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

ЭТИКА И БЕЗОПАСНОСТЬ ИСПОЛЬЗОВАНИЯ 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 «Этика и безопасность использования 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 3 семестре 2 курса. Дисциплину реализует Кафедра прикладного искусственного интеллекта. Дисциплина состоит из 3 разделов и 26 тем и направлена на изучение этических принципов и нормативно-правовых основ разработки и применения систем искусственного интеллекта: философских оснований этики технологий, принципов ответственного ИИ (справедливость, прозрачность, подотчётность, безопасность), проблем алгоритмической предвзятости и дискриминации, методов обеспечения объяснимости и интерпретируемости моделей, правового регулирования ИИ (AI Act, национальные стандарты, ГОСТ), требований к документированию и аудиту ИИ-систем, социальных последствий внедрения ИИ в различных отраслях, а также методик оценки и управления этическими и социальными рисками на всех стадиях жизненного цикла ИИ-систем.

Целью освоения дисциплины является формирование у студентов системного понимания этических, правовых и социальных аспектов разработки и применения ИИ-систем, способности выявлять и анализировать этические дилеммы и ценностные конфликты при проектировании ИИ, применять методики оценки рисков алгоритмической предвзятости и дискриминации, учитывать регуляторные требования и стандарты в технической документации, проектировать ИИ-системы с учётом инклюзивности и доступности, а также осуществлять метарефлексию относительно последствий внедрения ИИ для общества и профессиональной деятельности.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Освоение дисциплины «Этика и безопасность использования 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-11	Способен формировать нетерпимое отношение к проявлениям экстремизма, терроризма, коррупционному поведению и противодействовать им в профессиональной деятельности	УК-11.3 Соблюдает правила общественного взаимодействия на основе соблюдения действующего законодательства и нетерпимого отношения к коррупции;
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 необходимые для осуществления профессиональной деятельности правовые нормы и методологические основы принятия управленческого решения;
УК-3	Способен осуществлять социальное взаимодействие и реализовывать свою роль в команде	УК-3.2 Умеет действовать в духе сотрудничества; принимать решения с соблюдением этических принципов их реализации; проявлять уважение к мнению и культуре других; определять цели и работать в направлении личностного, образовательного и профессионального роста;

Шифр	Компетенция	Индикаторы достижения компетенции (в рамках данной дисциплины)
УК-5	Способен воспринимать межкультурное разнообразие общества в социально-историческом, этическом и философском контекстах	УК-5.2 Придерживается принципов недискриминационного взаимодействия при личном и массовом общении в целях выполнения профессиональных задач и усиления социальной интеграции; УК-5.3 Учитывает при социальном и профессиональном общении по заданной теме историческое наследие и социокультурные традиции различных социальных групп, этносов и конфессий, включая мировые религии, философские и этические учения;
УК-9	Способен использовать базовые дефектологические знания в социальной и профессиональной сферах	УК-9.2 Умеет дифференцированно использовать базовые дефектологические знания 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-6	Способен анализировать и разрабатывать организационно-технические процессы с применением методов системного анализа, математического моделирования и технологий искусственного интеллекта	ОПК-6.2 Умеет анализировать предметную область с позиции системного подхода, определять требования к ИИ-системе, формализовывать бизнес-задачи в задачи машинного обучения;
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;
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;
SS-1	Способен учитывать философские, когнитивные и социальные основания концепций ИИ в профессиональной	SS-1.2 Применяет методики работы с этическими и социальными рисками, возникающими на разных стадиях жизненного цикла ИИ;

Шифр	Компетенция	Индикаторы достижения компетенции (в рамках данной дисциплины)
	деятельности	
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); SS-3.2 Определяет релевантность применения ИИ для решения конкретных задач, анализирует поведение ИИ в техническом, социальном и правовом контекстах, переносит идеи и методы за пределы исходной предметной области; SS-3.3 Осуществляет метарефлексию при анализе систем и принятии решений, предсказывает возможные эффекты от внедрения ИИ через несколько уровней влияния, переосмысляет ИИ в своей профессиональной роли и в обществе;

3. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОП ВО

Дисциплина «Этика и безопасность использования 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 «Этика и безопасность использования 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; Безопасность систем искусственного интеллекта;
УК-11	Способен формировать нетерпимое отношение к проявлениям экстремизма, терроризма, коррупционному поведению и противодействовать им в профессиональной деятельности	Правоведение; Основы военной подготовки. Безопасность жизнедеятельности; Основы российской государственности;	
УК-3	Способен осуществлять социальное взаимодействие и реализовывать свою роль в команде		Философия; Практическая подготовка на проектах отраслевых индустриальных партнеров; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP); Эксплуатационная практика

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
			(производственная); Технологическая (проектно-технологическая) практика (производственная);
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	Правоведение;	Эксплуатационная практика (учебная); Технологическая (проектно-технологическая) практика (производственная); Методы разработки решений на основе искусственного интеллекта (Git, Docker); Оптимизация моделей машинного обучения; Практическая подготовка на проектах отраслевых индустриальных партнеров; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP); MLOps и промышленная разработка систем искусственного интеллекта;
УК-5	Способен воспринимать межкультурное разнообразие общества в социально-историческом, этическом и философском контекстах	История России; Основы российской государственности;	Философия;
ОПК-6	Способен анализировать и разрабатывать организационно-технические процессы с применением методов системного анализа, математического моделирования и технологий искусственного интеллекта	Введение в искусственный интеллект;	Онтология и графы знаний; Методы машинного обучения; Практическая подготовка на проектах отраслевых индустриальных партнеров; Оптимизация моделей машинного обучения;
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); Технологическая (проектно-технологическая) практика (производственная); Безопасность систем искусственного интеллекта; Методы разработки решений на основе искусственного интеллекта (Git, Docker); MLOps и промышленная разработка систем искусственного интеллекта; Практическая подготовка на проектах отраслевых индустриальных партнеров;

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
ПК-1	Способен анализировать требования к программному обеспечению систем ИИ, разрабатывать технические спецификации и техническое задание на систему	Правоведение; Введение в искусственный интеллект; История и теория программирования; Технологическая (проектно-технологическая) практика (учебная);	Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Преддипломная практика; Технологическая (проектно-технологическая) практика (производственная); Параллельное и распределенное программирование; Методы машинного обучения; Массово-параллельные вычисления в машинном обучении (GPU); Оптимизация моделей машинного обучения; Основы глубокого обучения; Безопасность систем искусственного интеллекта; Практическая подготовка на проектах отраслевых индустриальных партнеров; *Большие языковые модели***; Программирование на языке C++; Методы разработки решений на основе искусственного интеллекта (Git, Docker); Введение в базы данных; MLOps и промышленная разработка систем искусственного интеллекта; Нейронные сети; Онтология и графы знаний; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP);
ПК-3	Способен разрабатывать и реализовывать стратегии тестирования и контроля качества программного обеспечения систем ИИ	Технологическая (проектно-технологическая) практика (учебная); Программирование на языке Python;	Преддипломная практика; Технологическая (проектно-технологическая) практика (производственная); Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Методы машинного обучения; Нейронные сети; Безопасность систем искусственного интеллекта; Обработка и анализ изображений и видео с помощью методов искусственного интеллекта;

