ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«07» марта 2024 г.
Протокол №1

Рабочая программа дисциплины (модуля)
«Безопасность систем ИИ»

Направление подготовки: *02.03.01 Математика и компьютерные науки*

Направленность (профиль) подготовки: *Искусственный интеллект*

Квалификация (степень) выпускника: *бакалавр*

Форма обучения: *очная*

Срок освоения программы: *4 года*

Год набора: *2024*

Москва
2024

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 0215B80C000D1B3FEB44A37EE631E0C27D4
Владелец: Ивашкевич Евгений Васильевич
Действителен: с 13.01.2026 до 13.04.2027

Электронный документ

Содержание

1. **Краткая характеристика дисциплины (модуля)** 3
2. **Перечень планируемых результатов обучения** 5
3. **Тематический план** 7
4. **Содержание дисциплины (модуля)** 7
5. **Учебно-методическое обеспечение** 8
6. **Материально-техническое обеспечение** 8
7. **Методические и оценочные материалы** 10

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Безопасность систем ИИ» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по направлению 02.03.01 «Математика и компьютерные науки», профиль «Искусственный интеллект», утвержденный приказом Министерства науки и высшего образования Российской Федерации № 807 от 23.08.2017 года.

Изучение дисциплины (модуля) «Безопасность систем ИИ» позволяет подготовить специалистов, способных обеспечивать надёжность и безопасность ИИ-систем в условиях растущих киберугроз. Полученные знания и навыки формируют ключевые компетенции для работы в области безопасного машинного обучения, анализа уязвимостей и разработки защищённых ИИ-решений в критически важных отраслях.

Место дисциплины (модуля) в структуре образовательной программы

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки бакалавриата по направлению 02.03.01 «Математика и компьютерные науки», профиль «Искусственный интеллект» и входит в вариативную часть Блока 1 формируемую участниками образовательных отношений как дисциплина по выбору.

Дисциплина (модуль) изучается на 3 курсе в 6 семестре. Доступна к изучению после успешного освоения дисциплины (модуля): «Глубокое обучение».

Цель изучения дисциплины (модуля): сформировать у обучающихся системное понимание угроз и механизмов защиты в системах искусственного интеллекта, а также навыки проектирования, анализа и тестирования ИИ-решений с учётом требований кибербезопасности.

Задачи изучения дисциплины (модуля):

— изучить архитектуры и модели искусственного интеллекта, применяемые как в кибератаках, так и в системах защиты, и классифицировать основные типы угроз;

— освоить схемы атак на модели машинного обучения, включая отравление данных, уклонение и атаки на интеллектуальную собственность, а также методы их детекции и предотвращения;

— научиться оценивать безопасность наборов данных и моделей ИИ с точки зрения уязвимостей, связанных с отравлением, инференсом и эксплуатацией генеративных моделей;

— развить навыки использования современных открытых фреймворков для тестирования и повышения устойчивости моделей машинного обучения к атакам;

— сформировать способность обоснованно выбирать архитектурные и алгоритмические решения при разработке безопасных ИИ-систем и оценивать сопутствующие риски.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

— основные области использования систем искусственного интеллекта в кибербезопасности;

— архитектуру и модели систем искусственного интеллекта, используемых при организации кибератак;

— архитектуру и модели систем искусственного интеллекта, используемых при защите от кибератак;

— схемы организации атак на модели машинного обучения;

— классификацию атак на модели машинного обучения (искусственного интеллекта);

— схемы атак отравления данных и методы борьбы с ними;

— схемы атак отравления моделей машинного обучения;

— схемы атак уклонением и методы борьбы с ними;

— схемы атак на интеллектуальную собственность и методы борьбы с ними;

— схемы использования порождающих моделей в атаках на системы машинного обучения (искусственного интеллекта).

уметь:

— строить модели систем искусственного интеллекта для использования в кибербезопасности;

— организовывать тестирование систем искусственного интеллекта для оценки их безопасности;

— оценивать наборы данных, используемые в моделях искусственного интеллекта с точки зрения безопасности;

— формировать обоснованную оценку организации и функционирования тех или иных компонентов систем искусственного интеллекта, как в контексте их базовых функций, так и с точки зрения рисков кибербезопасности;

— использовать современные открытые фреймворки для задач кибербезопасности систем искусственного интеллекта.

