Федеральное государственное автономное образовательное учреждение высшего образования «Российский университет дружбы народов имени Патриса Лумумбы»

Факультет искусственного интеллекта

(наименование основного учебного подразделения (ОУП)-разработчика ОП ВО)

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

БЕЗОПАСНОСТЬ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

(наименование дисциплины/модуля)

Рекомендована МССН для направлений подготовки:

02.03.02 ФУНДАМЕНТАЛЬНАЯ ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ;
09.03.03 ПРИКЛАДНАЯ ИНФОРМАТИКА

(код и наименование направления подготовки/специальности)

Освоение дисциплины ведется в рамках реализации основной профессиональной образовательной программы высшего образования (ОП ВО):

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: РАЗРАБОТКА И ОБУЧЕНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ

(наименование (профиль/специализация) ОП ВО)

2026 г.

1. ЦЕЛЬ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Дисциплина «Безопасность систем искусственного интеллекта» входит в программу бакалавриата «Искусственный интеллект: разработка и обучение интеллектуальных систем» по направлениям подготовки 02.03.02 Фундаментальная информатика и информационные технологии и 09.03.03 Прикладная информатика, и изучается в 7 семестре 4 курса. Дисциплину реализует Кафедра прикладного искусственного интеллекта. Дисциплина состоит из 3 разделов и 26 тем и направлена на изучение комплексных аспектов безопасности систем искусственного интеллекта: угроз на всех этапах жизненного цикла ML-систем (отравление данных, adversarial attacks, кража моделей, инверсия, prompt injection), методов защиты и обеспечения робастности (adversarial training, certified defenses, differential privacy, federated learning), безопасности больших языковых моделей (jailbreaking, prompt injection, toxicity, alignment), нормативно-правового регулирования ИИ (EU AI Act, российское законодательство, стандарты ISO/IEC 42001, ГОСТ), методов оценки рисков и аудита ИИ-систем, проектирования безопасной инфраструктуры ML-сервисов (управление доступом, секреты, логирование, incident response), а также формирования культуры ответственной разработки и эксплуатации ИИ-систем.

Целью освоения дисциплины является формирование у студентов системных знаний об угрозах безопасности ИИ-систем и практических навыков проектирования, тестирования и эксплуатации защищённых ML-решений, включая способность классифицировать и моделировать угрозы, применять методы adversarial robustness и privacy-preserving ML, обеспечивать безопасность LLM-систем, проводить аудит ИИ-систем на соответствие нормативным требованиям, проектировать безопасную инфраструктуру, формулировать политику безопасности и критерии приёмки с учётом специфики ML, а также оценивать риски и экономические последствия нарушений безопасности.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Освоение дисциплины «Безопасность систем искусственного интеллекта» направлено на формирование у обучающихся следующих компетенций (части компетенций):

Таблица 2.1. Перечень компетенций, формируемых у обучающихся при освоении дисциплины (результаты освоения дисциплины)

Шифр	**Компетенция**	**Индикаторы достижения компетенции** (в рамках данной дисциплины)
УК-9	Способен использовать базовые дефектологические знания в социальной и профессиональной сферах	УК-9.3 Владеет навыками применения базовых дефектологических знаний в социальной и профессиональной сферах;
ОПК-4	Способен участвовать в разработке технической документации, стандартов, норм и правил, а также в управлении проектами создания информационных систем и систем ИИ на стадиях жизненного цикла	ОПК-4.1 Знает стандарты и нормы оформления технической документации программных продуктов и систем ИИ, принципы управления жизненным циклом ИС;
ОПК-5	Способен инсталлировать и сопровождать программное и аппаратное обеспечение информационных систем и систем ИИ, в том числе отечественного происхождения, с учётом требований информационной безопасности	ОПК-5.1 Знает принципы инсталляции, конфигурирования и сопровождения программного обеспечения ИС и систем ИИ, основные требования информационной безопасности; ОПК-5.2 Умеет развёртывать и сопровождать среды разработки и эксплуатации систем ИИ (контейнеризация, оркестрация, CI/CD), обеспечивать информационную безопасность данных и моделей;

