

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
Факультет компьютерных технологий и прикладной математики

УТВЕРЖДАЮ:

Проректор по учебной работе,
качеству образования – первый
проректор

_____ Хагуров Т.А.

подпись

«02» июня 2026 г

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Б1. О.37 Этика и социальная ответственность в ИИ

Направление подготовки 02.03.03 Математическое обеспечение и администрирование информационных систем

Профиль Искусственный интеллект и аналитика данных

Форма обучения очная

Квалификация бакалавр

Краснодар 2026

Рабочая программа дисциплины Этика и социальная ответственность в ИИ составлена в соответствии с федеральным государственным образовательным стандартом высшего образования (ФГОС ВО) по направлению подготовки 02.03.03 Математическое обеспечение и администрирование информационных систем

Программу составил(и):

Г.В. Калайдина, доцент, к.ф.-м.н.

И.О. Фамилия, должность, ученая степень, ученое звание

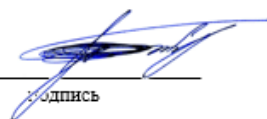


подпись

Рабочая программа дисциплины утверждена на заседании центра
искусственного интеллекта

протокол № 04 «14» мая 2026 г.

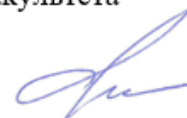
Руководитель центра ИИ Коваленко А.В.



подпись

Утверждена на заседании учебно-методической комиссии факультета
компьютерных технологий и прикладной математики
протокол № 03 «15» мая 2026 г.

Председатель УМК факультета Добровольская Н.Ю.



подпись

Рецензенты:

Мостовой Евгений Викторович, генеральный директор ООО «Портал-Юг»,
e-mail: mostovoy@portal-yug.ru

Лутенко Евгений Вениаминович, доктор экономических наук, кандидат технических наук, профессор кафедры компьютерных технологий и систем Федерального государственного бюджетного образовательного учреждения высшего образования «Кубанский государственный аграрный университет имени И.Т. Трубилина», e-mail: prof.lutsenko@gmail.com

1 Цели и задачи изучения дисциплины (модуля)

1.1 Цель освоения дисциплины Формирование у будущих специалистов в области искусственного интеллекта и аналитики данных системного понимания этических принципов, правовых норм и социальной ответственности, связанных с разработкой и применением ИИ, а также выработка практических навыков для выявления, анализа и минимизации этических рисков в профессиональной деятельности.

1.2 Задачи дисциплины

- Сформировать понимание ключевых этических проблем и дилемм, порождаемых современными ИИ-технологиями (предвзятость, приватность, безопасность, подотчетность).
- Изучить основные принципы, frameworks и регуляторные инициативы в области ответственного ИИ (на международном и национальном уровнях).
- Освоить методики выявления ценностных предпосылок, когнитивных искажений и культурно-обусловленных предвзятостей в данных и алгоритмах.
- Развить навыки применения инструментов и методов профессиональной коммуникации для обсуждения, презентации и аргументации этических аспектов ИИ-проектов.
- Научить интегрировать этические соображения и методики управления рисками на различных стадиях жизненного цикла ИИ-систем.

1.3 Место дисциплины (модуля) в структуре образовательной программы

Дисциплина «Этика и социальная ответственность в ИИ» относится к Обязательной части Блока 1 «Дисциплины (модули учебного плана).

Материал курса тесно связан с дисциплинами (пререквизиты) Машинное обучение, Глубокое обучение, Нейросетевые технологии, Обработка данных на Python, Компьютерная графика, Технологии обработки больших данных, Технологии компьютерного зрения, Технологии обработки языка, Генеративный искусственный интеллект.

Пререквизиты – Философия (SS-1.1 (Б), SS-1.2 (Б)), Введение в специальность (LC-8.1 (Б))

Постреквизиты – Математические модели защиты информации SS-1.1 (П), Правовые основы оценки проектных решений, SS-1.1 (П), SS-1.2 (П)), Правовые основы рынка программного обеспечения SS-1.1 (П), SS-1.2 (П).

Дисциплина напрямую определяет **качество и глубину** выпускной работы.

1.4 Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы

Роль 1: Data Analyst (Аналитик данных)

Задачи:

1. Статистический анализ, визуализация данных, предварительная обработка.
2. Создание прогнозных моделей
3. Построение аналитических моделей для поддержки бизнес-решений.

Роль 2: MLOps (Специалист по эксплуатации ИИ)

Задачи:

1. DevOps для ML.
2. Автоматизация, мониторинг ML-систем.
3. Операционное управление жизненным циклом ML-моделей.

Роль 3: AI PM (Менеджер проектов ИИ)

Задачи:

1. Управление ИИ-проектами от идеи до внедрения
2. Анализ бизнес-требований и постановка задач

3. Оценка эффективности и ROI ИИ-решений

Изучение данной учебной дисциплины направлено на формирование у обучающихся следующих компетенций:

Код и наименование индикатора	Результаты обучения по дисциплине
ОПК-2 Способен применять современный математический аппарат, связанный с проектированием, разработкой, реализацией и оценкой качества программных продуктов и программных комплексов в различных областях человеческой деятельности	
ОПК-2.1 Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС	Знать основные понятия системного подхода, методы выявления требований к ИС (интервью, сценарии, анализ стейкхолдеров, risk-воркшопы)
	Уметь Проводить анализ проблемной области, выделять стейкхолдеров, формулировать требования к ИИ-системе с учетом этических аспектов
	Владеть методами системного анализа применительно к этическому аудиту ИИ-проектов
ЛС-8 Способен применять знания об истории развития и трендах современного ИИ для формулирования корректных постановок задач и поиска перспективных способов решения проблем с помощью ИИ	
ЛС-8.1 Позиционирует собственную задачу в заданной области знания с точки зрения трендов современного искусственного интеллекта	Средний уровень Знать: расширенные тренды Уметь: анализировать и сопоставлять задачу с современными трендами, выделять специфику задачи в контексте последних достижений ИИ Владеть: навыком классификации задач по отношению к трендам
	Продвинутый уровень Знать Тренды и историю ИИ Уметь Выявлять потенциал новых подходов, формулировать задачи на стыке дисциплин Владеть Методами позиционирования задач в контексте трендов
СС-1 Способен осуществлять свою трудовую деятельность с учетом определения корректной роли ИИ в различных процессах, критического анализа последствий применения ИИ-технологий, этических принципов	
СС-1.1 Учитывает в разработке и эксплуатации систем ИИ философские основания концепций интеллекта, языка, знания, агентности	Средний уровень Знать: основы эпистемологии машинного обучения (как «обучаются» алгоритмы и что они «знают») Уметь: обосновывать выбор терминологии и интерпретации при описании поведения ИИ, применять философские подходы для анализа конкретных моделей ИИ (сравнение символического ИИ и нейросетей) Владеть: философским анализом архитектур ИИ
	Продвинутый уровень Знать

	Философские основы ИИ (интеллект, знание, агентность) Уметь Применять философские модели для анализа ИИ, определять границы машинного мышления Владеть Навыками философской рефлексии в инженерных решениях
SS-1.2 Применяет методики работы с этическими и социальными рисками, возникающими на разных стадиях жизненного цикла ИИ.	Средний уровень Знать: подходы к этической оценке ИИ и их ограничения (риск-ориентированный подход, бенчмаркинг, анкетирование) Уметь: <i>применять эти подходы к типовым системам</i> Владеть: навыком выбора методики для конкретного ЖЦ Продвинутый уровень Знать Методики работы с этическими рисками на ЖЦ ИИ Уметь Проводить этический аудит, участвовать в экспертизе и регуляторных инициативах Владеть Встраиванием этических принципов в проектирование и тестирование
LLM-1 Способен применять и (или) разрабатывать генеративные модели и БЯМ	
LLM-1.6 Учитывает этические аспекты генерации	Базовый уровень Знать: что такое галлюцинации, фильтрация, небезопасная генерация Уметь: распознавать этически проблемные ответы БЯМ на уровне пользователя Владеть: использованием готовых фильтров (например, moderation API) Средний уровень Знать: основные способы постобработки генераций и виды фильтров Уметь: настраивать фильтры и правила постобработки под задачу Владеть: применением фильтров в коде или конфигурации Продвинутый уровень Знать Этические аспекты генерации (галлюцинации, фильтрация) Уметь Разрабатывать и внедрять механизмы контроля этичности генеративных моделей Владеть Методами оценки этических рисков БЯМ

2. Структура и содержание дисциплины

2.1 Распределение трудоёмкости дисциплины по видам работ

Общая трудоёмкость дисциплины составляет 2 зач. ед. (72 час.), их распределение по видам работ представлено в таблице

Вид учебной работы		Всего часов	Семестры (часы)
			7
Контактная работа, в том числе:		36,2	36,2
Аудиторные занятия (всего):		34	34
Занятия лекционного типа		16	16
Лабораторные занятия		18	18
Занятия семинарского типа (семинары, практические занятия)			
Иная контактная работа:		2,2	2,2
Контроль самостоятельной работы (КСР)		2	2
Промежуточная аттестация (ИКР)		0,2	0,2
Самостоятельная работа, в том числе:		35,8	35,8
Проработка учебного (теоретического) материала		20	20
Выполнение индивидуальных заданий (типовой расчет)		10	10
Подготовка к текущему контролю		5,8	5,8
Контроль:		-	-
Подготовка к экзамену		-	-
Общая трудоемкость	час.	72	72
	в том числе контактная работа	36,2	36,2
	зач. ед	2	2

2.2 Структура дисциплины

Распределение видов учебной работы и их трудоемкости по разделам дисциплины.