владеть:

— навыком выбора архитектурных решений для систем искусственного интеллекта, используемых в кибербезопасности;

— навыком анализа моделей машинного обучения (искусственного интеллекта) с точки зрения кибербезопасности (защиты от атак);

— навыком выбора моделей и алгоритмов для систем искусственного интеллекта, используемых в кибербезопасности;

— навыком оценки рисков безопасности систем искусственного интеллекта.

2. Перечень планируемых результатов обучения

Компетенции, формируемые в результате освоения дисциплины (модуля) при проведении учебных занятий в форме контактной работы обучающихся с педагогическими работниками Университета и в форме самостоятельной работы обучающихся:

Компетенция	Содержание компетенции	Индикатор компетенции	Перечень планируемых результатов обучения по дисциплине (модулю)
УК-1.	Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1.	Знает методы поиска и анализа информации в области искусственного интеллекта, основные принципы критической оценки источников информации и их релевантности.
		УК-1.2.	Умеет критически оценивать источники информации и синтезировать данные из различных источников для решения задач, применять системный подход к анализу и решению комплексных проблем
		УК-1.3.	Имеет практический опыт работы с современными инструментами и технологиями для обработки информации, формулировании и структурировании задач на основе полученной информации
УК-2.	Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.1.	Знает действующие правовые нормы, регулирующие деятельность в области решения задач, основные методы и подходы к определению круга задач
		УК-2.2.	Умеет определять круг задач в рамках поставленной цели, выбирать оптимальные способы решения задач, учитывая имеющиеся ресурсы и ограничения
		УК-2.3.	Имеет практический опыт применения знаний о правовых нормах и ресурсах в реальных ситуациях, разработки и реализации решений в соответствии с установленными ограничениями
ОПК-1.	Способен находить, формулировать и решать актуальные и значимые проблемы прикладной и компьютерной математики	ОПК-1.1.	Знает основные методы и подходы к решению задач прикладной и компьютерной математики, включая алгоритмы, математическое моделирование и теорию оптимизации, а также современные инструменты и технологии, используемые в этой области

		ОПК-1.2.	Умеет анализировать и формулировать математические задачи, применять соответствующие методы и алгоритмы для их решения, а также интерпретировать и представлять результаты в понятной и доступной форме
		ОПК-1.3.	Имеет практический опыт работы над проектами или исследованиями в области прикладной и компьютерной математики, включая участие в конкурсах, олимпиадах или научных публикациях, где были решены актуальные и значимые задачи
ПК-1.	Способен формулировать задачи с математической точностью, обосновывать утверждения строго и анализировать полученные результаты в области математики и компьютерных наук	ПК-1.1.	Знает методы и подходы к формулированию задач, а также основные принципы математического доказательства и анализа результатов.
		ПК-1.2.	Умеет корректно ставить и формулировать математические задачи, применять строгие методы доказательства и анализировать полученные результаты.
		ПК-1.3.	Имеет опыт работы с задачами в области математики и компьютерных наук, включая применение математических методов для решения практических задач
ПК-2.	Способен решать типовые задачи профессиональной деятельности в области искусственного интеллекта, опираясь на информационную и библиографическую культуру, используя информационно-коммуникационные технологии и учитывая основные требования информационной безопасности	ПК-2.1.	Знает основы информационной и библиографической культуры, а также принципы информационной безопасности и применения информационно-коммуникационных технологий в профессиональной деятельности
		ПК-2.2.	Умеет эффективно использовать информационно-коммуникационные технологии для решения стандартных задач профессиональной деятельности, учитывая требования информационной безопасности
		ПК-2.3.	Имеет опыт работы с информационными ресурсами и технологиями в области искусственного интеллекта, включая соблюдение норм информационной безопасности