Шифр	Компетенция	Индикаторы достижения компетенции (в рамках данной дисциплины)
ПК-1	Способен анализировать требования к программному обеспечению систем ИИ, разрабатывать технические спецификации и техническое задание на систему	ПК-1.1 Анализирует возможности реализации функциональных и нефункциональных требований к ПО систем ИИ, выявляет противоречия и ограничения;
ПК-3	Способен разрабатывать и реализовывать стратегии тестирования и контроля качества программного обеспечения систем ИИ	ПК-3.1 Верифицирует требования к ПО систем ИИ, определяет требования к тестам и критерии приёмки; ПК-3.2 Разрабатывает план тестирования и организационные документы для тестирования ПО систем ИИ; ПК-3.3 Оценивает результаты тестирования, реализует процесс контроля качества ПО систем ИИ;
AI S-1	Способен управлять рисками при разработке и использовании систем ИИ, выстраивать управление безопасностью ИИ в организации с учетом принципов этического использования ИИ	AI S-1.1 Выявляет и моделирует угрозы на всём жизненном цикле ИИ-систем, оценивает и приоритизирует риски; AI S-1.2 Обеспечивает соответствие нормативным требованиям и принципам доверенного/этичного ИИ;
FC-5	Способен проводить передовые исследования в области безопасности, доверия и объяснимости	FC-5.1 Обеспечивает защиту от использования моделей искусственного интеллекта во вред человеку и обществу; FC-5.2 Обеспечивает объяснения причин принятия тех или иных решений в результатах работы искусственного интеллекта; FC-5.3 Обеспечивает отсутствие случайных или добавленных уязвимостей в системах искусственного интеллекта;
LLM-1	Способен применять и (или) разрабатывать генеративные модели и БЯМ	LLM-1.5 Оценивает защищённость моделей генерации;
ML-6	Способен применять алгоритмы обучения с подкреплением	ML-6.2 Применяет методы повышения устойчивости, надежности, безопасности алгоритмов обучения с подкреплением для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ;
SS-1	Способен учитывать философские, когнитивные и социальные основания концепций ИИ в профессиональной деятельности	SS-1.2 Применяет методики работы с этическими и социальными рисками, возникающими на разных стадиях жизненного цикла ИИ;
SS-3	Способен к критическому анализу, метарефлексии и переносу знаний при работе с системами ИИ	SS-3.1 Учитывает в работе когнитивные искажения человека и примеры их проявления при работе с данными и ИИ, выявляет предвзятости систем ИИ, аргументированно оценивает надежность данных и выдачи ИИ, применяет базовые принципы критического мышления (оценка источников, проверка аргументов, отличие факта от интерпретации);

3. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОП ВО

Дисциплина «Безопасность систем искусственного интеллекта» относится к обязательной части блока 1 «Дисциплины (модули)» образовательной программы высшего образования.

В рамках образовательной программы высшего образования обучающиеся также осваивают другие дисциплины и/или практики, способствующие достижению запланированных результатов освоения дисциплины «Безопасность систем искусственного интеллекта».

Таблица 3.1. Перечень компонентов ОП ВО, способствующих достижению запланированных результатов освоения дисциплины