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
			Анализ естественного языка с помощью методов искусственного интеллекта; Методы разработки решений на основе искусственного интеллекта (Git, Docker); MLOps и промышленная разработка систем искусственного интеллекта; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP); Оптимизация моделей машинного обучения; Практическая подготовка на проектах отраслевых индустриальных партнеров;
SS-1	Способен учитывать философские, когнитивные и социальные основания концепций ИИ в профессиональной деятельности	История и теория программирования; Введение в искусственный интеллект;	Философия; Онтология и графы знаний; Методы машинного обучения; MLOps и промышленная разработка систем искусственного интеллекта; Основы глубокого обучения; Нейронные сети; Лингвистические основы анализа естественного языка; *Основы робототехники***; *Большие языковые модели***; *Генеративные модели***; Безопасность систем искусственного интеллекта; Практическая подготовка на проектах отраслевых индустриальных партнеров; Проектирование и разработка систем компьютерного зрения; Практикум по обработке естественного языка (NLP); *Рекомендательные системы***;
SS-3	Способен к критическому анализу, метарефлексии и переносу знаний при работе с системами ИИ	Правоведение; Введение в искусственный интеллект;	Эксплуатационная практика (учебная); Эксплуатационная практика (производственная); Технологическая (проектно-технологическая) практика (производственная); Преддипломная практика; Методы машинного обучения; Нейронные сети; Безопасность систем

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
			искусственного интеллекта; Обработка и анализ изображений и видео с помощью методов искусственного интеллекта; Анализ естественного языка с помощью методов искусственного интеллекта; *Вайб-кодинг***; Оптимизация моделей машинного обучения; MLOps и промышленная разработка систем искусственного интеллекта; Практическая подготовка на проектах отраслевых индустриальных партнеров; 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***; *Информационный поиск***; *Рекомендательные системы***; *Обработка сигналов***; *Анализ временных рядов***; Философия; *Большие языковые модели***;
AI S-1	Способен управлять рисками при разработке и использовании систем ИИ, выстраивать управление безопасностью ИИ в организации с учетом принципов этического использования ИИ		Преддипломная практика; Эксплуатационная практика (производственная); Безопасность систем искусственного интеллекта;
FC-5	Способен проводить передовые исследования в области безопасности, доверия и объяснимости	Правоведение;	Безопасность систем искусственного интеллекта; Практическая подготовка на проектах отраслевых индустриальных партнеров; Методы машинного обучения; MLOps и промышленная разработка систем искусственного интеллекта;

Шифр	Наименование компетенции	Предшествующие дисциплины/модули, практики*	Последующие дисциплины/модули, практики*
			Эксплуатационная практика (производственная); Преддипломная практика; Эксплуатационная практика (учебная);

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

** - элективные дисциплины /практики

4. ОБЪЕМ ДИСЦИПЛИНЫ И ВИДЫ УЧЕБНОЙ РАБОТЫ

Общая трудоемкость дисциплины «Этика и безопасность использования 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)**
			3
Контактная работа, ак.ч.	*51*		*51*
Лекции (**ЛК**)	17		17
Лабораторные работы (**ЛР**)	0		0
Практические/семинарские занятия (**СЗ**)	34		34
Самостоятельная работа обучающихся, ак.ч.	*39*		*39*
Контроль (экзамен/зачет с оценкой), ак.ч.	*18*		*18*
Общая трудоемкость дисциплины	**ак.ч.**	***108***	*108*
	зач.ед.	***3***	*3*

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	Этические основания и принципы ответственного ИИ	1.1	Введение: этика технологий и ИИ	Предмет этики ИИ: почему технологии требуют этической рефлексии. Историческая перспектива: от законов робототехники Азимова до современных этических кодексов. Утилитаризм, деонтология, этика добродетели и этика заботы применительно к ИИ. Ключевые этические принципы: справедливость, прозрачность, подотчётность, безопасность, приватность	ЛК	SS-1.2, УК-5.3, SS-3.3
		1.2	Алгоритмическая предвзятость и справедливость	Источники предвзятости в ИИ-системах: данные (исторические смещения, недопредставленность групп), разметка (субъективность аннотаторов), архитектура модели, метрики оценки. Формальные определения справедливости: демографический паритет, равенство возможностей, предиктивное равенство. Невозможность одновременного выполнения всех критериев. Кейсы: COMPAS, Amazon recruiting, распознавание лиц	ЛК	SS-3.1, УК-5.2, AIS-1.1
		1.3	Прозрачность, объяснимость и подотчётность	Проблема «чёрного ящика» в глубоком обучении. Объяснимый ИИ (XAI): LIME, SHAP, attention maps, concept-based explanations. Уровни объяснимости: для разработчика, для пользователя, для регулятора. Подотчётность: кто несёт ответственность за решения ИИ? Модели подотчётности: разработчик, оператор, пользователь. «Пробел ответственности» (responsibility gap)	ЛК	SS-1.2, AIS-1.2, ОПК-6.2
		1.4	Практикум: анализ этических дилемм в ИИ	Разбор классических этических дилемм: проблема вагонетки и автономное вождение (Moral Machine), медицинский ИИ и распределение ресурсов, предиктивная полиция и профилирование. Применение различных этических теорий к каждой дилемме. Групповая дискуссия: существует ли «правильный» ответ?	СЗ	SS-1.2, УК-3.2, SS-3.3
		1.5	Практикум: измерение и митигация алгоритмической предвзятости	Работа с набором данных с заведомыми смещениями. Вычисление метрик справедливости (demographic parity, equalized odds) на примере бинарного классификатора. Применение методов митигации: ребалансировка выборки,	СЗ	SS-3.1, AIS-1.1, ПК-3.1