3. Тематический план

№п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостоятельная работа	
		Лекции	Семинары			
1	Атаки на системы ИИ	18	18	4	8	Домашние задания, Коллоквиум
2	Защита систем ИИ	4	4	2	8	Домашние задания, Коллоквиум
3	Современные вызовы и автоматизация в безопасности ИИ	8	8	2	16	Домашние задания, Контрольная работа
	Экзамен			4		
	Итого:	***30***	***30***	***12***	***118***	
Объем дисциплины (модуля) (в ак. ч.)		***190***				
Объем дисциплины (модуля) (в зач. ед.)		***5***				

4. Содержание дисциплины (модуля)

№п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Атаки на системы ИИ	Наступательный ИИ (атаки) Защита с помощью ИИ Зловредные воздействия Введение в атаки на системы ИИ Схемы атак на системы ИИ Атаки отравления данных и моделей Атаки отравления – бэкдоры Атаки уклонения – белый ящик Атаки уклонения – черный ящик
2	Защита систем ИИ	Атаки уклонения – защита Атаки на приватные данные
3	Современные вызовы и автоматизация в безопасности ИИ	Атаки с порождающими элементами Автоматизация атак (фреймворки) Атаки на большие языковые и мультимодальные модели

5. Учебно-методическое обеспечение

Университет располагает полным набором лицензионного и свободно распространяемого программного обеспечения, включая продукты отечественного производства.

Каждый студент в течение всего периода обучения получает индивидуальный неограниченный доступ к электронно-библиотечной системе и электронной информационно-образовательной среде университета. Эти системы предоставляют возможность доступа к ресурсам из любой точки, где есть подключение к сети Интернет, как на территории университета, так и за его пределами.

Студентам обеспечен удаленный доступ к современным профессиональным базам данных и информационным справочным системам.

Основная литература:

1. Козырь, Н. С. Информационные технологии в кибербезопасности : учебник для вузов / Н. С. Козырь, А. М. Левченко. — Москва : Издательство Юрайт, 2026. — 217 с. — (Высшее образование). — ISBN 978-5-534-22095-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/600794.

2. Платонов, В. В. Технологии машинного обучения в кибербезопасности : учебное пособие / В. В. Платонов. - Москва ; Вологда : Инфра-Инженерия, 2024. - 140 с. - ISBN 978-5-9729-2048-8. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2170891.

Дополнительная литература:

1. Партыка, Т. Л. Информационная безопасность : учебное пособие / Т.Л. Партыка, И.И. Попов. — 5-е изд., перераб. и доп. — Москва : ФОРУМ : ИНФРА-М, 2021. — 432 с. — (Среднее профессиональное образование). - ISBN 978-5-00091-473-1. - Текст : электронный. - URL: https://znanium.ru/catalog/product/1189328.

6. Материально-техническое обеспечение

Университет располагает материально-технической базой, соответствующей действующим противопожарным правилам и нормам и обеспечивающей проведение всех видов дисциплинарной и междисциплинарной подготовки, практической и научно-исследовательской работ обучающихся, предусмотренных учебным планом.

Помещения, которые представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского (практического) типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования. Помещения укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Изучение дисциплины (модуля) обеспечивается в учебных аудиториях, оснащенных:

- столами и стульями;
- компьютерной техникой;
- механическими калькуляторами;
- специализированным оборудованием, включая демонстрационное оборудование.

Помещения для самостоятельной работы обучающихся, в том числе приспособленные для использования инвалидами и лицами с ограниченными возможностями здоровья, оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду Университета.

Обучающимся предоставляется доступ (в том числе удаленный) к ресурсам информационно-телекоммуникационной сети «Интернет», электронным ресурсам (в том числе электронным библиотечным системам, современным профессиональным базам данных и информационным справочным системам):

№	Наименование портала (издания, курса, документа)	Ссылка
1.	Научная электронная библиотека elibrary.ru библиотека	https://elibrary.ru/defaultx.asp
2.	База данных для IT-специалистов	https://habr.com
3.	База данных ScienceDirect	https://www.sciencedirect.com
4.	Официальный сайт Министерства науки и высшего образования Российской Федерации	https://minobrnauki.gov.ru/
5.	Федеральный портал «Российское образование"	https://www.edu.ru/
6.	Информационная система "Единое окно доступа к образовательным ресурсам"	http://window.edu.ru/
7.	Единая коллекция цифровых образовательных ресурсов	http://school-collection.edu.ru/
8.	Федеральный центр информационно - образовательных ресурсов	http://fcior.edu.ru/