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
УК-9	Способен использовать базовые дефектологические знания в социальной и профессиональной сферах	Философия; Искусственный интеллект и когнитивная психология; Этика и безопасность использования искусственного интеллекта;	
ОПК-4	Способен участвовать в разработке технической документации, стандартов, норм и правил, а также в управлении проектами создания информационных систем и систем ИИ на стадиях жизненного цикла	Эксплуатационная практика (учебная); Этика и безопасность использования искусственного интеллекта; Методы разработки решений на основе искусственного интеллекта (Git, Docker);	MLOps и промышленная разработка систем искусственного интеллекта;
ОПК-5	Способен инсталлировать и сопровождать программное и аппаратное обеспечение информационных систем и систем ИИ, в том числе отечественного происхождения, с учётом требований информационной безопасности	Введение в базы данных; Методы разработки решений на основе искусственного интеллекта (Git, Docker); Массово-параллельные вычисления в машинном обучении (GPU); Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Технологическая (проектно-технологическая) практика (учебная);	MLOps и промышленная разработка систем искусственного интеллекта;
ПК-1	Способен анализировать требования к программному обеспечению систем ИИ, разрабатывать технические спецификации и техническое задание на систему	Правоведение; Параллельное и распределенное программирование; Введение в искусственный интеллект; Искусственный интеллект и когнитивная психология; Этика и безопасность использования искусственного интеллекта; Методы машинного обучения; Массово-параллельные вычисления в машинном обучении (GPU); Основы глубокого обучения; Большие языковые модели**; История и теория программирования; Программирование на языке C++; Методы разработки решений на основе искусственного интеллекта (Git, Docker); Введение в базы данных; Онтология и графы знаний; Проектирование и разработка систем компьютерного зрения; Практикум по обработке	Преддипломная практика; Методы машинного обучения; MLOps и промышленная разработка систем искусственного интеллекта; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP);

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
		естественного языка (NLP); Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Технологическая (проектно-технологическая) практика (учебная);	
ПК-3	Способен разрабатывать и реализовывать стратегии тестирования и контроля качества программного обеспечения систем ИИ	Технологическая (проектно-технологическая) практика (учебная); Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Теория вероятностей и математическая статистика; Этика и безопасность использования искусственного интеллекта; Статистические методы и первичный анализ данных; Методы машинного обучения; Обработка и анализ изображений и видео с помощью методов искусственного интеллекта; Анализ естественного языка с помощью методов искусственного интеллекта; Программирование на языке Python; Методы разработки решений на основе искусственного интеллекта (Git, Docker); Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP);	Преддипломная практика; Методы машинного обучения; MLOps и промышленная разработка систем искусственного интеллекта; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP);
SS-1	Способен учитывать философские, когнитивные и социальные основания концепций ИИ в профессиональной деятельности	Философия; История и теория программирования; Введение в искусственный интеллект; Искусственный интеллект и когнитивная психология; Онтология и графы знаний; Методы машинного обучения; Основы глубокого обучения; Лингвистические основы анализа естественного языка; Основы робототехники**; Большие языковые модели**; Этика и безопасность использования искусственного интеллекта; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP);	Методы машинного обучения; MLOps и промышленная разработка систем искусственного интеллекта; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP);

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
SS-3	Способен к критическому анализу, метарефлексии и переносу знаний при работе с системами ИИ	Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Теория вероятностей и математическая статистика; Искусственный интеллект и когнитивная психология; Этика и безопасность использования искусственного интеллекта; Статистические методы и первичный анализ данных; Методы машинного обучения; Обработка и анализ изображений и видео с помощью методов искусственного интеллекта; Анализ естественного языка с помощью методов искусственного интеллекта; Правоведение; Введение в искусственный интеллект; Введение в компьютерное зрение; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP); Основы программирования HTML - CSS - JavaScript**; Основы программирования на языке NodeJS**; Основы программирования на языке Go**; Основы программирования на языке Julia**; Основы робототехники**; Цифровые двойники**; Философия; Большие языковые модели**;	Преддипломная практика; Методы машинного обучения; Вайб-кодинг**; MLOps и промышленная разработка систем искусственного интеллекта; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP); Обработка сигналов**; Анализ временных рядов**;
ML-6	Способен применять алгоритмы обучения с подкреплением	Методы машинного обучения; Основы робототехники**;	Преддипломная практика; Методы машинного обучения;
LLM-1	Способен применять и (или) разрабатывать генеративные модели и БЯМ	Основы глубокого обучения; Большие языковые модели**;	MLOps и промышленная разработка систем искусственного интеллекта; Вайб-кодинг**; Преддипломная практика;
AI S-1	Способен управлять рисками при разработке и использовании систем ИИ, выстраивать управление безопасностью ИИ в организации с учетом принципов этического использования ИИ	Этика и безопасность использования искусственного интеллекта; Эксплуатационная практика (производственная);	Преддипломная практика;
FC-5	Способен проводить передовые исследования в области безопасности, доверия и объяснимости	Эксплуатационная практика (производственная); Эксплуатационная практика (учебная);	Методы машинного обучения; MLOps и промышленная разработка систем