Разделы (темы) дисциплины, изучаемые в 7 семестре

№	Наименование разделов (тем)	Количество часов				
		Всего	Аудиторная работа			Внеаудиторная работа
			Л	ПЗ	ЛР	
1	2	3	4	5	6	7
1.	Введение в этику ИИ. Основные концепции и вызовы	10	2	2		6
2.	Философские основания ИИ: интеллект, знание, агентность	12	2	2		6
3.	История ИИ и современные тренды. Позиционирование задач	8	2	2		4
4.	Предвзятость, справедливость, прозрачность и конфиденциальность	8	2	2		4
5.	Регулирование ИИ и управление социальными рисками	8	2	2		4
6.	Этические аспекты генеративных моделей и БЯМ	8	2	2		4

№	Наименование разделов (тем)	Количество часов				
		Всего	Аудиторная работа			Внеаудиторная работа
			Л	ПЗ	ЛР	
1	2	3	4	5	6	7
7.	Инструменты и практики внедрения ответственного ИИ. Этический аудит	7,8	2	2		3,8
8.	Профессиональная коммуникация в проектах ИИ: отчеты, презентации, дебаты	8	2	2		4
ИТОГО по разделам дисциплины		69,8	16	18		35,8
Контроль самостоятельной работы (КСР)		2				
Промежуточная аттестация (ИКР)		0,2				
Подготовка к текущему контролю						
Общая трудоемкость по дисциплине		72				

Примечание: Л – лекции, ПЗ – практические занятия/семинары, ЛР – лабораторные занятия, СРС – самостоятельная работа студента

2.3 Содержание разделов (тем) дисциплины

2.3.1 Занятия лекционного типа

2.3.1 Занятия лекционного типа

№	Наименование раздела (темы)	Содержание раздела (темы)	Соответствие индикаторам компетенций
1	2	3	
1	Введение в этику ИИ. Основные концепции и вызовы	Определение этики ИИ, ключевые проблемы: предвзятость, конфиденциальность, прозрачность, подотчетность, социальные риски.	SS-1.2 (Б) SS-1.2 (С)
2	Философские основания ИИ: интеллект, знание, агентность	Естественное vs искусственное познание. Онтологические и эпистемологические основы ИИ. Критическое переосмысление понятий «мышление», «понимание», «субъект».	SS-1.1 (Б) SS-1.1 (С) SS-1.1 (П)
3	История ИИ и современные тренды. Позиционирование задач	От логического подхода к глубокому обучению. Генеративный ИИ, БЯМ, мультимодальность. Как формулировать задачи на стыке дисциплин.	LC-8.1 (Б) — LC-8.1 (С) — LC-8.1 (П) —
4	Предвзятость, справедливость, прозрачность и конфиденциальность	Algorithmic Fairness, дискриминация в данных, методы интерпретации (SHAP/LIME), DP, дифференциальная приватность.	SS-1.2 (Б) SS-1.2 (С)
5	Регулирование ИИ и управление социальными рисками	EU AI Act, отраслевые стандарты. Социальные последствия ИИ: манипуляции, слежка, автономные системы.	SS-1.2 (С) SS-1.2 (П)
6	Этические аспекты генеративных моделей и БЯМ	Галлюцинации, фильтрация контента, токсичность, авторские права. Механизмы контроля качества и соответствия этическим нормам.	LLM-1.6 (Б) LLM-1.6 (С) LLM-1.6 (П)

№	Наименование раздела (темы)	Содержание раздела (темы)	Соответствие индикаторам компетенций
1	2	3	
7	Инструменты и практики внедрения ответственного ИИ. Этический аудит	Этический чек-лист, междисциплинарная экспертиза, методики работы с рисками на стадиях ЖЦ ИИ. Проведение аудита ИИ-систем.	SS-1.2 (П) ОПК-2.1
8	Профессиональная коммуникация в проектах ИИ: отчеты, презентации, дебаты	Структура отчета по этическому аудиту, аргументация в дискуссиях, подготовка презентаций для заинтересованных сторон.	ОПК-2.1

2.3.2 Занятия семинарского типа / лабораторные занятия

№	Наименование раздела (темы)	Наименование лабораторных работ	Соответствие индикаторам компетенций
1	2	3	4
1	Разбор кейса Cambridge Analytica	Дискуссия о роли данных и алгоритмов в манипуляции общественным мнением. Выявление этических нарушений.	SS-1.2 (Б) ОПК-2.1 (С)
2	Философские дебаты: может ли ИИ обладать субъектностью?	Командные дебаты на основе философских моделей (Сёрль, Деннет, китайская комната). Аргументация границ машинного мышления.	SS-1.1 (С) SS-1.1 (П) ОПК-2.1 (П)
3	Этический аудит датасета + Fairlearn	Анализ Adult Census на предвзятости. Сравнение «справедливой» и «обычной» модели в библиотеке Fairlearn.	SS-1.2 (С) SS-1.2 (П)
4	Анализ объяснимости модели (SHAP/LIME)	Применение SHAP/LIME для интерпретации предсказаний ML-модели. Оценка прозрачности.	SS-1.2 (С) SS-1.2 (П)
5	Симуляция этического ревью по EU AI Act	Ролевая игра: оценка гипотетического ИИ-стартапа (распознавание эмоций, найм, кредитный скоринг) на соответствие регуляторным нормам.	SS-1.2 (С) SS-1.2 (П)
6	Работа с галлюцинациями LLM и фильтрацией контента	Практический анализ ответов LLM (ChatGPT, LLaMa). Разработка простого механизма фильтрации / промптов для снижения этических рисков.	LLM-1.6 (Б) LLM-1.6 (С) LLM-1.6 (П)
7	Дебаты по этическим дилеммам ИИ	Командные дебаты: «Распознавание лиц в публичных пространствах», «Автономный транспорт и	ОПК-2.1

№	Наименование раздела (темы)	Наименование лабораторных работ	Соответствие индикаторам компетенций
1	2	3	4
		проблема вагонетки». Аргументация позиции.	
8.	Разработка этического чек-листа для ИИ-проекта	В группах создание чек-листа для сквозного этического контроля (от сбора данных до внедрения). Защита чек-листа.	SS-1.2 (С) SS-1.2 (П) ОПК-2.1
9.	Презентация и защита итогового проекта: этический аудит ИИ-системы	Командная презентация отчета по этическому аудиту выбранной ИИ-системы (или синтетического кейса). Вопросы, обсуждение, оценка.	LC-8.1 (П) SS-1.1 (П) SS-1.2 (П) LLM-1.6 (П) ОПК-2.1

Примечание: ЛР – отчет/защита лабораторной работы, КП – выполнение курсового проекта, КР – курсовой работы, РГЗ – расчетно-графического задания, Р – написание реферата, Э – эссе, К – коллоквиум, Т – тестирование, РЗ – решение задач.

2.3.4 Примерная тематика курсовых работ (проектов)

Курсовые работы по данному предмету не предусмотрены

2.4 Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)

№	Вид СРС	Перечень учебно-методического обеспечения дисциплины по выполнению самостоятельной работы
1	Проработка и повторение лекционного материала, материала учебной и научной литературы, подготовка к семинарским занятиям	Методические указания для подготовки к лекционным и семинарским занятиям, лабораторным работам, утвержденные на заседании кафедры анализа данных и искусственного интеллекта факультета компьютерных технологий и прикладной математики ФГБОУ ВО «КубГУ», протокол №7 от 22.03.2023 г
2	Подготовка к текущему контролю	Методические указания для подготовки к лекционным и семинарским занятиям, утвержденные на заседании кафедры анализа данных и искусственного интеллекта факультета компьютерных технологий и прикладной математики ФГБОУ ВО «КубГУ», протокол №7 от 22.03.2023 г

Учебно-методические материалы для самостоятельной работы обучающихся из числа инвалидов и лиц с ограниченными возможностями здоровья (ОВЗ) предоставляются в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

- в печатной форме увеличенным шрифтом,
- в форме электронного документа,
- в форме аудиофайла,
- в печатной форме на языке Брайля.

Для лиц с нарушениями слуха:

- в печатной форме,
- в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

- в печатной форме,
- в форме электронного документа,

– в форме аудиофайла.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

3. Образовательные технологии

В соответствии с требованиями ФГОС в программа дисциплины предусматривает использование в учебном процессе следующих образовательные технологии: чтение лекций с использованием мультимедийных технологий; лабораторные занятия.

С точки зрения применяемых методов используются как традиционные информационно-объяснительные лекции, так и интерактивная подача материала с мультимедийной системой. Компьютерные технологии в данном случае обеспечивают возможность разнопланового отображения алгоритмов и демонстрационного материала. Такое сочетание позволяет оптимально использовать отведенное время и раскрывать логику и содержание дисциплины.

Лабораторное занятие позволяет научить студента применять теоретические знания при решении и исследовании конкретных задач. Лабораторные занятия проводятся в компьютерных классах. Подход разбора конкретных ситуаций широко используется как преподавателем, так и студентами при проведении анализа результатов самостоятельной работы. Это обусловлено тем, что в процессе исследования часто встречаются задачи, для которых единых подходов не существует. Каждая конкретная задача при своем исследовании имеет множество подходов, а это требует разбора и оценки целой совокупности конкретных ситуаций.

Для лиц с ограниченными возможностями здоровья предусмотрена организация консультаций с использованием электронной почты.

