

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО

**Директор физтех-школы
прикладной математики и
информатики**

А.М. Райгородский

	Рабочая программа дисциплины (модуля)
по дисциплине:	Безопасный искусственный интеллект
по направлению:	Информатика и вычислительная техника
профиль подготовки:	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра проблем передачи информации и анализа данных
курс:	1
квалификация:	магистр

Семестр, формы промежуточной аттестации: 1 (осенний) - Дифференцированный зачет

Аудиторных часов: 30 всего, в том числе:

лекции: 15 час.

семинары: 15 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 15 час.

Всего часов: 45, всего зач. ед.: 1

Программу составил: А.П. Кулешов, д-р техн. наук, профессор

Программа обсуждена на заседании кафедры проблем передачи информации и анализа данных 13.01.2025

Аннотация

Курс предназначен для ознакомления студентов с наукой защиты информации. В курсе рассказываются основные понятия и основные направления науки. Рассматриваются основные виды угроз и нарушителей, причины уязвимости информационных систем. Рассматриваются вопросы построения информационной структуры в РФ, нормативное и правовое регулирование защиты информации в РФ. Так же рассматривается криптографическая защита данных и модели ИБ. Для успешного прохождения курса необходимо посещение и конспектирование лекций, выполнение самостоятельная работа с дополнительными литературными источниками.

1. Цели и задачи

Цель дисциплины

Ознакомление студентов магистратуры с современными подходами в области доверенного ис-кусственного интеллекта (ДИИ).

Задачи дисциплины

- Конфиденциальность в машинном обучении. Модели угроз (например, кража данных, отравление данных, вывод о принадлежности к обучающей выборке и т. д.). Атаки на федера-тивное машинное обучение (в таких модальностях, как изображения, естественный язык, ви-део, тексты). Дифференциальная приватность для защиты машинного обучения. Обеспечение соблюдения правил с гарантиями (например, посредством доказуемой минимизации данных).
- Надежность в глубоком обучении. Состязательные атаки и защита моделей глубокого обу-чения. Автоматизированная сертификация моделей глубокого обучения (охватывает основные тенденции: выпуклые релаксации и методы ветвей и границ, а также случайное сглаживание). Сертифицированное обучение глубоких нейронных сетей.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области информатики и вычислительной техники, способен на научном языке формулировать профессиональные задачи	ОПК-2.1 Имеет представление о современном состоянии исследований в рамках тематической области своей профессиональной деятельности
	ОПК-2.2 Способен оценивать актуальность исследований в области информатики и вычислительной техники и их практическую значимость
	ОПК-2.3 Владеет профессиональной терминологией, используемой в современной научно-технической литературе, обладает навыками устного и письменного изложения результатов научной деятельности в рамках профессиональной коммуникации
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности основные законы естествознания, современный математический аппарат и алгоритмы, современные информационно-коммуникационные технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационных технологий, владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности
	ПК-2.3 Имеет практический опыт научно-исследовательской деятельности в области информационно-коммуникационных технологий

ПК-1 Готов к включению в профессиональное сообщество; способен проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности	ПК-1.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации; владеет навыками подготовки научных обзоров, публикаций, рефератов и библиографий по тематике проводимых исследований на русском и английском языке
	ПК-1.2 Умеет решать научные задачи с пониманием существующих подходов к верификации моделей программного обеспечения в связи с поставленной целью и в соответствии с выбранной методикой

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- модели угроз в рамках уязвимостей «белого», «серого» и «черного» ящика;
- методы, используемые в задачах дифференциальной приватности, случайном сглаживании;
- методы быстрого градиентного знака, проектируемый градиентный спуск, базовый итерационный метод, атаки Carlini-Wagner, распределенно-сопоставительная атака;
- методы цифровой маркировки нейросетей путем создания триггерных датасетов;
- методы цифровой маркировки генеративного контента;
- методы, обеспечивающие интерпретируемость нейросетей;
- защитные методы для обнаружения и классификации сопоставительных атак, включая эвристические и сертифицированные средства защиты.

уметь:

- обеспечить гарантированную устойчивость моделей машинного обучения в отношении возмущений во входных данных;
- выявлять отдельные классы атак на системы MlaaS (машинное обучение как сервис);
- пользоваться современными подходами цифровой маркировки нейросетей и данных;
- составлять примеры сопоставительности в обработке звука в задачах «речь-в-текст».