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				adversarial debiasing (концепция). Использование библиотеки Fairlearn / AI Fairness 360		
		1.6	Практикум: объяснимость моделей на практике	Обучение модели на табличных данных (решающее дерево, случайный лес). Применение методов объяснимости: SHAP-values, feature importance, partial dependence plots. Интерпретация результатов для разных аудиторий: технический отчёт vs. объяснение для пользователя. Обсуждение ограничений объяснимости	СЗ	AIS-1.2, ОПК-6.2, SS-3.2
		1.7	Практикум: ИИ и дискриминация — мультикультурный контекст	Анализ кейсов алгоритмической дискриминации в разных культурных контекстах: распознавание лиц и расовая принадлежность, NLP-модели и гендерные стереотипы, рекомендательные системы и фильтровые пузыри. Различие культурных норм справедливости. Разработка чек-листа проверки на дискриминацию	СЗ	УК-5.2, УК-5.3, SS-3.1
		1.8	Практикум: проектирование ИИ с учётом инклюзивности	Анализ ИИ-продукта с позиции доступности для людей с ОВЗ: незрячие пользователи и системы CV, глухие пользователи и голосовые ассистенты, пользователи с когнитивными особенностями. Формулирование требований к ИИ-системе с учётом инклюзивности. Принципы универсального дизайна и ИИ	СЗ	УК-9.2, ОПК-6.2, ПК-1.1
		1.9	Практикум: этическая экспертиза ИИ-проекта (часть 1)	Групповая работа: выбор ИИ-проекта (медицина, образование, финансы, HR, правосудие). Идентификация стейкхолдеров. Выявление потенциальных этических рисков. Применение матрицы «вероятность × воздействие». Формулирование этических требований к системе. Начало составления отчёта	СЗ	SS-1.2, SS-3.2, ПК-1.1
Раздел 2	Правовое регулирование, стандарты и управление жизненным циклом ИИ	2.1	Правовое регулирование ИИ: международный ландшафт	Европейский подход: AI Act (классификация по уровням риска, запрещённые практики, требования к высокорисковым системам). Американский подход: Blueprint for an AI Bill of Rights, NIST AI RMF. Китайский подход: регулирование генеративного ИИ. Российский подход: Национальная стратегия ИИ, экспериментальные правовые режимы, 152-ФЗ (персональные данные)	ЛК	УК-2.1, ОПК-4.1, FC-5.1
		2.2	Стандарты и нормативные документы для ИИ-систем	Международные стандарты: ISO/IEC 42001 (система управления ИИ), ISO/IEC 23894 (управление рисками ИИ),	ЛК	ОПК-4.1, FC-5.1,