Информационно-коммуникационные технологии (ИКТ) - расширяют рамки образовательного процесса, повышая его практическую направленность, способствуют интенсификации самостоятельной работы учащихся и повышению познавательной активности. В рамках ИКТ выделяются 2 вида технологий:

Технология использования компьютерных программ – позволяет эффективно дополнить процесс обучения языку на всех уровнях.

Интернет-технологии – предоставляют широкие возможности для поиска информации, разработки научных проектов, ведения научных исследований.

проектная технология - индивидуальная или коллективная деятельность по отбору, распределению и систематизации материала по определенной теме, в результате которой составляется проект;

анализ конкретных ситуаций - анализ реальных проблемных ситуаций, имевших место в соответствующей области профессиональной деятельности, и поиск вариантов лучших решений;

развитие критического мышления – образовательная деятельность, направленная на развитие у студентов разумного, рефлексивного мышления, способного выдвинуть новые идеи и увидеть новые возможности.

Подход разбора конкретных задач и ситуаций широко используется как преподавателем, так и студентами во время лекций, лабораторных занятий и анализа результатов самостоятельной работы. Это обусловлено тем, что при решении каждой конкретной задачи имеется, как правило, несколько методов, а это требует разбора и оценки целой совокупности конкретных ситуаций.

4. Оценочные и методические материалы

4.1 Оценочные средства для текущего контроля успеваемости и промежуточной аттестации

Оценочные средства предназначены для контроля и оценки образовательных достижений обучающихся, освоивших программу учебной дисциплины «название дисциплины».

Оценочные средства включает контрольные материалы для проведения **текущего контроля** в форме тестовых заданий, разноуровневых заданий, типовых расчетов и **промежуточной аттестации** в форме вопросов и заданий к экзамену.

Оценочные средства для инвалидов и лиц с ограниченными возможностями здоровья выбираются с учетом их индивидуальных психофизических особенностей.

– при необходимости инвалидам и лицам с ограниченными возможностями здоровья предоставляется дополнительное время для подготовки ответа на экзамене;

– при проведении процедуры оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья предусматривается использование технических средств, необходимых им в связи с их индивидуальными особенностями;

– при необходимости для обучающихся с ограниченными возможностями здоровья и инвалидов процедура оценивания результатов обучения по дисциплине может проводиться в несколько этапов.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине (модулю) предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

- в печатной форме увеличенным шрифтом,
- в форме электронного документа.

Для лиц с нарушениями слуха:

- в печатной форме,
- в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

- в печатной форме,
- в форме электронного документа.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

Структура оценочных средств для текущей и промежуточной аттестации

№ п/п	Контролируемые разделы (темы) дисциплины	Код контролируемой компетенции (или её части) с указанием уровня	Наименование оценочного средства
1	Введение в этику ИИ. Основные концепции и вызовы	SS-1.2 (Б), SS-1.2 (С)	Тест №1
2	Философские основания ИИ: интеллект, знание, агентность	SS-1.1 (Б), SS-1.1 (С), SS-1.1 (П)	Эссе/рефлексия, философские дебаты (Пр.2)
3	История ИИ и современные тренды. Позиционирование задач	LC-8.1 (Б), LC-8.1 (С), LC-8.1 (П)	Аналитический обзор, итоговый проект (глава 1)
4	Предвзятость, справедливость, прозрачность и конфиденциальность	SS-1.2 (Б), SS-1.2 (С)	Тест №2, аудит датасета (Пр.3), SHAP/LIME (Пр.4)
5	Регулирование ИИ и управление социальными рисками	SS-1.2 (С), SS-1.2 (П)	Кейс-анализ (EU AI Act), симуляция ревью (Пр.5)

6	Этические аспекты генеративных моделей и БЯМ	LLM-1.6 (Б), LLM-1.6 (С), LLM-1.6 (П)	Тест №3, работа с галлюцинациями (Пр.6)
7	Инструменты и практики внедрения ответственного ИИ. Этический аудит	SS-1.2 (П), ОПК-2.1	Контрольные вопросы, разработка чек-листа (Пр.8)
8	Профессиональная коммуникация в проектах ИИ: отчёты, презентации, дебаты	ОПК-2.1	Разбор коммуникативного кейса, дебаты (Пр.2, Пр.7), защита проекта (Пр.9)
	Все разделы	Все компетенции (Б, С, П)	Итоговый групповой проект (письменный отчёт + презентация + защита)

Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы

Соответствие практических работ и индикаторов компетенций

В таблице ниже представлено, как каждая практическая работа способствует формированию заявленных компетенций.

Оценочные средства для лекций (текущий контроль)

Тест №1 (Лекция 1. Введение в этику ИИ)

Формат: 10 вопросов (8 закрытых, 2 открытых). Время — 15 минут.

Примеры вопросов:

- Что из перечисленного НЕ является предметом этики ИИ?**
 - Алгоритмическая предвзятость
 - Конфиденциальность данных
 - Оптимизация гиперпараметров модели
 - Проблема подотчетности при автономных решениях
- Какое понятие обозначает ситуацию, когда модель дает систематически разные результаты для различных демографических групп при одинаковых объективных характеристиках?**
 - Галлюцинация
 - Алгоритмическая дискриминация
 - Дрейф концепта
 - Переобучение

- Открытый вопрос (2-3 предложения):**

Почему высокая точность модели на тестовой выборке не гарантирует ее этическую приемлемость? Приведите один пример.

2.2. Эссе/рефлексия (Лекция 2. Философские основания ИИ)

Задание: (тема на выбор, 300–400 слов):

- «Может ли ИИ обладать субъектностью?» (опираясь на Сёрля и Деннета).
- В чем разница между «знанием» человека и «знанием» ИИ с эпистемологической точки зрения?

Аналитический обзор (Лекция 3. История и тренды ИИ)

Задание: Выбрать одну из технологий (трансформеры, GAN, RLHF, нейросимволические системы) и описать:

Какой тренд современного ИИ она представляет.

Потенциал интеграции в смежную область (например, медицина, право, образование).

Какие этические вопросы порождает (LC-8.1).

Формат: 400–500 слов.

Тест №2 (Лекция 4. Предвзятость, справедливость, прозрачность)

Вопросы (примеры):

1. **Подберите метрику:** какая метрика fairness сравнивает долю положительных исходов в разных группах?
 - a) Equalized Odds
 - b) Demographic Parity Difference
 - c) Predictive Rate Parity
 - d) SHAP
2. **Что такое «гипотеза Томаса» применительно к ИИ?**
(открытый: если ИИ предсказывает человеку, что он преступник, это влияет на его поведение → self-fulfilling prophecy)
3. **Почему LIME и SHAP считаются методами «объяснимого ИИ»?**

Максимум: 5 баллов (1 закрытый = 1, открытые = 2+1).

Кейс-анализ (Лекция 5. Регулирование ИИ и социальные риски)

Ситуация:

Стартап разрабатывает ИИ для оценки «вероятности рецидива» в уголовном судопроизводстве (аналог COMPAS). Система классифицирована как высокорисковая по EU AI Act.

Задания:

1. Какие требования EU AI Act применимы? (назвать 3)
2. Какие социальные риски возникают (коренное население vs мигранты)?
3. Предложите одну регуляторную инициативу (что именно должно контролировать государство?).

Тест №3 (Лекция 6. Этические аспекты генеративных моделей)

Примеры вопросов:

1. **Что такое «галлюцинация» LLM?**
 - a) Переобучение на шуме
 - b) Генерация правдоподобного, но ложного контента
 - c) Утечка данных обучения
 - d) Снижение производительности
2. **Какой метод снижает токсичность генерации?**
 - a) Reinforcement Learning from Human Feedback (RLHF)
 - b) Data augmentation
 - c) Dropout
 - d) Early stopping
3. **Открытый:** Предложите два способа фильтрации опасного контента на выходе LLM.

Контрольные вопросы (Лекция 7. Инструменты и практики)

Устно/письменно (4 вопроса, на каждый 1–2 минуты):

1. Какие стадии ЖЦ ИИ наиболее критичны для этического аудита и почему?
2. Назовите 3 инструмента этического контроля и для чего каждый.
3. Что такое «этический чек-лист»? Приведите 2 примера пунктов.
4. Кто должен проводить междисциплинарную экспертизу ИИ-проекта?

Разбор коммуникативного кейса (Лекция 8. Профессиональная коммуникация)

Ситуация:

Вы провели этический аудит и нашли дискриминацию в системе кредитного скоринга. Руководитель продукта говорит: «У нас точность 95%, не надо саботировать запуск».

Задание:

Напишите **диалог / аргумент** (5–7 реплик), как вы аргументируете необходимость

исправлений. Используйте формулу:

«Не только точность, но и *fairness* / *риски* / *регуляторика*».

Соответствие практических работ и индикаторов компетенций

В таблице ниже представлено, как каждая практическая работа способствует формированию заявленных компетенций.

