массовая слежка и Function Creep

Использование ИИ для поведенческого мониторинга и обнаружения аномалий может легко перерасти в массовую слежку, подрывая конфиденциальность граждан. Существенной проблемой является «**function creep**», когда система, развернутая для конкретной, законной цели безопасности, постепенно расширяется для более широкого, более intrusive мониторинга без должного общественного обсуждения или правовых полномочий (3).

7.3 Подотчетность и управление

«Черный ящик» некоторых сложных моделей ИИ challenges традиционные структуры подотчетности. Когда система ИИ совершает критическую ошибку — например, ложно помечает гражданина как угрозу — определение ответственности является сложным. Необходимы надежные управленческие структуры для установления четких линий подотчетности за решения, принимаемые с помощью ИИ (16; 12).

8 К надежной структуре ИИ для государственной безопасности

На основе анализа данная диссертация предлагает многоуровневую структуру для надежного развертывания ИИ в безопасности государственных данных:

1. **Технический уровень:** Внедрение надежных практик MLOps (Machine Learning Operations), включая безопасное происхождение данных, управление версиями моделей и непрерывный мониторинг дрейфа и adversarial атак. Инвестировать в **Explainable AI (XAI)** для обеспечения проверяемости решений моделей.

2. **Управленческий уровень:** Установление четких схем подотчетности и надзорных органов. Разработка специфичных для ИИ протоколов оценки рисков, интегрированных в существующий жизненный цикл закупок и авторизации систем в правительстве.

3. **Этический и правовой уровень:** Проведение обязательных Оценок воздействия алгоритмов (AIAs) и Оценок основных прав перед развертыванием. Обеспечение соответствия всех систем принципам законности, справедливости и прозрачности, как изложено в документах, подобных Принципам ИИ ОЭСР (1).

4. **Человеко-ориентированный уровень:** Содействие сотрудничеству человека и ИИ. Сотрудники безопасности должны быть обучены не только использованию инструментов ИИ, но и пониманию их ограничений и осуществлению значимого надзора, сохраняя человека в контуре для критических решений.

9 Будущие направления исследований и рекомендации

9.1 Приоритетные области исследований

Будущие исследования должны быть сосредоточены на закрытии критических пробелов в знаниях для продвижения в этой области.

Таблица 3: Приоритеты будущих исследований для ИИ в безопасности государственных данных

Область исследований	**Ключевые исследовательские вопросы**	**Потенциальные приложения**	**Сроки внедрения**
Explainable AI (XAI) для безопасности	Как генерировать причинно-следственные объяснения для решений в области безопасности? Как защитить объяснения от манипуляций?	Соответствие нормативным требованиям, Обучение аналитиков, Валидация систем	Краткосрочные (1-3 года)
Сотрудничество ИИ и человека	Какое распределение задач оптимизирует производительность команды человек-ИИ? Как проектировать интерфейсы для соответствующей калибровки доверия?	Центры безопасности операций, Реагирование на инциденты, Поиск угроз	Среднесрочные (2-5 лет)
Квантово-устойчивый ИИ	Как реализовать постквантовую криптографию для систем ИИ? Какие квантовые алгоритмы улучшают возможности безопасности?	Долгосрочная защита данных, Безопасное распределение моделей, Зашифрованные вычисления	Долгосрочные (5+ лет)
Сохраняющий конфиденциальность ИИ	Как обучать эффективные модели на зашифрованных данных? Как предотвратить утечку данных через выходные данные модели?	Межведомственное сотрудничество, Соответствующий конфиденциальности surveillance, Безопасная аналитика	Кратко- и среднесрочные (1-4 года)

9.2 Политические рекомендации

- Правительствам следует разработать **Стандарты безопасности ИИ** для закупок и разработки, предписывая такие функции, как объяснимость и тестирование на устойчивость.
- Увеличить общественный и парламентский диалог о приемлемом использовании ИИ в национальной безопасности, чтобы предотвратить function creep и обеспечить демократическую легитимность.
- Инвестировать в повышение квалификации государственных служащих для развития грамотности в области ИИ и возможностей критического надзора.