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
		Этика и безопасность использования искусственного интеллекта; Правоведение; Методы машинного обучения;	искусственного интеллекта; Преддипломная практика;

* - заполняется в соответствии с матрицей компетенций и СУП ОП ВО

** - элективные дисциплины /практики

4. ОБЪЕМ ДИСЦИПЛИНЫ И ВИДЫ УЧЕБНОЙ РАБОТЫ

Общая трудоемкость дисциплины «Безопасность систем искусственного интеллекта» составляет «3» зачетные единицы.

Таблица 4.1. Виды учебной работы по периодам освоения образовательной программы высшего образования для очной формы обучения.

Вид учебной работы	**ВСЕГО, ак.ч.**		**Семестр(-ы)**
			7
Контактная работа, ак.ч.	*52*		*52*
Лекции (**ЛК**)	26		26
Лабораторные работы (**ЛР**)	0		0
Практические/семинарские занятия (**СЗ**)	26		26
Самостоятельная работа обучающихся, ак.ч.	*38*		*38*
Контроль (экзамен/зачет с оценкой), ак.ч.	*18*		*18*
Общая трудоемкость дисциплины	**ак.ч.**	**108**	**ак.ч.**
	зач.ед.	**3**	**зач.ед.**

5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Таблица 5.1. Содержание дисциплины (модуля) по видам учебной работы

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
Раздел 1	Угрозы ML-системам и методы атак	1.1	Введение: ландшафт угроз ИИ-систем	Безопасность ИИ как междисциплинарная область: пересечение ML, кибербезопасности, этики и права. Таксономия угроз по жизненному циклу: обучение (data poisoning, backdoor), инференс (adversarial, model stealing), развёртывание (инфраструктура, API abuse). Модель угроз (threat model): возможности атакующего (white-box, black-box, grey-box), цели (нарушение целостности, конфиденциальности, доступности). MITRE ATLAS: фреймворк тактик и техник атак на ML. Обзор резонансных инцидентов: Tay chatbot, adversarial patches, prompt injection в production	ЛК	AIS-1.1, SS-1.2, ПК-1.1
		1.2	Adversarial attacks: evasion-атаки	Adversarial examples: определение, L_p нормы (L_0, L_2, L_∞). White-box атаки: FGSM (Fast Gradient Sign Method), PGD (Projected Gradient Descent), C&W (Carlini-Wagner). Black-box атаки: transfer-based (adversarial transferability), query-based (finite differences, NES). Физические adversarial examples: патчи, стикеры, 3D-объекты (adversarial glasses, stop sign attacks). Adversarial для NLP: TextFooler, BAE, character-level perturbations. Формальное определение робастности: ∀δ ∈ B_ε: f(x+δ) = f(x)	ЛК	ML-6.2, AIS-1.1
		1.3	Data poisoning и backdoor-атаки	Data poisoning: внедрение вредоносных примеров в обучающую выборку. Типы: availability poisoning (ухудшение качества), targeted poisoning (ошибка на конкретных примерах). Backdoor attacks: BadNets (триггер → целевой класс), clean-label backdoors. Trojan attacks: модификация модели. Poisoning в federated learning: модельные обновления от злоумышленника. Supply chain attacks: вредоносные предобученные модели на Hugging Face Hub / Model Zoo. Обнаружение: activation clustering, spectral signatures, Neural Cleanse	ЛК	AIS-1.1, ML-6.2, ОПК-5.1
		1.4	Атаки на конфиденциальность: model stealing и data extraction	Model stealing: клонирование модели через API-запросы (Tramèr et al., 2016). Hyperparameter stealing. Membership	ЛК	AIS-1.1, ОПК-5.1,