Индикатор компетенции	Соответствующие практические работы		Обоснование
ОПК-2.1 Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС	Практическая работа № 1 (в составе практики 3 — аналитический обзор тренда ИИ): системный анализ проблемной области Практическая работа № 5 (в составе практики 5): симуляция этического ревью — выявление требований к ИИ-системе Практическая работа № 7 (в составе практики 8): разработка чек-листа — этап системного анализа Практическая работа № 10 (в составе итогового проекта): раздел «Системный анализ проблемной области и выявление требований»		В ходе лабораторных работ студенты: Применяют системный подход: выделяют компоненты ИИ-системы, внешнюю среду, обратные связи. Идентифицируют заинтересованные стороны (пользователи, разработчики, регуляторы, общество) и их противоречивые требования. Формулируют функциональные и нефункциональные требования к ИИ-системе с учетом этических и правовых норм. Используют методы выявления требований (анализ сценариев, риск-воркшопы, карта стейкхолдеров). Создают основу для последующего этического аудита
ИС-8.1 Позиционирует собственную задачу в заданной области знания с точки зрения трендов современного искусственного интеллекта	Базовый (Б)	Практическая работа №1 (в составе аналитического обзора на лекции 3) — Описание тренда ИИ	Студенты учатся находить и описывать актуальные тренды ИИ (трансформеры, GAN, RLHF, мультимодальность) в письменной форме, выделяя ключевые характеристики каждого направления
	Средний (С)	Практическая работа №3 (раздел аналитической части) — Выбор предметной области для этического аудита	При выборе системы для аудита студенты должны обосновать, почему выбранная задача актуальна в контексте современных трендов (например, почему аудит LLM важнее аудита линейной регрессии в 2025 г.)
	Продвинутый (П)	Практическая работа №9 (Глава 1 итогового проекта) — «Анализ трендов и позиционирование задачи»	Студенты формулируют собственную задачу на стыке дисциплин (например, этический аудит генеративной модели в медицине), выявляют потенциал новых подходов и обосновывают перспективность решения методами позиционирования задач в контексте трендов
СС-1.1 Учитывает в разработке и эксплуатации систем ИИ философские основания концепций интеллекта, языка, знания, агентности	Базовый (Б)	Практическая работа №2 (подготовительный этап) — Изучение философских концепций (Сёрль, Деннет) перед дебатами	Студенты знакомятся с базовыми философскими понятиями: «китайская комната», «интенциональность», «субъектность». Формируется первичное понимание границ машинного мышления
	Средний (С)	Практическая работа №2 — Философские	Студенты применяют философские модели (Сёрль, Деннет, Тьюринг) для

		дебаты («Может ли ИИ обладать субъектностью?»)	анализа ИИ-систем. Обосновывают выбор терминологии (например, чем «знание» ИИ отличается от человеческого знания). Сравнивают символический и нейросетевой подходы с эпистемологической точки зрения
	Продвинутый (П)	Практическая работа №9 (Глава 2 итогового проекта) — «Философская рефлексия инженерного решения»	Студенты встраивают философскую рефлексия в обсуждение конкретного инженерного решения: определяют границы машинного мышления для выбранной ИИ-системы, отвечают на вопрос «можно ли считать данную систему агентной?», делают выводы о субъектности и понимании
SS-1.2 Применяет методики работы с этическими и социальными рисками на разных стадиях жизненного цикла ИИ	Базовый (Б)	Практическая работа №1 — Анализ кейса Cambridge Analytica Практическая работа №3 (подготовка) — Идентификация protected attributes в датасете	Студенты учатся выявлять этические риски по чек-листу: отсутствие информированного согласия, манипуляция, нарушение приватности. Осваивают навык анкетирования для предварительной оценки рисков
	Средний (С)	Практическая работа №3 — Этический аудит датасета Adult Census с Fairlearn Практическая работа №4 — Анализ объяснимости модели (SHAP/LIME) Практическая работа №5 — Симуляция этического ревью по EU AI Act	Студенты применяют конкретные методики: расчёт метрик fairness (Demographic Parity, Equalized Odds), интерпретацию предсказаний через SHAP/LIME, категоризацию рисков по EU AI Act. Формируется навык выбора методики для конкретного этапа ЖЦ ИИ (данные → модель → тестирование → внедрение)
	Продвинутый (П)	Практическая работа №8 — Разработка этического чек-листа для ИИ-проекта Практическая работа №9 — Этический аудит ИИ-системы (полный отчёт)	Студенты самостоятельно проводят сквозной этический аудит по стадиям ЖЦ (сбор данных → разработка → тестирование → внедрение → мониторинг). Разрабатывают чек-лист из 7–10 измеримых пунктов. Участвуют в экспертизе и формулируют регуляторные рекомендации. Встраивают этические принципы в проектирование и тестирование
LLM-1.6 Учитывает этические аспекты генерации (галлюцинации, фильтрация)	Базовый (Б)	Практическая работа №6 (этап 1) — Воспроизведение галлюцинации LLM через промпт-инжиниринг	Студенты учатся распознавать этически проблемные ответы БЯМ: галлюцинации, токсичность, ложные факты. Понимают, в чём заключаются риски генеративных моделей
	Средний (С)	Практическая работа №6 (этапы 2–3) — Разработка механизмов фильтрации и постобработки	Студенты настраивают фильтры (moderation API, регулярные выражения) и правила постобработки для снижения этических рисков. Применяют готовые инструменты фильтрации в коде или конфигурации LLM
	Продвинутый (П)	Практическая работа №6 (этап 4) —	Студенты разрабатывают и внедряют механизмы контроля соответствия

		Этический анализ допустимости галлюцинаций Практическая работа №9 (раздел для LLM- кейсов) — Разработка механизмов контроля этичности генеративных моделей	этическим нормам (RAG + фильтрация по источникам + закрытый домен). Оценивают этические риски БЯМ в различных контекстах (медицина vs творческие задачи). Дифференцируют допустимость галлюцинаций в зависимости от области применения
--	--	--	---

ТЕСТ №4 (Итоговый / Рубежный)

по дисциплине «Этика и социальная ответственность в ИИ»

Форма проведения: письменно / в LMS Moodle

Время выполнения: 25 минут

Максимальный балл: 12 (по 1 баллу за вопрос)

Цель: проверка сформированности компетенций по всем разделам курса

Вопрос 1 (Базовый уровень, SS-1.2 Б)

Что из перечисленного НЕ является предметом этики искусственного интеллекта?

- A) Алгоритмическая предвзятость
- B) Конфиденциальность персональных данных
- C) Оптимизация гиперпараметров нейронной сети
- D) Проблема подотчётности при автономных решениях

Правильный ответ: C

Вопрос 2 (Базовый уровень, LC-8.1 Б)

Расположите следующие этапы развития ИИ в хронологическом порядке (от раннего к позднему):

1. Правила и экспертные системы (символический ИИ)
2. Статистическое машинное обучение (SVM, случайный лес)
3. Глубокое обучение и свёрточные сети
4. Генеративные модели и большие языковые модели (LLM)

Правильный ответ: 1 → 2 → 3 → 4

Вопрос 3 (Базовый уровень, SS-1.1 Б)

Какое понятие в философии ИИ обозначает способность системы быть носителем целей, убеждений и намерений, а также действовать на их основе?

- A) Интеллект
- B) Агентность (agency)
- C) Эпистемология
- D) Феноменология

Правильный ответ: B

Вопрос 4 (Базовый уровень, LLM-1.6 Б)

Что такое «галлюцинация» в контексте работы больших языковых моделей (LLM)?

- A) Переобучение модели на шумовых данных
- B) Генерация правдоподобного, но фактически ложного контента
- C) Снижение производительности модели после длительной работы
- D) Утечка данных из обучающей выборки в ответ модели

Правильный ответ: B

Вопрос 5 (Средний уровень, SS-1.2 C)

Какая метрика fairness оценивает разницу в доле положительных исходов между защищённой и привилегированной группами?

- A) Equalized Odds Difference
- B) Demographic Parity Difference
- C) Predictive Rate Parity
- D) SHAP value

Правильный ответ: B

Вопрос 6 (Средний уровень, SS-1.1 C)

В чём, согласно эпистемологии машинного обучения, заключается ключевое отличие «знания» нейросетевой модели от «знания» человека?

- A) Нейросеть не может хранить большие объёмы информации
- B) Нейросеть не обладает интенциональностью и пониманием смысла
- C) Нейросеть обучается только на структурированных данных
- D) Нейросеть не способна к обобщению

Правильный ответ: B

Вопрос 7 (Средний уровень, LC-8.1 C)

Какое из перечисленных направлений ИИ относится к современному тренду «мультимодальности»?

- A) Модели, способные обрабатывать и генерировать текст, изображения и аудио одновременно
- B) Алгоритмы, использующие только числовые признаки
- C) Методы, основанные исключительно на логических правилах
- D) Алгоритмы кластеризации без учителя

Правильный ответ: A

Вопрос 8 (Средний уровень, LLM-1.6 C)

Какой метод НЕ используется для снижения токсичности и галлюцинаций в ответах LLM?

- A) Reinforcement Learning from Human Feedback (RLHF)
- B) Постобработка выходного контента с фильтрацией
- C) Увеличение количества слоёв нейронной сети
- D) Промпт-инжиниринг с уточнением границ ответа

Правильный ответ: C

Вопрос 9 (Продвинутый уровень, SS-1.2 П)

Разработчик создаёт систему ИИ для автоматического найма персонала. Согласно EU AI Act, к какой категории риска, скорее всего, будет отнесена эта система, и какое требование станет обязательным?

- A) Минимальный риск — уведомление пользователя
- B) Высокий риск — требование прозрачности и объяснимости решений
- C) Неприемлемый риск — запрет на использование
- D) Ограниченный риск — добровольная сертификация

Правильный ответ: B

Вопрос 10 (Продвинутый уровень, LC-8.1 П)

Вы работаете над проектом в области медицинской диагностики и хотите применить современный подход ИИ, который позволяет не только ставить диагноз, но и объяснять ход рассуждений, используя как нейросетевую обработку изображений, так и символьные правила. Как называется этот тренд?