владеть:

- компонентами фреймворков adversarial-robustness-toolbox, foolbox, cleverhans для реализации атак и защит в рамках курса.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Сопоставительные атаки и защиты, ослабление ограничений	2	1		2
2	Модели угроз и типы атак/защит. Атаки на основе вывода о принадлежности к обучающей выборке и защита конфиденциальности данных	3	4		1
3	Интерпретируемость. Интерпретация решений модели: на основе примеров. LLM.	2	3		4
4	Противодействие атакам на голосовую биометрию. Примеры и задачи. Безопасные инструменты и генеративные приложения ИИ.	3	2		3

5	Извлечение модели и защита от кражи функциональности. Проблемы извлечения модели в облачных платформах. Доказуемая цифровая маркировка нейросетей	1	2		3
6	Перспективные методы защиты каналов связи. Устойчивость нейросетевых кодеков сжатия видеоданных и атаки.	4	3		2
Итого часов		15	15		15
Подготовка к экзамену		0 час.			
Общая трудоёмкость		45 час., 1 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 1 (Осенний)

1. Состязательные атаки и защиты, ослабление ограничений

Основные модели угроз для ИИ-систем. Типы атак и механизмов защиты в машинном обучении. Состязательные атаки и их применение. Методы ослабления ограничений атак и противодействие им.

2. Модели угроз и типы атак/защит. Атаки на основе вывода о принадлежности к обучающей выборке и защита конфиденциальности данных

Атаки на основе вывода о принадлежности к обучающей выборке (Membership Inference Attacks). Методы защиты данных: дифференциальная приватность, федеративное обучение. Атаки на размытие границ между приватными и публичными данными. Практические кейсы атак на приватность.

3. Интерпретируемость. Интерпретация решений модели: на основе примеров. LLM.

Интерпретация решений моделей на основе примеров. Анализ чувствительности модели и методы визуализации принятия решений. Интерпретация решений LLM (Large Language Models). Связь интерпретируемости и безопасности: угроза атакам на интерпретируемые модели.

4. Противодействие атакам на голосовую биометрию. Примеры и задачи. Безопасные инструменты и генеративные приложения ИИ.

Типы атак на голосовую биометрию: имитация, спуфинг, переобучение. Методы защиты биометрических систем: антиспуфинг. Генеративные модели для атаки и защиты голосовых биометрических систем. Безопасные инструменты для голосовой биометрии и генеративных приложений ИИ.

5. Извлечение модели и защита от кражи функциональности. Проблемы извлечения модели в облачных платформах. Доказуемая цифровая маркировка нейросетей

Методы извлечения моделей машинного обучения. Атаки на модели с использованием API облачных платформ. Методы защиты от кражи модели: дистилляция, watermarking. Практические кейсы атак и защиты от извлечения моделей.

6. Перспективные методы защиты каналов связи. Устойчивость нейросетевых кодеков сжатия видеоданных и атаки.

Цифровая водяная маркировка нейросетей и контента (digital watermarking). Доказуемые методы идентификации нейросетей. Методы обнаружения кражи модели в облачных сер-висах. Перспективы развития цифровой идентификации моделей ИИ. Методы защиты передачи данных в ИИ-системах. Устойчивость нейросетевых видеокодеков к атакам.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система) или личный ноутбук. Рекомендуется использование общедоступных сервисов вида Colab Google, Kaggle.

6.Перечень рекомендуемой литературы

Основная литература

Литература кафедры:

1. Adversarial Learning and Secure AI 9781009315647.
2. Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations, 2024 <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf> 10.6028/NIST.AI.100-2e2023.
- .

Дополнительная литература

Литература кафедры:

1. Nicholas Carlini and David A. Wagner. “Towards Evaluating the Robustness of Neural Networks”. In: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017. IEEE Computer Society, <https://arxiv.org/abs/1608.04644>.
2. Clear: Character unlearning in textual and visual modalities, ACL 2025 in press <https://arxiv.org/pdf/2410.18057>.
3. AASIST3: KAN- Enhanced AASIST Speech Deepfake Detection using SSL Features and Additional Regularization for the ASVspoof 2024 Challenge https://www.isca-archive.org/asvspoof_2024/borodin24_asvspoof.html.
4. Certification of Speaker Recognition Models to Additive Perturbations D Korzh, E Karimov, M Pautov, OY Rogov, I Oseledets, AAAI 2025 Proceedings.
5. Probabilistically Robust Watermarking of Neural Networks M Pautov, N Bogdanov, S Pyatkin, O Rogov, I Oseledets.
6. Quantum Cryptography and Machine Learning: Enhancing Security in AI Systems DOI: 10.4018/979-8-3693-5961-7.ch006.
7. Machine Learning Method with Applications in Hardware Security of Post-Quantum Cryptography <https://link.springer.com/article/10.1007/s10723-023-09643-4>.