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				Inference Attack (MIA): определение, был ли пример в обучающей выборке. Model Inversion: восстановление обучающих данных из модели. Training Data Extraction из LLM: воспроизведение персональных данных (Carlini et al., 2021). Watermarking: защита интеллектуальной собственности модели. Связь с GDPR и правом на забвение		УК-9.3
		1.5	Безопасность LLM: prompt injection, jailbreaking, toxicity	Prompt injection: direct (вредоносный промпт от пользователя), indirect (вредоносный контент в извлекаемых документах). Jailbreaking: обход alignment через creative prompts, DAN, multi-turn attacks. Типы вредного поведения: toxicity, bias, hallucination, harmful instructions. Атаки на RAG: poisoning knowledge base. Обзор реальных инцидентов: Bing Chat, ChatGPT jailbreaks, DPD chatbot. Red teaming: систематическое тестирование LLM на безопасность	ЛК	LLM-1.5, AIS-1.2, FC-5.1
		1.6	Практикум: adversarial attacks — FGSM и PGD	Реализация FGSM на PyTorch: вычисление градиента по входу, perturbation. PGD: итеративная атака с проекцией на ε-шар. Атака на CNN (CIFAR-10): визуализация adversarial examples, clean vs. adversarial accuracy. Анализ: зависимость success rate от ε. Torchattacks library. Практика: генерация adversarial examples для предобученной модели	СЗ	ML-6.2, AIS-1.1
		1.7	Практикум: adversarial attacks на NLP	TextAttack library: рецепты атак (TextFooler, BAE, DeepWordBug). Атака на BERT classifier: замена слов → изменение предсказания. Визуализация: какие слова изменены, семантическая близость. Метрики: attack success rate, perturbation rate, semantic similarity. Практика: атака на fine-tuned BERT, анализ уязвимостей	СЗ	ML-6.2, LLM-1.5
		1.8	Практикум: membership inference attack	Реализация MIA: shadow model approach (Shokri et al., 2017). Обучение shadow models, сбор in/out predictions, обучение attack model. Оценка: AUC атаки. Факторы: overfitting увеличивает уязвимость. Практика: MIA на простой модели, анализ зависимости от размера данных и overfitting	СЗ	AIS-1.1, ОПК-5.1
		1.9	Практикум: prompt injection и jailbreaking LLM	Эксперимент: direct prompt injection на open-source LLM (local deployment). Типы: ignore previous instructions, role-playing, encoding tricks. Indirect injection: внедрение	СЗ	LLM-1.5, AIS-1.2