- A) Мультиmodalность
- B) Нейросимволический ИИ
- C) Трансформерная архитектура
- D) Обучение с подкреплением

Правильный ответ: B

Вопрос 11 (Продвинутый уровень, SS-1.1 П)

В дискуссии о субъектности ИИ сторонник позиции «сильный ИИ невозможен» ссылается на мысленный эксперимент, где человек, не знающий китайского языка, с помощью инструкций имитирует понимание китайского языка, но реального понимания не возникает. На какой философский аргумент он опирается?

- A) Тест Тьюринга
- B) Китайская комната Джона Сёрля
- C) Философский зомби
- D) Проблема квалиа

Правильный ответ: B

Вопрос 12 (Продвинутый уровень, LLM-1.6 П)

Вы разрабатываете механизм контроля этичности генеративной модели в системе поддержки принятия врачебных решений. Галлюцинации здесь недопустимы.

Какую комбинацию методов вы выберете?

- A) Только увеличение температуры генерации
- B) RAG (поиск по базе знаний) + фильтрация ответов по источникам + закрытый домен
- C) Исключительно постобработка регулярными выражениями
- D) Полное отключение всех фильтров для скорости работы

Правильный ответ: B

Оценочные средства для практических занятий

Практика 1 — Кейс Cambridge Analytica (дискуссия)

Формат: Групповая дискуссия (10 мин подготовки, 15–20 мин дискуссии).

Вопросы для направляющего:

1. Были ли этичными действия Cambridge Analytica?
2. Кто несет ответственность — платформа (Facebook), разработчики, заказчики?
3. Какой механизм регулирования предотвратил бы это?

Оценивание (4 балла):

- Активность (0–1)
- Аргументы по этике данных (0–2)
- Предложения по регулированию (0–1)

Практика 2 — Философские дебаты (SS-1.1 + ОПК-2.1)

Тема: «Может ли ИИ в принципе обладать субъектностью?»

Команды: «Да» (субъектность возможна — эмерджентна) vs «Нет» (субъектность требует биологии/феноменологии).

Регламент:

- 3 мин аргументы «за»
- 3 мин «против»
- 2 мин вопросы и возражения

Критерии (5 баллов):

- Философская аргументация (ссылки на Сёрля, Деннета, Тьюринга) — 2 балла
- Риторика и командная работа — 1 балл
- Ответы на вопросы — 1 балл

Этическая рефлексия (границы спора) — 1 балл

Практика 3 — Аудит датасета + Fairlearn

Отчет:

Краткое описание Adult Census (не более 1 стр)

Какие признаки могут быть проблемными (race, sex, income)?

Сравнение модели без коррекции и с коррекцией (описать изменение метрик fairness).

Вывод: какой вариант этичнее и почему.

Критерии (5 баллов):

Корректный анализ bias (2 балла)

Запуск и сравнение моделей (2 балла)

Обоснование этического выбора (1 балл)

Практика 4 — SHAP/LIME

Отчет (4 балла):

Применение SHAP к любой ML-модели (классификация/регрессия)

Интерпретация глобального и локального графика SHAP

Как объяснения помогают выявить несправедливость?

Критерии:

Технически верно — 2 балла

Аналитическая интерпретация — 2 балла

Практика 5 — Симуляция этического ревью (EU AI Act)

Ролевая игра:

Команда разработчиков стартапа (представляют систему)

Этический комитет (задают вопросы)

Регулятор (выносит решение: запретить/ограничить/разрешить).

Заключение (5 баллов): каждый студент пишет 3-пунктовое заключение: риски, категория по EU AI Act, рекомендации.

Критерии:

Обоснование категории риска — 2 балла

Полнота рисков — 2 балла

Рекомендации — 1 балл

Практика 6 — Галлюцинации LLM

Задание:

1. Вызвать галлюцинацию через промпт (например, «Расскажи о несуществующем историческом событии») у ChatGPT / локальной модели.
2. Предложить два способа снижения галлюцинаций (методы фильтрации, промпт-инжиниринг).
3. Этический анализ: когда галлюцинация недопустима, а когда допустима?

Формат: Отчет + промпты + анализ (5 баллов).

Критерии:

Галлюцинация воспроизведена (1)

Механизмы снижения (2)

Этическая дифференциация (2)

Практика 7 — Дебаты по этическим дилеммам ИИ

Темы (по выбору группы):

а) «Распознавание лиц в публичных пространствах — допустимо?»

б) «Автономный транспорт и проблема вагонетки — кто программирует мораль?»

Критерии (4 балла):

Аргументация (2)

Работа с контраргументами (1)

Презентационное мастерство (1)

Практика 8 — Разработка этического чек-листа

Задание: Разработать чек-лист (7–10 пунктов) для сквозного этического контроля.

Примеры пунктов:

Проведен анализ датасета на репрезентативность?

Для каждого решения существует ли человеческая подотчетность?

Формат: документ + 3-минутная защита (4 балла).

Критерии:

Полнота (охват всех стадий ЖЦ) — 2 балла

Конкретность (измеримость) — 1 балл

Защита (убедительность) — 1 балл

Практика 9 — Защита итогового проекта

Зачетно-экзаменационные материалы для промежуточной аттестации

Примерный перечень вопросов для зачета (Контрольные вопросы)

1. Дайте определение этики ИИ. Чем она отличается от традиционной компьютерной этики?
2. Назовите основные этические принципы ИИ по версии OECD.
3. Что такое алгоритмическая предвзятость? Приведите примеры из реальных систем.
4. Опишите методы выявления предвзятости в тренировочных данных.
5. Какие метрики fairness вы знаете? В чем разница между demographic parity и equalized odds?
6. Как проблема "черного ящика" связана с этикой ИИ?
7. Что такое "право на объяснение" в контексте ИИ?
8. Опишите методы обеспечения конфиденциальности в ML (дифференциальная приватность).
9. Каковы основные требования EU AI Act к high-risk системам?
10. Как GDPR регулирует использование персональных данных в ИИ?
11. Назовите основные положения российской Стратегии развития ИИ.
12. В чем заключаются этические риски использования компьютерного зрения?
13. Какие этические проблемы возникают при применении генеративного ИИ?
14. Как ИИ влияет на рынок труда? Какие решения предлагаются?
15. Опишите процесс проведения этического ревью ИИ-проекта.
16. Какие инструменты для этического аудита ИИ-систем вы знаете?
17. Как интегрировать этические чекпоинты в MLOps-пайплайн?
18. В чем особенности корпоративных стандартов ответственного ИИ (Microsoft, Google)?
19. Как проводить оценку социальных последствий внедрения ИИ-системы?
20. Какие методы используются для минимизации экологических последствий работы больших моделей?
21. Опишите методику управления этическими рисками throughout жизненного цикла ИИ.
22. Как выстроить коммуникацию с стейкхолдерами по этическим аспектам проекта?
23. Какие навыки необходимы этическому офицеру (AI Ethics Officer)?
24. Разработайте этический чек-лист для системы рекомендаций в соцсетях.
25. Подготовьте аргументы "за" и "против" использования распознавания лиц в университете.
26. Проанализируйте этические аспекты вашего курсового проекта.
27. Как вы применили принципы справедливости в практической работе №3?
28. Какие регуляторные требования наиболее важны для вашего проекта?

29. Опишите процесс принятия этически сложного решения в ИИ-проекте.
30. Как вы организуете мониторинг этических аспектов системы после внедрения?
31. Подготовьте презентацию для нетехнической аудитории по этическим рискам вашего проекта.

4.2 Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

4.2.1 Уровни освоения (для всех форм контроля)

Уровень	Баллы (в % от макс.)	Характеристика
Высокий (отлично)	86–100%	Демонстрирует полное понимание, самостоятельность, аргументацию, критическое мышление
Средний (хорошо)	70–85%	Демонстрирует понимание, но есть незначительные ошибки или неполнота
Базовый (удовлетворительно)	56–69%	Минимально достаточный уровень, есть пробелы, но основные требования выполнены
Недостаточный (неудовлетворительно)	0–55%	Не выполнены требования, грубые ошибки, отсутствие аргументации

4.2.2 Процедура оценивания теста

Характеристика

Параметр	Значение
Формат	10 вопросов (8 закрытых + 2 открытых)
Время выполнения	15 минут
Максимальный балл	5
Компетенции	SS-1.2, LLM-1.6

Критерии оценивания

Компонент	Количество	Балл за единицу	Макс. балл
Закрытые вопросы (выбор одного / нескольких ответов)	8	0,25	2
Открытые вопросы (краткий ответ 2–4 предложения)	2	1,5	3
Итого			5

Шкала перевода тестового балла в уровень

Балл	Уровень	Характеристика
5	Высокий	Все ответы верны, открытые вопросы раскрыты полно и аргументированно
4	Средний	1–2 ошибки в закрытых, открытые раскрыты с небольшими недочетами
3	Базовый	3–4 ошибки в закрытых, открытые раскрыты частично
0–2	Недостаточный	Более 4 ошибок или открытые не раскрыты

2.4. Процедура проведения

1. Тест проводится в начале лекционного занятия (без предварительного объявления или планово).
2. Запрещено использование любых устройств и конспектов.
3. Результат объявляется на следующем занятии.