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				IEEE 7000 (этическое проектирование). Российские стандарты: ГОСТ Р 59277 (ИИ — терминология), ГОСТ Р 59898 (оценка качества ИИ-систем). Требования к документированию ИИ: model cards, datasheets for datasets, FactSheets		FC-5.2
		2.3	Управление рисками ИИ на стадиях жизненного цикла	Жизненный цикл ИИ-системы: постановка задачи → сбор данных → обучение → валидация → развёртывание → мониторинг → вывод из эксплуатации. Этические и социальные риски на каждой стадии. NIST AI Risk Management Framework: Govern, Map, Measure, Manage. Оценка воздействия ИИ (AI Impact Assessment). Аудит ИИ-систем: внутренний и внешний	ЛК	SS-1.2, FC-5.2, FC-5.3
		2.4	Практикум: анализ AI Act — классификация рисков	Практическое применение AI Act: классификация ИИ-систем из реальной практики по уровням риска (минимальный, ограниченный, высокий, неприемлемый). Определение обязательств разработчика для каждого уровня. Анализ: какие требования предъявляются к высокорисковым системам (документация, данные, прозрачность, человеческий надзор)	СЗ	УК-2.1, ОПК-4.1, FC-5.1
		2.5	Практикум: составление Model Card и Datasheet	Составление Model Card для обученной модели по шаблону Mitchell et al. (2019): назначение, ограничения, метрики, этические аспекты. Составление Datasheet for Datasets по шаблону Gebru et al. (2021): источник данных, процесс сбора, состав, предвзятости. Обсуждение: как документирование снижает риски	СЗ	ОПК-4.1, FC-5.2, AIS-1.2
		2.6	Практикум: приватность и защита данных в ИИ	152-ФЗ и GDPR: основные требования к обработке персональных данных. Принцип минимизации данных. Право на объяснение алгоритмического решения. Технические методы обеспечения приватности: дифференциальная приватность (концепция), федеративное обучение (обзор), анонимизация и псевдонимизация	СЗ	УК-2.1, AIS-1.1, ПК-1.1
		2.7	Практикум: оценка воздействия ИИ-системы (AI Impact Assessment)	152-ФЗ и GDPR: основные требования к обработке персональных данных. Принцип минимизации данных. Право на объяснение алгоритмического решения. Технические методы обеспечения приватности:	СЗ	FC-5.2, FC-5.3, SS-1.2

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				дифференциальная приватность (концепция), федеративное обучение (обзор), анонимизация и псевдонимизация		
		2.8	Практикум: формулирование критериев приёмки с учётом этических требований	Трансляция этических принципов в верифицируемые технические требования: пороги метрик справедливости, требования к объяснимости, ограничения на использование данных. Формулирование критериев приёмки для ИИ-системы с учётом AI Act и стандартов. Обсуждение: можно ли формализовать этику?	СЗ	ПК-1.1, ПК-3.1, ОПК-6.2
		2.9	Практикум: этическая экспертиза ИИ-проекта (часть 2)	Продолжение группового проекта: анализ правовых требований к выбранному проекту. Определение применимых стандартов. Заполнение AI Impact Assessment. Составление Model Card и Datasheet. Формулирование	СЗ	SS-1.2, FC-5.1, ОПК-4.1
Раздел 3	Социальные последствия ИИ, безопасность и ответственная разработка	3.1	Социальные последствия ИИ: труд, неравенство, демократия	ИИ и рынок труда: автоматизация задач, поляризация, новые профессии. ИИ и неравенство: цифровой разрыв, концентрация данных и вычислительных ресурсов. ИИ и демократия: дезинформация, дипфейки, манипуляция общественным мнением. ИИ и экология: энергопотребление обучения больших моделей. Связь с целями устойчивого развития	ЛК	SS-3.3, SS-3.2, УК-5.3
		3.2	Безопасность ИИ: от adversarial attacks до alignment	Обзор угроз безопасности ИИ: adversarial attacks, data poisoning, model stealing, prompt injection. Проблема выравнивания (alignment): что значит «цели ИИ соответствуют целям человека»? Reward hacking и спецификация целей. Экзистенциальные риски (обзор дискуссии). Принцип предосторожности. Связь безопасности ИИ с этикой	ЛК	AIS-1.1, AIS-1.2, FC-5.3
		3.3	Ответственная разработка: от принципов к практике	Интеграция этики в процесс разработки: ethics by design, value-sensitive design. Роли в команде: этический офицер, ответственный за данные. Этические комитеты и ревью-борды. Инструменты ответственной разработки: чек-листы, ethical canvas, consequence scanning. Кодексы профессиональной этики в ИТ. Ответственность разработчика: юридическая и моральная	ЛК	SS-1.2, УК-11.3, УК-3.2
		3.4	Практикум: анализ социальных последствий генеративного ИИ	Кейс-стади: социальные последствия генеративного ИИ (LLM, генерация изображений). Дезинформация и	СЗ	SS-3.2, SS-3.3,