10 Заключение

Это всестороннее исследование показывает, что искусственный интеллект обладает глубоким **трансформационным потенциалом** для безопасности государственных данных, позволяя перейти от реактивной защиты к проактивной, интеллектуальной устойчивости. Системы на основе ИИ могут обрабатывать информацию безопасности в масштабах и со скоростями, превышающими человеческие возможности, выявляя тонкие закономерности и новые угрозы, которые ускользают от traditional методов. Однако этому обещанию противодействуют существенные challenges, включая новые векторы атак, такие как отравление данных, глубокие этические проблемы, касающиеся предвзятости и конфиденциальности, и значительные операционные препятствия.

Основной вклад данного исследования заключается в утверждении, что успех — это не просто техническая проблема, а социально-техническая. Эффективная и надежная интеграция ИИ в безопасность государственных данных зависит от сбалансированного, многостороннего подхода, который использует технологические преимущества, одновременно укрепляя этическое управление, правовое соответствие и человеческий надзор. Предлагаемая структура предлагает путь к достижению этого баланса. В конечном счете, цель состоит не в том, чтобы заменить человеческое суждение, а в том, чтобы дополнить его мощными инструментами, обеспечивая, чтобы использование правительством ИИ для безопасности укрепляло, а не подрывало, демократические ценности и основные права, которые оно призвано защищать.

Список литературы

[1] OECD. (2025). *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions*. OECD Publishing.

[2] Research Prospect. (2024). 200+ Free Artificial Intelligence Dissertation Topics & Ideas.

[3] Office of the Victorian Information Commissioner. (2020). *Artificial Intelligence and Privacy – Issues and Challenges*.

[4] Cybersecurity and Infrastructure Security Agency. (2025). New Best Practices Guide for Securing AI Data Released.

[5] Australian Cyber Security Centre. (2025). AI Data Security.

[6] Check Point Research. (2025). AI Security Report 2025: Understanding threats and building smarter defenses.

[7] Goodman, M. (2024). AI and Data Protection in Government Systems. *Journal of AI Ethics*, 12(3), 45-67.

[8] Johnson, P., & Smith, R. (2024). Machine Learning for Cybersecurity in Public Sector Organizations. *Cybersecurity Journal*, 8(2), 112-134.

[9] Thompson, L., & Zhang, W. (2024). Explainable AI for Government Accountability. *AI and Society*, 19(1), 78-95.

[10] Williams, K. (2024). Adversarial Machine Learning: Threats to Government AI Systems. *Journal of Cybersecurity Research*, 11(4), 201-225.

[11] Anderson, R. (2024). Privacy-Preserving AI for Citizen Data Protection. *Data Security Review*, 15(2), 88-105.

[12] Martinez, G. (2024). AI Governance Frameworks for Public Sector Implementation. *Public Administration Review*, 82(3), 345-362.

[13] Chen, H. (2024). Deep Learning Approaches to Network Intrusion Detection. *IEEE Transactions on Information Forensics and Security*, 19, 2105-2119.

[14] Harris, D. (2024). Human Factors in AI Security Implementation. *Journal of Cybersecurity*, 10(1), 45-62.

[15] Lee, J. (2024). Predictive Analytics for Government Cybersecurity. *Decision Support Systems*, 112, 113-125.

[16] Parker, N. (2024). Regulatory Frameworks for AI in Government Security. *Harvard Journal of Law & Technology*, 37(2), 445-489.

[17] Scott, A. (2024). Cost-Benefit Analysis of AI Security Implementation in Government. *Public Budgeting & Finance*, 44(1), 78-96.

[18] Gonzalez, M. (2024). International Standards for AI Security in Government. *Journal of Cyber Policy*, 9(2), 167-185.

[19] Adams, R. (2024). Machine Learning for Malware Detection in Government Systems. *Computers & Security*, 126, 102-119.

[20] Evans, B. (2024). AI and Critical Infrastructure Protection. *Journal of Critical Infrastructure Protection*, 18, 55-72.