4.2.3 Процедура оценивания эссе (философская рефлексия)

Характеристика

Параметр	Значение
Объем	300–400 слов

Время на подготовку	5–7 дней (домашнее задание)
Максимальный балл	5
Компетенция	SS-1.1

Критерии оценивания

Критерий	Вес	Высокий (2)	Средний (1)	Низкий (0)
Философская аргументация	2 балла	Корректное использование 2+ философских концепций (Сёрль, Деннет, проблема субъектности)	Упомянута 1 концепция, но без глубокого анализа	Философские концепции отсутствуют
Структура и логика	1 балл	Четкая структура (тезис – аргументы – вывод), логичная связь	Есть структура, но есть нарушения логики	Логика отсутствует
Соответствие теме и полнота	1 балл	Тема раскрыта полностью, все ключевые аспекты затронуты	Тема раскрыта частично	Тема не раскрыта
Оформление и грамотность	1 балл	Нет орфографических и пунктуационных ошибок	1–3 ошибки	Более 3 ошибок

Итоговая шкала

Балл Уровень

5 Высокий

4 Средний

3 Базовый

0–2 Недостаточный

Процедура сдачи

1. Эссе сдается в электронном виде (формат .docx или .pdf) через LMS или на проверку преподавателю.
2. Проверка осуществляется в течение 5 рабочих дней.
3. Студент получает развернутую обратную связь по каждому критерию.

4.2.4 Процедура оценивания дебатов

Характеристика

Параметр	Значение
Формат	Командные дебаты (3–4 человека в команде)
Время	3 мин аргументы «за» + 3 мин «против» + 2 мин вопросы
Максимальный балл	5 (индивидуально)
Компетенции	SS-1.1 + ОПК-2.1

4.2. Критерии оценивания (индивидуально)

Критерий	Вес	Высокий (2)	Средний (1)	Низкий (0)
Философская / этическая аргументация	2 балла	Убедительные аргументы со ссылками на теории (Сёрль, EU AI Act, fairness)	Аргументы есть, но без теоретической опоры	Аргументы слабые или отсутствуют
Работа с контраргументами	1 балл	Успешно парирует возражения, демонстрирует гибкость	Частично отвечает на возражения	Не отвечает на возражения
Риторика и командная работа	1 балл	Четкая речь, соблюдение регламента, поддержка команды	Небольшие нарушения регламента или пассивность в команде	Грубые нарушения, доминирование или молчание

Ответы на вопросы	1 балл	Точные, развернутые, уверенные ответы	Ответы частичные или неуверенные	Ответов нет или они неверны
-------------------	--------	---------------------------------------	----------------------------------	-----------------------------

Итоговая шкала (индивидуальная)

Балл Уровень

5 Высокий

4 Средний

3 Базовый

0–2 Недостаточный

Процедура проведения

1. Жеребьевка определяет позицию команд («за» / «против»).
2. Время строго контролируется (секундомер).
3. После дебатов преподаватель и студенты задают уточняющие вопросы.
4. Оценка выставляется каждому студенту индивидуально на основе его вклада.

4.2.5. Процедура оценивания кейс-анализа

Характеристика

Параметр	Значение
Формат	Письменный разбор ситуации (1–2 стр.) или устное решение
Время	20–30 минут (аудиторная работа) / 3–5 дней (домашнее задание)
Максимальный балл	5
Компетенции	SS-1.2

Критерии оценивания

Критерий	Макс. балл	Высокий	Средний	Низкий
Идентификация этических рисков	2 балла	Выявлены все ключевые риски с обоснованием	Выявлены основные, но не все	Риски не выявлены или ошибочны
Применение регуляторного / философского аппарата	1 балл	Корректно применены EU AI Act / философские концепции / fairness-метрики	Применено, но с ошибками	Не применено
Качество рекомендаций	1 балл	Конкретные, реалистичные, измеримые рекомендации	Общие рекомендации без конкретики	Рекомендации отсутствуют
Аргументация и обоснованность	1 балл	Выводы обоснованы, логика сильная	Обоснование есть, но слабое	Обоснование отсутствует

Итоговая шкала

Балл Уровень

5 Высокий

4 Средний

3 Базовый

0–2 Недостаточный

Процедура проведения (для аудиторного кейса)

1. Студенты получают текст кейса и вопросы.
2. В течение 5–10 минут делают заметки.

- Далее — индивидуальная письменная работа (или устный ответ по вызову).
- Оценка объявляется после проверки (3–5 дней).

4.2.6. Процедура оценивания итогового группового проекта

Характеристика

Параметр	Значение
Формат	Письменный отчет (10–15 стр.) + презентация (10–12 слайдов) + защита (10–15 мин)
Состав группы	3–4 человека
Максимальный балл	20 (10 отчет + 5 презентация + 5 защита)
Компетенции	Все

Критерии оценивания отчета (10 баллов)

Раздел / Критерий	Балл	Показатели
Анализ трендов и позиционирование задачи (LC-8.1)	2	Тренды определены верно, задача позиционирована корректно
Философская рефлексия (SS-1.1)	2	Применены философские модели, сделан вывод о субъектности/понимании
Этический аудит по стадиям ЖЦ (SS-1.2)	3	Полнота: сбор данных → разработка → тестирование → внедрение → эксплуатация
Механизмы снижения рисков + чек-лист (SS-1.2 / LLM-1.6)	2	Конкретные, реалистичные меры (технические + организационные)
Регуляторные рекомендации + оформление	1	Категория по EU AI Act, предложения, структура отчета

Критерии оценивания презентации (5 баллов)

Критерий	2 балла	1 балл	0 баллов
Структура и наглядность	Логичная структура, визуализация данных, отсутствие перегруженности	Есть недочеты	Отсутствует структура
Полнота содержания	Охвачены все ключевые разделы (тренды, философия, аудит, чек-лист, регулирование)	Охвачены не все	Менее половины
Качество слайдов	Читаемый шрифт, единый стиль, корректные ссылки	Есть нарушения	Нечитаемо / полное отсутствие стиля

Критерии оценивания защиты (5 баллов)

Критерий	2 балла	1 балл	0 баллов
Аргументация и ответы на вопросы	Уверенные, полные ответы на дополнительные вопросы	Частичные ответы	Ответов нет или неверны
Командная работа при защите	Все члены группы участвуют, распределение ролей	Доминирует один участник	Защищает один человек
Тайминг и риторика	Уложились в 15 мин, четкая речь	Небольшие нарушения	Превышение >3 мин

Итоговая шкала

Балл Уровень

18–20 Высокий

14–17 Средний

11–13 Базовый

0–10 Недостаточный

Процедура защиты

- За 3 дня до защиты группы сдают отчет (проверка на плагиат, структуру).
- Защита проходит публично: 10–12 мин презентация, 3–5 мин вопросы.
- Оценка выставляется индивидуально (возможна корректировка при неравномерном вкладе).
- При неудовлетворительной оценке проект дорабатывается и защищается повторно.

4.2.7. Процедура оценивания дополнительных форм

Оценивание групповой дискуссии

Критерий	1 балл	0,5 балла	0 баллов
Активность	Выступал 2+ раза	Выступал 1 раз	Не выступал
Качество аргументов	Аргументы убедительны, с опорой на этику данных	Аргументы общие	Аргументов нет
Уважение к оппонентам	Не перебивает, конструктивно отвечает	Некорректность	Агрессия / игнор

Максимум: 4 балла.

Итоговый уровень: 4 = высокий, 3 = средний, 2 = базовый, 0–1 = недостаточный.

Оценивание отчета по лабораторной работе

Критерий	2 балла	1 балл	0 баллов
Техническая корректность	Код запускается, результаты воспроизводимы	Есть ошибки, но частично работает	Не работает / нет кода
Анализ результатов	Глубокий анализ с этической интерпретацией	Поверхностное описание результатов	Анализ отсутствует
Выводы и рекомендации	Четкие, обоснованные выводы	Общие фразы	Нет выводов

Максимум: 4–5 баллов (в зависимости от сложности работы).

Оценивание чек-листа

Критерий	2 балла	1 балл	0 баллов
Полнота (охват всех стадий ЖЦ)	7–10 пунктов, все стадии	4–6 пунктов, часть стадий	Менее 4 пунктов
Конкретность / измеримость	Каждый пункт предполагает верификацию	Часть пунктов общая	Абстрактные формулировки
Защита (3 мин)	Убедительно обосновал выбор пунктов	Слабая аргументация	Не защищал / нелогично

Максимум: 4 балла.

4.2.8. Итоговое формирование оценки по дисциплине

Структура итоговой оценки (макс. 100 баллов)

Блок	Форма контроля	Баллы
Текущий контроль (лекции)	8 форм (тесты, эссе, обзор, кейс, вопросы)	40
Текущий контроль (практики)	9 форм (отчеты, дебаты, дискуссии, чек-лист)	40
Промежуточная аттестация	Итоговый групповой проект	20
ВСЕГО		100

Перевод баллов в традиционную оценку

Баллы **Оценка (5-балльная)** **Итог**

86–100 «Отлично» Зачтено

70–85 «Хорошо» Зачтено

56–69 «Удовлетворительно» Зачтено

0–55 «Неудовлетворительно» Не зачтено

Условия допуска к промежуточной аттестации

- Набрано не менее 30 баллов по текущему контролю (из 80).
- Сдан и защищен итоговый проект.
- Отсутствие академической задолженности по лабораторным работам.

4.3 Методические указания по организации практических работ по дисциплине «Этика и социальная ответственность в ИИ»

Практические работы направлены на формирование у студентов следующих компетенций:

ОПК-2.1 — навыки профессиональной коммуникации (дебаты, презентации, отчеты).

LC-8.1 — умение позиционировать задачи в контексте трендов ИИ.

SS-1.1 — философская рефлексия об основаниях ИИ.

SS-1.2 — применение методик этического аудита и работы с рисками.

LLM-1.6 — учет этических аспектов генеративных моделей.