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), в том числе комплект лицензионного программного обеспечения, современные профессиональные базы данных и информационные справочные системы:

Наименование ПО	Производство	Лицензионное / свободно распространяемое
Операционные системы:		
Microsoft Imagine (Windows Client, Server)	зарубежное	лицензионное
Браузеры:		
Яндекс.Браузер	отечественное	свободно распространяемое
Google Chrome	зарубежное	свободно распространяемое
Офисные приложения:		
Microsoft Imagine (Visio, OneNote)	зарубежное	лицензионное
TeXstudio	зарубежное	свободно распространяемое
Adobe Acrobat Reader	зарубежное	свободно распространяемое
Программное обеспечение для планирования и учета времени:		
Toggle app	зарубежное	свободно распространяемое
Системы управления проектами:		
Microsoft Imagine (Project)	зарубежное	лицензионное
Системы управления базами данных:		
Microsoft Imagine (SQL Server)	зарубежное	лицензионное
Системы резервного копирования (backup):		
Acronis Backup Advanced for HyperV	зарубежное	лицензионное
Справочно-правовые системы:		
КонсультантПлюс: справочно-правовая система	отечественное	лицензионное
Средства антивирусной защиты:		
Kaspersky Endpoint Security для бизнеса Стандартный Russian Edition	отечественное	лицензионное
Среды разработки:		
Visual Studio Code	зарубежное	свободно распространяемое
Bash (Unix shell)	зарубежное	свободно распространяемое

Anaconda	зарубежное	свободно распространяемое
Robotic Operating System	зарубежное	свободно распространяемое
CopelliaSim	зарубежное	свободно распространяемое
Google Colaboratory	зарубежное	свободно распространяемое
Пакеты программных средств и библиотек:		
AutoPsy	зарубежное	свободно распространяемое
Interactive Disassembler (IDA)	зарубежное	свободно распространяемое
Системы управления библиографической информацией:		
Zotero	зарубежное	свободно распространяемое
Сервисы и службы:		
Bind	зарубежное	свободно распространяемое
Docker	зарубежное	свободно распространяемое

7. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Безопасность систем ИИ» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары, домашние задания, контрольная работа, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков профессиональной лексики, закрепление практических профессиональных компетенций, поощрение инициатив.

Лекция – систематическое, последовательное, монологическое изложение преподавателем учебного материала, как правило, теоретического характера.

В процессе лекций рекомендуется вести конспект лекций: кратко и схематично фиксировать основные идеи, выводы и обобщения лекции; выделять важные мысли, ключевые слова и термины. Необходимо отметить вопросы или материалы, которые вызывают затруднения, и попытаться найти ответы в рекомендованной литературе. Если разобраться в материале не удается, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

Семинар – это форма учебной деятельности, проводимая в учебном заведении под руководством преподавателя, где студенты активно участвуют в обсуждениях, практических заданиях и других формах взаимодействия.

Для успешной подготовки к семинару рекомендуется заранее ознакомиться с темой занятия и основными материалами, чтобы иметь возможность активно участвовать в обсуждении. Также полезно подготовить вопросы и идеи для обсуждения, что поможет глубже понять материал и продемонстрировать заинтересованность.

Аудиторная работа – активная работа студента на семинаре, его ответы на вопросы преподавателя и участие в дискуссии.

Для успешного участия в семинаре студентам рекомендуется заранее ознакомиться с темой обсуждения, прочитать необходимые материалы и подготовить вопросы. Важно активно слушать и вовлекаться в дискуссию, высказывая свои мнения и аргументируя их. При ответах на вопросы преподавателя стоит быть уверенным, четким и логичным, опираясь на изученный материал. Также полезно поддерживать диалог с однокурсниками, чтобы обогатить обсуждение и расширить свои знания.

Домашнее задание – набор задач по темам недели.