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				инструкций в документ для RAG. Оценка: success rate, обнаружение. Обсуждение: ответственное disclosure, этика red teaming. Практика: систематическое тестирование LLM на prompt injection		
		1.10	Практикум: threat modeling для ML-системы	STRIDE для ML: Spoofing (data poisoning), Tampering (adversarial), Repudiation (отсутствие логов), Information Disclosure (model stealing), Denial of Service (resource exhaustion), Elevation of Privilege (prompt injection). Составление threat model: assets, threats, mitigations. MITRE ATLAS mapping. Практика: threat modeling для учебного ML-проекта, формирование матрицы рисков	СЗ	AIS-1.1, ПК-1.1, ПК-3.1
Раздел 2	Методы защиты и обеспечения робастности	2.1	Adversarial robustness: training и certified defenses	Adversarial training: обучение на adversarial examples (Madry et al., PGD-AT). Trade-off: robustness vs. clean accuracy. TRADES: оптимизация trade-off. Certified defenses: randomized smoothing (Cohen et al., 2019) — гарантированный радиус робастности. Interval Bound Propagation (обзор). Практические аспекты: adversarial training дорого, certified defense консервативна. Когда adversarial robustness критична (автономное вождение, медицина) и когда — нет	ЛК	ML-6.2, FC-5.1, AIS-1.2
		2.2	Privacy-preserving ML: differential privacy и federated learning	Differential privacy (DP): ε-DP, определение, интуиция. DP-SGD: добавление шума к градиентам, клиппирование. Trade-off: privacy budget ε vs. utility. Opacus (PyTorch). Federated Learning: обучение без передачи данных, FedAvg, communication rounds. Угрозы FL: gradient inversion, model poisoning. Secure aggregation. Практические применения: медицина, финансы. Связь с GDPR и ФЗ-152	ЛК	AIS-1.2, ОПК-5.1, FC-5.2
		2.3	Защита LLM: alignment, guardrails, red teaming	Alignment: RLHF (Reinforcement Learning from Human Feedback), Constitutional AI, DPO. Guardrails: input filtering (toxicity detection, PII detection), output filtering (safety classifier), system prompt hardening. Red teaming: систематическое тестирование, automated red teaming (Perez et al.). Content moderation: OpenAI Moderation API, LlamaGuard. Watermarking LLM output: статистические паттерны для обнаружения AI-generated текста.	ЛК	LLM-1.5, FC-5.1, AIS-1.2

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				Мониторинг: обнаружение аномальных паттернов использования		
		2.4	Защита инфраструктуры ML-сервисов	Модель угроз инфраструктуры: API abuse (rate limiting, authentication), model exfiltration (access control, watermarking), data leakage (encryption at rest/transit). Secrets management: API keys, model weights, credentials (Vault, обзор). Контейнерная безопасность: сканирование образов (Trivy), минимальные образы, non-root user. Сетевая безопасность: HTTPS, mTLS, network policies. Audit logging: кто, когда, какой запрос. Compliance: SOC 2, ISO 27001 (обзор). Связь с ОПК-5	ЛК	ОПК-5.2, ОПК-5.1, ПК-3.2
		2.5	Практикум: adversarial training	PGD adversarial training на PyTorch: генерация PGD examples в каждом батче, обучение на них. Сравнение: standard training vs. adversarial training (clean accuracy, robust accuracy под PGD-20 attack). Визуализация: loss landscape (adversarially trained vs. standard). Практика: AT для CNN на CIFAR-10, анализ trade-off	СЗ	ML-6.2, AIS-1.2
		2.6	Практикум: differential privacy (Opacus)	Opacus: PrivacyEngine, make_private (model, optimizer, data_loader). Параметры: target_epsilon, max_grad_norm. Обучение с DP-SGD. Мониторинг: epsilon по эпохам (privacy accountant). Сравнение: accuracy с DP vs. без DP при разных ε. Практика: обучение классификатора с differential privacy, анализ utility-privacy trade-off	СЗ	AIS-1.2, ОПК-5.1
		2.7	Практикум: guardrails для LLM	Реализация guardrails: input validator (regex, toxicity classifier), output validator (safety check, PII scrubbing). System prompt hardening: инструкции против prompt injection. LangChain output parsers. Rate limiting. Практика: добавление guardrails к LLM-сервису (FastAPI + vLLM), тестирование на prompt injection из раздела 1	СЗ	LLM-1.5, FC-5.1
		2.8	Практикум: безопасная инфраструктура ML-сервиса	Hardening Docker: non-root user, read-only filesystem, resource limits, secrets через Docker secrets / environment. HTTPS для FastAPI (certbot). Rate limiting (slowapi). Authentication: API key / JWT. Audit logging: structured logs с request_id, user_id, prediction. Сканирование контейнера (Trivy). Практика: hardening учебного ML-сервиса по чек-	СЗ	ОПК-5.2, ПК-3.2