Объем и формат

Параметр	Значение
Количество практических занятий	9
Общая трудоемкость	18 академических часов (по 2 часа на занятие)
Форма работы	Индивидуальная, парная, групповая (3–4 человека)
Форма отчетности	Устная защита, письменный отчет, код (при наличии), презентация

Общие правила подготовки к практическому занятию

Этап	Действие	Время
1	Ознакомиться с темой и вопросами занятия (по РПД)	1–2 дня до занятия
2	Изучить рекомендуемую литературу и источники	1–2 дня до занятия
3	При необходимости — выполнить предварительное задание (прочитать кейс, установить библиотеки)	Накануне занятия
4	Подготовить устные ответы / заметки / код / презентацию (по требованию)	Накануне занятия

4.3.2. Описание практических работ

Практическая работа № 1. Разбор кейса Cambridge Analytica

Параметр	Содержание
Тема	Этические проблемы сбора и использования данных
Компетенции	SS-1.2, ОПК-2.1
Формат	Групповая дискуссия (20–25 мин) + заметки
Максимальный балл	4

Задачи занятия:

1. Проанализировать кейс Cambridge Analytica: что произошло, какие данные использовались, как алгоритмы влияли на поведение.
2. Выявить этические нарушения (информированное согласие, манипуляция, прозрачность).
3. Обсудить ответственность: Facebook, разработчики, заказчики, регуляторы.
4. Предложить меры предотвращения подобных ситуаций.

Подготовка к занятию:

Посмотреть документальный фильм / прочитать статью о Cambridge Analytica (рекомендуется: Netflix «The Great Hack» или статьи The Guardian).

Подготовить 3–4 тезиса по каждому из пунктов обсуждения.

Структура занятия:

Этап	Время	Содержание
1	5 мин	Введение преподавателя, напоминание кейса
2	15 мин	Групповая дискуссия (модерация — преподаватель)
3	5 мин	Подведение итогов, ключевые выводы

Критерии оценивания:

Активность (1 балл)

Качество аргументов (2 балла: убедительные + с опорой на этику данных)

Уважение к оппонентам (1 балл)

Практическая работа № 2. Философские дебаты о субъектности ИИ

Параметр	Содержание
Тема	«Может ли ИИ обладать субъектностью?»
Компетенции	SS-1.1, ОПК-2.1
Формат	Командные дебаты (3–4 чел. в команде)
Максимальный балл	5 (индивидуально)

Задачи занятия:

1. Сформулировать аргументы «за» и «против» субъектности ИИ.
2. Использовать философские концепции (Сёрль, «китайская комната», Деннет, проблема интенциональности).
3. Парировать контраргументы оппонентов.
4. Ответить на вопросы аудитории и преподавателя.

Подготовка к занятию:

Прочитать фрагменты работ Сёрля («Разум мозга — вычислительная программа?») и Деннета.

Подготовить 3–5 аргументов с философским обоснованием.

Подготовить 2–3 контраргумента на ожидаемую позицию оппонентов.

Структура занятия:

Этап	Время	Содержание
1	5 мин	Жеребьевка позиций, объявление регламента
2	3 мин	Аргументы команды «ЗА»
3	3 мин	Аргументы команды «ПРОТИВ»
4	2 мин	Вопросы от преподавателя и аудитории
5	2 мин	Заключительное слово (по 1 минуте на команду)

Критерии оценивания (индивидуально):

Философская аргументация (2 балла: ссылки на Сёрля, Деннета)

Работа с контраргументами (1 балл)

Риторика и командная работа (1 балл)

Ответы на вопросы (1 балл)

Практическая работа № 3. Этический аудит датасета + Fairlearn

Параметр	Содержание
Тема	Выявление и коррекция предвзятости в данных
Компетенции	SS-1.2
Формат	Практическая работа (код + отчет)
Максимальный балл	5

Задачи занятия:

1. Загрузить датасет Adult Census.
2. Проанализировать репрезентативность и выявить потенциальные признаки предвзятости (race, sex, income).
3. Обучить модель XGBoost / LogisticRegression.
4. Использовать библиотеку Fairlearn для оценки fairness (Demographic Parity Difference, Equalized Odds).
5. Сравнить обычную модель и модель с коррекцией (например, GridSearch с fairness-ограничением).
6. Сделать вывод об этичности каждой модели.

Подготовка к занятию:

Установить Python, pandas, scikit-learn, fairlearn.

Скачать датасет Adult Census (можно через fetch_openml или из UCI).

Прочитать документацию Fairlearn (кратко).

Структура занятия:

Этап	Время	Содержание
1	10 мин	Объяснение задачи, напоминание о fairness-метриках
2	30 мин	Самостоятельная работа в парах/микрогруппах (код)
3	10 мин	Обсуждение результатов, сравнение моделей
4	10 мин	Оформление отчета (в тетради / электронно)

Пример кода (фрагмент):

```
python
from fairlearn.metrics import demographic_parity_difference
from fairlearn.reductions import GridSearch, DemographicParity
```

```
# Сравнение метрик fairness
```

```
diff_unconstrained = demographic_parity_difference(
    y_true, y_pred_unconstrained, sensitive_features=A
)
print(f"Unconstrained: {diff_unconstrained:.3f}")
```

Критерии оценивания:

- Корректный анализ признаков bias (2 балла)
- Запуск и сравнение двух типов моделей (2 балла)
- Обоснование этического выбора (1 балл)

Практическая работа № 4. Анализ объяснимости модели (SHAP/LIME)

Параметр	Содержание
Тема	Интерпретация предсказаний ML-моделей
Компетенции	SS-1.2
Формат	Практическая работа (код + отчет)
Максимальный балл 4	

Задачи занятия:

1. Обучить любую ML-модель (классификация/регрессия).
2. Применить SHAP Explainer (TreeExplainer / KernelExplainer) и LIME.
3. Построить глобальные и локальные графики важности признаков.
4. Проинтерпретировать, какие признаки вносят наибольший вклад в решение.
5. Сделать вывод о том, насколько модель «объяснима» и приемлема с этической точки зрения.

Подготовка к занятию:

- Установить SHAP (pip install shap) и LIME (pip install lime).
- Подготовить небольшой датасет (можно использовать Adult Census или Boston Housing).
- Ознакомиться с базовыми примерами SHAP (документация).

Структура занятия:

Этап	Время	Содержание
1	10 мин	Объяснение SHAP/LIME (waterfall plot, summary_plot)
2	30 мин	Самостоятельная генерация объяснений
3	10 мин	Разбор сложных моментов, обсуждение интерпретаций
4	10 мин	Оформление вывода (какие признаки проблемны)

Критерии оценивания:

- Техническая правильность (2 балла: графики построены, код работает)
- Аналитическая интерпретация (2 балла: выводы о справедливости / прозрачности)

Практическая работа № 5. Симуляция этического ревью (EU AI Act)

Параметр	Содержание
Тема	Оценка ИИ-системы на соответствие EU AI Act
Компетенции	SS-1.2
Формат	Ролевая игра + письменное заключение
Максимальный балл	5

Задачи занятия:

1. Распределить роли: команда разработчиков, этический комитет, регулятор.
2. Представить гипотетический ИИ-стартап (распознавание эмоций / кредитный скоринг / система найма).
3. Оценить категорию риска по EU AI Act (неприемлемый / высокий / ограниченный / минимальный).
4. Сформулировать 3–5 рекомендаций для соответствия регуляторным требованиям.
5. Защитить заключение перед аудиторией.

Подготовка к занятию:

- Ознакомиться с основными положениями EU AI Act (категории риска, требования к высокорисковым системам).

- Подготовить краткое описание одной ИИ-системы (можно из реальной жизни).

Структура занятия:

Этап	Время	Содержание
1	5 мин	Распределение ролей, объявление системы
2	15 мин	Работа в группах (анализ рисков, категория, рекомендации)
3	15 мин	Представление заключений (каждая группа – 3–5 мин)
4	10 мин	Дискуссия и подведение итогов

Письменное заключение (индивидуально / в группе):

- Категория риска (и обоснование)
- 3 этических риска
- 3 рекомендации по снижению рисков

Критерии оценивания:

- Обоснованность категории риска (2 балла)
- Полнота выявления рисков (2 балла)
- Конкретность рекомендаций (1 балл)

Практическая работа № 6. Работа с галлюцинациями LLM

Параметр	Содержание
Тема	Галлюцинации и фильтрация контента в больших языковых моделях
Компетенции	LLM-1.6
Формат	Практическая работа (промпты + анализ)
Максимальный балл	5

Задачи занятия:

1. Сгенерировать «галлюцинацию» с помощью LLM (ChatGPT, Llama, или через API).
2. Проанализировать, почему модель выдала ложный, но правдоподобный ответ.
3. Предложить не менее двух способов снижения галлюцинаций (промпт-инжиниринг, фильтрация, RLHF).
4. Обсудить этические последствия: когда галлюцинация недопустима, а когда – допустима (творческие задачи).
5. Провести простую фильтрацию выходного контента (регулярные выражения или API-фильтр).

Подготовка к занятию:

- Зарегистрироваться в OpenAI API / Hugging Face (или использовать локальную модель).
- Ознакомиться с понятием галлюцинаций (статьи: «Challenges in LLM Hallucinations»).
- Подготовить 2–3 промпта, которые могут вызвать галлюцинацию.