При работе над домашними заданиями важно внимательно ознакомиться с требованиями и сроками выполнения. Рекомендуется разбивать задания на этапы, чтобы избежать перегрузки и лучше усвоить материал. Использовать различные источники

информации, включая учебники и онлайн-ресурсы, для более глубокого понимания темы.

Контрольная работа – письменная работа с набором задач, которые нужно решить за ограниченное время.

Цель контрольной работы – получить специальные знания по одной или нескольким темам дисциплины и продемонстрировать навыки их практического применения.

Коллоквиум – устные ответы на вопросы, список которых известен студенту заранее.

В процессе подготовки к коллоквиуму необходимо проанализировать учебные материалы, ознакомившись с лекциями, учебниками и дополнительными источниками, акцентируя внимание на ключевых темах. Рекомендуется создать структурированные конспекты, выделяя основные идеи, термины и формулы.

Самостоятельная работа – работа студентов, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы студенты взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи студента включают работу с конспектами лекций (обработка текста), повторное изучение учебных материалов планов и тезисов ответов, изучение дополнительных тем, выполнение учебно-исследовательских заданий и другое.

Система оценивания результатов обучения по дисциплине (модулю)

Критерии получения уровня и оценивания сформированности компетенций по дисциплине (модулю) «Безопасность систем ИИ»

Оценивание уровня учебных достижений, обучающихся по дисциплине (модулю), осуществляется в виде текущего контроля успеваемости и промежуточной аттестации.

Промежуточная аттестация по дисциплине (модулю) осуществляется в форме ***экзамена***, при этом проводится оценка компетенций, сформированных по дисциплине.

Для оценивания текущего контроля успеваемости и промежуточной аттестации используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом, если иное не указано в силлабусе дисциплины:

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
10	Отлично	Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину. Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя ключевые моменты и устанавливая причинно-следственные связи. Четко формулирует ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
9	Отлично	
8	Отлично	
7	Хорошо	Студент обладает знаниями предмета почти в полном

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
6	Хорошо	объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы, акцентируя внимание на наиболее значимых моментах. Он умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя его ключевые аспекты и устанавливая причинно-следственные связи. Формулирует свои ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные ситуационные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
5	Удовлетворительно	Студент обладает базовыми знаниями по дисциплине (модулю), но испытывает трудности при самостоятельных ответах и использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути вопросов. Студент способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
4	Удовлетворительно	
3	Не сдан	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	
1	Не сдан	

Дисциплина (модуль) «Безопасность систем ИИ» оценивается следующим образом:

Активность	Вес	Количество	Описание
Домашние задания	30%	12	Набор задач по темам недели
Коллоквиум	30%	3	Устные ответы на вопросы, список которых известен студенту заранее
Контрольная работа	20%	1	Письменная работа с набором задач, которые нужно решить за ограниченное время
Экзамен	20%	1	Письменная или устная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

Формула расчёта итоговой оценки по дисциплине (модулю) «Безопасность систем ИИ»: «0,3 × среднее за домашние задания + 0,3 × среднее за коллоквиум + 0,2 × контрольная работа + 0,2 × экзамен».

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание 1.

1. Опишите схему атаки отравления данных и приведите пример её применения против модели классификации изображений;

2. Объясните, как атака с бэкдором может быть внедрена в модель машинного обучения и как она активируется во время инференса;

3. Реализуйте простую атаку уклонения (evasion attack) в белом ящике с использованием метода FGSM против обученной модели на наборе данных MNIST или CIFAR-10;

4. Сравните принципы работы атак уклонения в условиях белого и чёрного ящика, укажите различия в доступе к модели и требованиях к информации;

5. Приведите пример реальной кибератаки, в которой использовался искусственный интеллект, и проанализируйте её архитектуру и цели.

Домашнее задание 2.

1. Смоделируйте атаку с использованием порождающей модели (например, GAN), создающей враждебные примеры, и оцените её эффективность против классификатора;

2. Используйте один из открытых фреймворков (например, ART, Foolbox или CleverHans) для автоматизации генерации атак на модель машинного обучения;

3. Проведите анализ уязвимости большой языковой модели (LLM) к prompt-инъекциям и опишите возможные сценарии эксплуатации;

4. Исследуйте, как мультимодальные модели (например, CLIP) могут быть подвержены атакам, сочетающим текстовые и визуальные триггеры;

5. Подготовьте отчёт, в котором обобщите риски безопасности, связанные с автоматизацией атак на ИИ, и предложите стратегии противодействия.