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				листу		
Раздел 3	Регулирование, аудит и ответственная разработка	3.1	Нормативно-правовое регулирование ИИ	EU AI Act: классификация по уровням риска (unacceptable, high, limited, minimal), требования к high-risk системам (risk management, data governance, transparency, human oversight, robustness). Российское законодательство: Национальная стратегия развития ИИ, ФЗ-149, ФЗ-152, Кодекс этики ИИ. Стандарты: ISO/IEC 42001 (AI management system), ISO/IEC 23894 (AI risk management), ГОСТ Р 59898 (ИИ. Термины). NIST AI RMF. Ответственность за вред, причинённый ИИ: текущее состояние и дискуссии	ЛК	FC-5.2, УК-9.3, ОПК-4.1
		3.2	Оценка рисков и аудит ИИ-систем	Risk assessment для ML: идентификация → анализ → оценка → митигация → мониторинг. Матрица рисков: вероятность × воздействие. Категории рисков: технические (adversarial, drift, failure), этические (bias, fairness, privacy), операционные (downtime, data loss), юридические (non-compliance). Аудит ИИ: internal vs. external, checklist-based, conformity assessment. Связь аудита с EU AI Act requirements. Algorithmic Impact Assessment (обзор)	ЛК	FC-5.3, AIS-1.1, ПК-3.3
		3.3	AI Safety: alignment, value alignment, existential risk	AI Safety как исследовательское направление: alignment problem (модель делает то, что мы хотим). Inner alignment vs. outer alignment. Reward hacking: модель оптимизирует proxy. Goal misgeneralization. Scalable oversight: как контролировать модели умнее контролёра. Existential risk: аргументы за и против (Bostrom, Russell, Marcus). AI governance: международные инициативы (Bletchley, Hiroshima). Связь с практикой: alignment в ChatGPT, Claude — ограничения текущих подходов	ЛК	AIS-1.2, SS-3.1, SS-1.2
		3.4	Культура безопасной разработки ИИ и экономика безопасности	Security by design для ML: безопасность на каждом этапе (данные, обучение, развёртывание). Shift-left security: раннее выявление проблем. DevSecOps для ML: интеграция security tests в CI/CD. Responsible disclosure: как сообщать об уязвимостях ML-систем. Экономика безопасности: стоимость инцидента vs. стоимость защиты, ROI security measures. Bug bounty для AI (обзор). Формирование security culture в ML-команде	ЛК	FC-5.3, FC-5.1, SS-1.2

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
		3.5	Практикум: оценка рисков ML-системы	Проведение risk assessment для учебного ML-проекта: идентификация активов (данные, модель, API), угроз (MITRE ATLAS), уязвимостей, митигаций. Заполнение матрицы рисков. Приоритизация: какие риски критичны, какие приемлемы. Формирование плана митигации. Документ: Risk Assessment Report	СЗ	FC-5.3, ПК-1.1, ПК-3.3
		3.6	Практикум: аудит ИИ-системы по чек-листу EU AI Act	Чек-лист для high-risk AI: risk management system, data governance, technical documentation, record-keeping, transparency, human oversight, accuracy/robustness/cybersecurity. Практика: проведение аудита учебного ML-проекта по чек-листу, формирование отчёта: compliant / partially compliant / non-compliant по каждому требованию, рекомендации	СЗ	FC-5.2, ОПК-4.1, ПК-3.1
		3.7	Практикум: формирование политики безопасности ML-системы	Политика безопасности: scope, roles, data handling policy, model access policy, incident response procedure, monitoring requirements, compliance requirements. Acceptable Use Policy для LLM (обзор OpenAI, Anthropic). Практика: составление Security Policy для ML-сервиса, включая специфичные для ML аспекты (adversarial testing, data poisoning prevention, model versioning)	СЗ	FC-5.1, ОПК-5.2, ПК-3.2
		3.8	Практикум: итоговый проект — комплексная оценка безопасности ИИ-системы	Финальная интеграция: выбор ML-системы (из предыдущих курсовых проектов) → threat modeling (STRIDE + MITRE ATLAS) → adversarial testing (evasion attacks, prompt injection) → privacy assessment (MIA, data extraction risk) → infrastructure security audit → risk assessment → compliance check (EU AI Act checklist) → формирование Security Policy + Risk Report + рекомендации по митигации. Презентация (15 мин) + отчёт (8–10 стр.). Peer review отчётов между студентами	СЗ	AIS-1.1, FC-5.3, ПК-3.3, ПК-3.1