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

- 1)<https://cvpr24-advml.github.io/>
- 2)<https://iclr.cc/virtual/2021/workshop/2127>

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

Рекомендуется использовать ОС Linux и фреймворк PyTorch для реализации практических задач.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, изучающий дисциплину, должен, с одной стороны, овладеть общими понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

В результате изучения дисциплины студент должен знать основные определения, понятия, методы доказательств.

Успешное освоение курса требует напряженной самостоятельной работы студента. В программе курса отведено минимально необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- чтение и конспектирование рекомендованной литературы, материалов конференций А/А* по рекомендации преподавателя;
- проработку учебного материала (по конспектам занятий, учебной и научной литературе), подготовку ответов на вопросы, предназначенные для самостоятельного изучения, доказательство отдельных утверждений, свойств, решение задач;
- подготовка к дифференцированному зачёту.

Руководство и контроль за самостоятельной работой студента осуществляется в форме индивидуальных консультаций.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов следует обращаться за консультациями к лектору.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Информатика и вычислительная техника
профиль подготовки:	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра проблем передачи информации и анализа данных
курс:	<u>1</u>
квалификация:	магистр
Семестр, формы промежуточной аттестации: 1 (осенний) - Дифференцированный зачет	
Разработчик:	А.П. Кулешов, д-р техн. наук, профессор

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области информатики и вычислительной техники, способен на научном языке формулировать профессиональные задачи	ОПК-2.1 Имеет представление о современном состоянии исследований в рамках тематической области своей профессиональной деятельности
	ОПК-2.2 Способен оценивать актуальность исследований в области информатики и вычислительной техники и их практическую значимость
	ОПК-2.3 Владеет профессиональной терминологией, используемой в современной научно-технической литературе, обладает навыками устного и письменного изложения результатов научной деятельности в рамках профессиональной коммуникации
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности основные законы естествознания, современный математический аппарат и алгоритмы, современные информационно-коммуникационные технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационных технологий, владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности
	ПК-2.3 Имеет практический опыт научно-исследовательской деятельности в области информационно-коммуникационных технологий
ПК-1 Готов к включению в профессиональное сообщество; способен проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности	ПК-1.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации; владеет навыками подготовки научных обзоров, публикаций, рефератов и библиографий по тематике проводимых исследований на русском и английском языке
	ПК-1.2 Умеет решать научные задачи с пониманием существующих подходов к верификации моделей программного обеспечения в связи с поставленной целью и в соответствии с выбранной методикой

2. Показатели оценивания компетенций

В результате изучения дисциплины «Безопасный искусственный интеллект» обучающийся должен:

знать:

- модели угроз в рамках уязвимостей «белого», «серого» и «черного» ящика;
- методы, используемые в задачах дифференциальной приватности, случайном сглаживании;
- методы быстрого градиентного знака, проектируемый градиентный спуск, базовый итерационный метод, атаки Carlini-Wagner, распределенно-сопоставительная атака;
- методы цифровой маркировки нейросетей путем создания триггерных датасетов;
- методы цифровой маркировки генеративного контента;
- методы, обеспечивающие интерпретируемость нейросетей;
- защитные методы для обнаружения и классификации сопоставительных атак, включая эвристические и сертифицированные средства защиты.

уметь:

- обеспечить гарантированную устойчивость моделей машинного обучения в отношении возмущений во входных данных;
- выявлять отдельные классы атак на системы MlaaS (машинное обучение как сервис);
- пользоваться современными подходами цифровой маркировки нейросетей и данных;
- составлять примеры сопоставительности в обработке звука в задачах «речь-в-текст».

владеть:

– компонентами фреймворков adversarial-robustness-toolbox, foolbox, cleverhans для реализации атак и защит в рамках курса.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

С целью контроля освоения обучающимися учебного материала проводится опрос в начале занятия по теме прошлой лекции или в конце занятия по пройденной теме.