Домашнее задание 3.

1. Реализуйте метод защиты модели от атак уклонения с помощью adversarial training на примере простой нейросети;

2. Опишите, как работает механизм дифференциальной приватности при обучении модели и какие риски он снижает;

3. Примените метод feature squeezing для детекции и отклонения враждебных примеров и оцените его эффективность;

4. Объясните, как использование ансамблей моделей может повысить устойчивость к атакам и снизить уязвимость отдельных компонентов;

5. Проанализируйте, какие меры защиты могут быть применены для предотвращения атак на приватные данные (например, атак типа membership inference).

Примерные вопросы к коллоквиуму

Коллоквиум 1.

1. Что такое наступательный ИИ и какие основные цели преследуют атаки с его использованием?

2. В чём заключается разница между защитой с помощью ИИ и использованием ИИ в наступательных целях?

3. Опишите общую схему зловредного воздействия на систему искусственного интеллекта.

4. Какие этапы включает атака отравления данных и как она влияет на процесс обучения модели?

5. Что такое бэкдор-атака в контексте моделей машинного обучения? Приведите пример.

6. Объясните принцип работы атаки уклонения (evasion attack) в условиях белого ящика.

7. Какие данные и модели требуются для реализации атаки уклонения в чёрном ящике?

8. В чём заключается разница между атаками в белом и чёрном ящике с точки зрения доступа и сложности реализации?

9. Как атака отравления модели может повлиять на её поведение в продакшене?

10. Приведите пример реальной системы, подвергшейся атаке на ИИ, и опишите использованный вектор атаки.

Коллоквиум 2.

1. Какие методы защиты от атак уклонения наиболее эффективны и как они работают?

2. Что такое adversarial training и как он помогает повысить устойчивость модели?

3. Опишите принцип работы feature squeezing как метода детекции враждебных примеров.

4. Как ансамбли моделей могут снизить уязвимость системы к атакам?

5. Что такое дифференциальная приватность и как она защищает приватные данные при обучении моделей?

6. Какие риски связаны с атаками типа membership inference и какие меры позволяют их снизить?

7. Как можно обнаружить наличие бэкдора в предобученной модели?

8. Какие ограничения имеет метод adversarial training при защите от атак в чёрном ящике?

9. Почему важно проводить аудит моделей перед их развертыванием в продакшене?

10. Какие инструменты (например, ART, CleverHans) используются для тестирования безопасности моделей ИИ?

Примерные задания для контрольной работы

Контрольная 1.

1. Опишите, что понимается под наступательным ИИ и приведите два примера его использования в кибератаках;

2. Объясните, в чём заключается разница между атаками уклонения в условиях белого и чёрного ящика;

3. Что такое атака отравления данных? Как она влияет на процесс обучения модели машинного обучения;

4. Опишите схему бэкдор-атаки и приведите пример её реализации в задаче распознавания изображений;

5. Какие признаки могут указывать на наличие вредоносного триггера в предобученной модели;

6. Объясните, как злоумышленник может использовать ИИ для автоматизации атак на системы с машинным обучением;

7. Приведите пример реальной уязвимости в системе ИИ, связанной с отравлением модели, и опишите последствия;

8. В чём заключается зловредное воздействие на систему ИИ и какие компоненты подвержены атаке;

9. Какие данные и модельные архитектуры особенно уязвимы к атакам в режиме чёрного ящика;

10. Объясните, почему использование открытых наборов данных может повысить риски атак отравления.

Контрольная 2.