* - заполняется только по **ОЧНОЙ** форме обучения: *ЛК – лекции; ЛР – лабораторные работы; СЗ – практические/семинарские занятия.*

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Таблица 6.1. Материально-техническое обеспечение дисциплины

Тип аудитории	**Оснащение аудитории**	**Специализированное учебное/лабораторное оборудование, ПО и материалы для освоения дисциплины (при необходимости)**
Лекционная	Аудитория для проведения занятий лекционного типа, оснащенная комплектом специализированной мебели; доской (экраном) и техническими средствами мультимедиа презентаций.	
Семинарская	Аудитория для проведения занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенная комплектом специализированной мебели и техническими средствами мультимедиа презентаций.	Персональные компьютеры, необходимое ПО
Для самостоятельной работы	Аудитория для самостоятельной работы обучающихся (может использоваться для проведения семинарских занятий и консультаций), оснащенная комплектом специализированной мебели и компьютерами с доступом в ЭИОС.	Персональные компьютеры, необходимое ПО

* - аудитория для самостоятельной работы обучающихся указывается **ОБЯЗАТЕЛЬНО**!

7. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Основная литература:

1. Митяков, Е. С. Искусственный интеллект и машинное обучение : учебное пособие для вузов / Е. С. Митяков, А. Г. Шмелева, А. И. Ладынин. — 2-е изд., стер. — Санкт-Петербург : Лань, 2026. — 252 с. — ISBN 978-5-507-51198-3. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/507451

2. Баланов, А. Н. Комплексная информационная безопасность : учебное пособие для вузов / А. Н. Баланов. — 2-е изд., стер. — Санкт-Петербург : Лань, 2025. — 400 с. — ISBN 978-5-507-52839-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/460715

Дополнительная литература:

1. Щербачева, Л. В. Правовое регулирование искусственного интеллекта в современном праве : учебное пособие для вузов / Л. В. Щербачева. — 2-е изд., стер. — Санкт-Петербург : Лань, 2026. — 140 с. — ISBN 978-5-507-56550-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/517172

Ресурсы информационно-телекоммуникационной сети «Интернет»:

1. ЭБС РУДН и сторонние ЭБС, к которым студенты университета имеют доступ

на основании заключенных договоров

- Электронно-библиотечная система РУДН – ЭБС РУДН https://mega.rudn.ru/MegaPro/Web
- ЭБС «Университетская библиотека онлайн» http://www.biblioclub.ru
- ЭБС «Юрайт» http://www.biblio-online.ru
- ЭБС «Консультант студента» www.studentlibrary.ru
- ЭБС «Знаниум» https://znanium.ru/

2. Базы данных и поисковые системы

- Sage https://journals.sagepub.com/
- Springer Nature Link https://link.springer.com/
- Wiley Journal Database https://onlinelibrary.wiley.com/
- Наукометрическая база данных Lens.org https://www.lens.org

Учебно-методические материалы для самостоятельной работы обучающихся при освоении дисциплины/модуля:*

1. Курс лекций по дисциплине «Безопасность систем искусственного интеллекта».

* - все учебно-методические материалы для самостоятельной работы обучающихся размещаются в соответствии с действующим порядком на странице дисциплины **в ТУИС**!