Номер раздела	Наименование раздела дисциплины	Наименование темы		Содержание темы	Вид учебной работы *	Формируемые индикаторы
				дипфейки: технические и социальные меры противодействия. Авторское право и ИИ: кому принадлежит сгенерированный контент? Влияние на образование: плагиат, копирайт, трансформация навыков. Групповая дискуссия с аргументацией позиций		УК-11.3
		3.5	Практикум: red teaming ИИ-системы	Концепция red teaming для ИИ: систематический поиск уязвимостей и нежелательного поведения. Практика: формулирование adversarial-промптов для LLM, попытки обхода ограничений, документирование найденных уязвимостей. Обсуждение: как red teaming связан с этическим аудитом. Ответственное раскрытие уязвимостей	СЗ	AIS-1.1, AIS-1.2, ПК-3.1
		3.6	Практикум: consequence scanning для ИИ-проекта	Метод consequence scanning (Doteveryone): систематический анализ последствий внедрения ИИ-продукта для разных групп стейкхолдеров. Работа в командах: идентификация предполагаемых и непредполагаемых последствий, позитивных и негативных. Формулирование мер предотвращения негативных последствий	СЗ	FC-5.3, SS-3.3, УК-3.2
		3.7	Практикум: разработка этического кодекса команды ИИ-разработчиков	Анализ существующих кодексов: ACM Code of Ethics, IEEE Ethically Aligned Design, Кодекс этики ИИ (Россия). Разработка собственного кодекса для команды ИИ-разработчиков: принципы, обязательства, процедуры разрешения конфликтов. Презентация и взаимная экспертиза кодексов между командами	СЗ	УК-3.2, УК-11.3, FC-5.1
		3.8	Практикум: защита итоговой этической экспертизы ИИ-проекта	Финальная презентация группового проекта: полный отчёт этической экспертизы ИИ-системы, включающий анализ стейкхолдеров, этические риски, правовые требования, AI Impact Assessment, Model Card, критерии приёмки, рекомендации. Перекрёстная экспертиза между командами. Рефлексия: чему я научился о своей ответственности как разработчика ИИ	СЗ	SS-1.2, FC-5.2, SS-3.3, ПК-1.1

* - 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. Баюк, Д. А. Правовые и этические проблемы искусственного интеллекта : учебник для магистратуры / Д. А. Баюк, А. В. Попова. - Москва : Прометей, 2022. - 300 с. - (Высшее образование: магистратура). - ISBN 978-5-00172-253-3. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2124861

2. Филипова И.А. Правовое регулирование искусственного интеллекта: учебное пособие, 3-е издание, обновленное и дополненное – Нижний Новгород: Нижегородский госуниверситет, 2025. – 321 с.

Дополнительная литература:

1. Этика искусственного интеллекта (Витрина технологий от SBER). - URL: https://developers.sber.ru/help/business-development/ethics-of-artificial-intelligence?ysclid=mogxmxuf7w166712740

2. Этические аспекты искусственного интеллекта (Рекомендация UNESCO). - URL: https://www.unesco.org/ru/artificial-intelligence/recommendation-ethics

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 «Этика и безопасность использования искусственного интеллекта».

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