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

Перечень вопросов:

Введение в безопасность искусственного интеллекта

1. Какие основные модели угроз существуют для ИИ-систем.
2. Чем состязательные атаки отличаются от других типов атак в машинном обучении.
3. Какие методы защиты могут применяться против состязательных атак.
4. Каковы основные способы ослабления ограничений атакующих в ИИ-системах.
5. Какие виды атак угрожают конфиденциальности данных в обученных моделях.

Атаки на конфиденциальность данных

6. Что такое атаки на основе вывода о принадлежности (Membership Inference Attacks) и какие риски они создают.
7. Как дифференциальная приватность помогает защитить данные.
8. Как работает гомоморфное шифрование в контексте защиты ИИ.
9. Какие недостатки есть у федеративного обучения с точки зрения безопасности.
10. Какие методы могут применяться для детекции атак на приватность данных.

Интерпретируемость и безопасность нейросетевых моделей

11. Какие методы существуют для интерпретации решений моделей машинного обучения.
12. Почему высокая интерпретируемость может повышать уязвимость модели.
13. Как можно атаковать интерпретируемые модели.
14. В чем особенности интерпретации решений больших языковых моделей (LLM).
15. Какие метрики применяются для оценки интерпретируемости нейросетевых моделей.

Защита голосовой биометрии

16. Какие типы атак угрожают голосовой биометрии.
17. Как работает атака на биометрическую систему методом имитации.
18. В чем суть спуфинга в голосовых биометрических системах.
19. Как работают методы защиты голосовой биометрии от атак.
20. Как можно использовать генеративные модели для атаки и защиты биометрии.

Извлечение модели и защита от кражи функциональности

21. Какие основные способы извлечения моделей ИИ существуют.
22. Как API-атаки позволяют извлекать функциональность модели.
23. Какие методы защиты от извлечения модели применяются на практике.
24. В чем суть метода дистилляции как способа защиты модели.
25. Как можно обнаружить факт кражи модели.

Доказуемая цифровая маркировка нейросетей

26. Какие методы цифровой маркировки нейросетей существуют.
27. Как работают цифровые водяные знаки в контексте нейросетей.
28. Какие критерии предъявляются к эффективной защите моделей от копирования.
29. Как можно доказать авторство модели в условиях облачных вычислений.

30. В чем преимущества и недостатки цифровой маркировки нейросетей. Какие подходы вероятностно устойчивой маркировки существуют сейчас.

Пример билетов:

Билет 1

1. Какие методы защиты каналов связи используются в нейросетевых системах.
2. Какие наиболее перспективные методы защиты ИИ разрабатываются в настоящее время.
3. Как регулируется безопасность искусственного интеллекта на международном уровне.

Билет 2

1. Какие проблемы безопасности могут возникнуть при масштабировании больших моделей.
2. Какие подходы к сертификации безопасных ИИ-систем существуют.
3. Как атаки могут изменять сжатие и восстановление видеоданных.

Критерии оценивания

Оценка отлично (10) выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины, проявляющему интерес к данной предметной области, продемонстрировавшему умение уверенно и творчески применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично (9) выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично (8) выставляется студенту, показавшему систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, правильное обоснование принятых решений, с некоторыми недочетами.

Оценка хорошо (7) выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но недостаточно грамотно обосновывает полученные результаты.

Оценка хорошо (6) выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности.

Оценка хорошо (5) выставляется студенту, если он в основном знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач достаточно большое количество неточностей.

Оценка удовлетворительно (4) выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он освоил основные разделы учебной программы, необходимые для дальнейшего обучения, и может применять полученные знания по образцу в стандартной ситуации.

Оценка удовлетворительно (3) выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, допускающему ошибки в формулировках базовых понятий, нарушения логической последовательности в изложении программного материала, слабо владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и с трудом применяет полученные знания даже в стандартной ситуации.

Оценка неудовлетворительно (2) выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных принципов и не умеет использовать полученные знания при решении типовых задач.

Оценка неудовлетворительно (1) выставляется студенту, который не знает основного содержания учебной программы дисциплины, допускает грубейшие ошибки в

формулировках базовых понятий дисциплины и вообще не имеет навыков решения типовых практических задач.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Дифференцированный зачет проводится в устной форме.

При проведении устного дифференцированного зачёта обучающемуся предоставляется 30 минут на подготовку.

Во время проведения дифференцированного зачета обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, вычислительной техникой и проч.