1. Опишите принцип работы обучения на враждебных примерах как метода защиты от атак уклонения;

2. Что такое сжатие признаков и как оно помогает в обнаружении враждебных примеров;

3. Объясните, как использование нескольких моделей (ансамбля) может повысить устойчивость к атакам на системы искусственного интеллекта;

4. Как дифференциальная приватность защищает конфиденциальные данные при обучении моделей? Укажите одно ограничение этого подхода;

5. Опишите, как работает атака membership inference и какие данные она может раскрыть;

6. Какие меры можно применить для защиты модели от атак на конфиденциальность обучающих данных;

7. Объясните, зачем проводится аудит моделей перед их развертыванием в производственной среде;

8. Как можно обнаружить наличие скрытого канала (бэкдора) в предварительно обученной модели? Приведите два подхода;

9. Почему важно тестировать модель на устойчивость к враждебным примерам до её внедрения в реальные системы;

10. Объясните, в чём заключается разница между пассивной и активной защитой моделей искусственного интеллекта.

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1	Определите, что понимается под наступательным ИИ в контексте кибербезопасности	использование искусственного интеллекта для организации и автоматизации кибератак	ОПК-1
2	Объясните, в чём заключается разница между защитой с помощью ИИ и использованием ИИ в наступательных целях	защита направлена на обнаружение и предотвращение угроз, наступательный ИИ — на эксплуатацию уязвимостей	ОПК-1
3	Опишите, что такое зловредное воздействие на систему искусственного интеллекта	преднамеренное искажение поведения модели с помощью вредоносных данных или алгоритмов	ОПК-1
4	Укажите основные цели атак на системы ИИ	нарушение целостности, доступности или конфиденциальности модели или данных	ОПК-1
5	Объясните, как работает атака отравления данных и на каком этапе она применяется	модифицируются обучающие данные с целью искажения поведения модели во время обучения	УК-1
6	Опишите схему бэкдор-атаки и как она активируется во время инференса	модель корректно работает, но при наличии определённого триггера выдаёт заданное злоумышленником поведение	УК-1

7	В чём разница между атаками уклонения в условиях белого и чёрного ящика	в белом ящике атакующий знает архитектуру и параметры модели, в чёрном — только входы и выходы	УК-1
8	Назовите два типа атак, относящихся к схемам атак на системы ИИ	атаки отравления и атаки уклонения	УК-1
9	Объясните, как метод обучения на враждебных примерах помогает защититься от атак уклонения	модель обучается на модифицированных входах, что повышает её устойчивость к малым вредоносным изменениям	УК-2
10	Опишите, как работает атака по определению членства и какие данные она может скомпрометировать	позволяет определить, был ли конкретный объект в обучающей выборке, угрожая приватности данных	УК-2
11	Что такое сжатие признаков и как оно используется для защиты от враждебных примеров	метод предварительной обработки входных данных, снижающий чувствительность модели к малым возмущениям	УК-2
12	Объясните, как дифференциальная приватность защищает приватные данные при обучении модели	добавляет шум в процесс обучения, чтобы невозможно было восстановить информацию об отдельных примерах	УК-2
13	Приведите пример использования порождающей модели (например, GAN) в атаке на систему ИИ	генерация враждебных примеров, неотличимых от реальных, для обмана классификатора	ПК-1
14	Назовите известный фреймворк для автоматизации атак на модели машинного обучения	cleverhans, art (adversarial robustness toolbox) или foolbox	ПК-1
15	Опишите, какие уязвимости имеют большие языковые модели к prompt-инъекциям	возможность перехвата управления моделью через специально сформулированный запрос, обход	ПК-1

		инструкций	
16	Объясните, как мультимодальные модели могут быть атакованы с помощью комбинированных триггеров	злоумышленник использует пару текст-изображение, вызывающую нежелательное поведение модели	ПК-1
17	Укажите, какие меры можно применить для защиты модели от атак отравления	аудит обучающих данных, использование методов очистки, обучение на защищённых выборках	ПК-2
18	Назовите инструмент, позволяющий тестировать модель на устойчивость к враждебным атакам	art (adversarial robustness toolbox)	ПК-2
19	Объясните, зачем проводится аудит предобученных моделей перед их использованием	для выявления скрытых бэкдоров, уязвимостей и признаков отравления	ПК-2
20	Опишите, почему важно оценивать безопасность модели не только до, но и после развертывания	угрозы могут проявляться только в условиях реальной эксплуатации и динамической среды	ПК-2