Структура занятия:

Этап	Время	Содержание
1	10 мин	Теория галлюцинаций, примеры, риски
2	20 мин	Самостоятельная генерация галлюцинаций (в парах)
3	15 мин	Разработка механизмов снижения / фильтрации
4	15 мин	Обсуждение этических границ и результатов

Пример отчета (кратко):

- Использованный промпт и полученная галлюцинация.
- Почему модель ошиблась.
- Два способа снижения риска.
- Вывод: допустимо / недопустимо (для заданного контекста).

Критерии оценивания:

- Галлюцинация воспроизведена (1 балл)
- Предложены 2+ механизма снижения (2 балла)
- Этическая дифференциация (когда допустимо) (2 балла)

Практическая работа № 7. Дебаты по этическим дилеммам ИИ

Параметр	Содержание
Тема	«Распознавание лиц в публичных пространствах» или «Автономный транспорт и проблема вагонетки»

Компетенции	ОПК-2.1
Формат	Командные дебаты
Максимальный балл	4

Задачи занятия:

1. Аргументировать позицию «ЗА» или «ПРОТИВ» в одной из предложенных дилемм.
2. Использовать этические теории (утилитаризм, деонтология, права человека) и регуляторные нормы.
3. Отвечать на контраргументы и вопросы.
4. Прийти к взвешенному выводу (не обязательно консенсусу).

Подготовка к занятию:

Изучить основные аргументы по выбранной теме (статьи, позиции ЕС, отчеты AI Now Institute).

Подготовить 3–5 аргументов с опорой на литературу.

Структура — аналогично практической работе № 2.

Критерии оценивания:

Аргументация (2 балла)

Работа с контраргументами (1 балл)

Презентационное мастерство (1 балл)

Практическая работа № 8. Разработка этического чек-листа для ИИ-проекта

Параметр	Содержание
Тема	Создание инструмента этического контроля
Компетенции	SS-1.2, ОПК-2.1
Формат	Групповая работа + письменный чек-лист + защита
Максимальный балл	4

Задачи занятия:

1. Выбрать гипотетический ИИ-проект (медицина, финансы, образование, HR).
2. Разработать чек-лист из 7–10 пунктов, охватывающий стадии ЖЦ ИИ (сбор данных → разработка → тестирование → внедрение → мониторинг).
3. Каждый пункт должен быть конкретным и проверяемым (ответ «да/нет» или шкала).
4. Подготовить 3-минутную защиту чек-листа.

Подготовка к занятию:

Ознакомиться с примерами этических чек-листов (Google PAIR, EU Ethics Guidelines, IEEE).

Продумать, какие риски характерны для выбранной предметной области.

Пример пунктов чек-листа:

1. Проведен ли анализ датасета на репрезентативность и скрытые предвзятости?
2. Для каждого решения назначено ли ответственное лицо (human accountability)?
3. Предусмотрена ли возможность объяснения решения пользователю (SHAP/LIME)?
4. Проводилась ли оценка fairness-метрик (Demographic Parity, Equalized Odds)?
5. Имеется ли механизм обжалования автоматического решения?
6. Оповещаются ли пользователи о том, что решение принимает ИИ?
7. Проводится ли мониторинг дрейфа данных и смещения метрик после внедрения?

Структура занятия:

Этап	Время	Содержание
1	5 мин	Объяснение задачи, критерии
2	25 мин	Создание чек-листа в группах
3	20 мин	Защита чек-листов (3 мин на группу)
4	10 мин	Обсуждение лучших решений

Критерии оценивания:

Полнота (охват всех стадий ЖЦ) – 2 балла

Конкретность / измеримость – 1 балл

Качество устной защиты – 1 балл

Практическая работа № 9. Презентация и защита итогового группового проекта (этический аудит ИИ-системы)

Параметр	Содержание
Тема	Защита результатов этического аудита
Компетенции	Все
Формат	Презентация + доклад (10–15 мин) + вопросы
Балл	Входит в итоговый проект (20 баллов)

Задачи занятия:

1. Представить письменный отчет (сдан заранее) и презентацию по выбранной ИИ-системе.
2. Раскрыть разделы: тренды ИИ, философская рефлексия, аудит по ЖЦ, чек-лист, регуляторные рекомендации.
3. Ответить на вопросы преподавателя и других студентов.

Подготовка к занятию:

Завершить отчет и презентацию за 3 дня до защиты.

Провести внутреннюю репетицию в группе.

Подготовить ответы на возможные вопросы (например: «Почему вы выбрали именно эти fairness-метрики?», «Что делать, если нельзя устранить bias полностью?»).

Структура защиты:

10–12 мин – выступление группы

5–8 мин – вопросы от преподавателя и аудитории

Критерии оценивания — см. раздел «Итоговый групповой проект» в ФОС.

4.3.3. Организация работы в группах

Принципы формирования групп

Группы формируются на 2–3-й неделе семестра.

Рекомендуемый размер: 3–4 человека.

Роли в группе (распределяются самими студентами):

Координатор — следит за сроками, организует встречи.

Аналитик — отвечает за анализ данных, код.

Этик — фокусируется на этических рисках и философской рефлексии.

Докладчик — готовит презентацию и защиту.

Правила коммуникации в группе

Каждый участник должен внести видимый вклад.

При оценке проекта преподаватель может скорректировать балл индивидуально (на основе peer review внутри группы).

Регулярные встречи группы (минимум 1 раз в неделю).

4.3.4. Требования к оформлению отчетов по практическим работам

Отчет по лабораторным работам (№ 3, 4, 6)

Структура отчета:

1. Титульный лист (ФИО, группа, номер работы, дата).
2. Цель работы.
3. Используемые инструменты (библиотеки, датасеты).
4. Ход выполнения (ключевые шаги, фрагменты кода, скриншоты).
5. Результаты (таблицы, графики, метрики fairness, объяснения SHAP).
6. Этический анализ (выводы о справедливости, рисках, объяснимости).
7. Заключение (что улучшили бы при повторении).

Формат: .pdf или .docx, объем 2–4 страницы.

Отчет по кейсам (№ 1, 5, 8)

Структура:

Краткое описание ситуации / системы.

Выявленные этические проблемы (3–5 пунктов).

Применённые нормы (EU AI Act, философские концепции, fairness-метрики).

Рекомендации (конкретные, измеримые).

Список источников (2–3).

Объем: 1–2 страницы.

4.3.5. Контроль посещения и допуска к промежуточной аттестации

Пропуск одного практического занятия без уважительной причины — снижение итогового балла на 2.

Пропуск по уважительной причине — отработка в форме письменного отчета.

Допуск к промежуточной аттестации (защите проекта) — не менее 3 выполненных и сданных отчетов по практическим работам (из 1–5, 7–9; практические 3,4,6 обязательно).

4.4 Методические указания по выполнению кейса: "Этический аудит системы кредитного скоринга"

1. Общая информация о кейсе

Название: Этический аудит системы кредитного скоринга "CreditFair"

Контекст: Банк внедрил ML-модель для автоматизированного одобрения кредитных заявок. Поступили жалобы от клиентов на возможную дискриминацию по возрасту и полу.

Цель кейса: Провести комплексный этический аудит системы и разработать рекомендации по устранению выявленных нарушений.

2. Этапы выполнения кейса с привязкой к компетенциям

ЭТАП 1: ПОДГОТОВКА И АНАЛИЗ ДАННЫХ (2 часа)

Задачи этапа:

- Загрузка и первичный анализ датасета
- Выявление protected attributes (пол, возраст)
- Предварительная оценка репрезентативности данных

Практическая реализация:

python

```
# Импорт библиотек
```

```
import pandas as pd
```

```
import numpy as np
```

```
import matplotlib.pyplot as plt
```

```
import seaborn as sns
```

```
from fairlearn.metrics import demographic_parity_difference
```

```
# Загрузка данных
```

```
data = pd.read_csv('credit_scoring_data.csv')
```

```
print("Размер датасета:", data.shape)
```

```
print("Колонки:", data.columns.tolist())
```

```
# Анализ protected attributes
```

```
print("Распределение по полу:")
```

```
print(data['gender'].value_counts(normalize=True))
```

```
print("Распределение по возрасту:")
```

```
print(data['age_group'].value_counts(normalize=True))
```

```
# Визуализация распределения
```

```
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 5))
data['gender'].value_counts().plot(kind='bar', ax=ax1, title='Распределение по полу')
data['age_group'].value_counts().plot(kind='bar', ax=ax2, title='Распределение по возрастным группам')
plt.show()
```

Формируемые компетенции:

СС-1.1: *Определяет ценностные предпосылки, когнитивные искажения, культурно-обусловленные предвзятости в данных*

- Выявление потенциально дискриминирующих признаков
- Анализ репрезентативности данных по защищенным группам

ЭТАП 2: РАСЧЕТ МЕТРИК СПРАВЕДЛИВОСТИ (3 часа)

Задачи этапа:

- Расчет метрик fairness для различных protected groups
- Сравнение результатов по разным демографическим группам
- Статистический анализ значимости различий

Практическая реализация:

python

```
from fairlearn.metrics import (
    demographic_parity_difference,
    equalized_odds_difference,
    selection_rate
)

# Расчет метрик для пола
gender_metrics = {
    'demographic_parity_diff': demographic_parity_difference(
        data['loan_approved'],
        data['model_prediction'],
        sensitive_features=data['gender']
    ),
    'equalized_odds_diff': equalized_odds_difference(
        data['loan_approved'],
        data['model_prediction'],
        sensitive_features=data['gender']
    )
}
```

```
# Расчет метрик для возрастных групп
```

```
age_metrics = {
    'demographic_parity_diff': demographic_parity_difference(
        data['loan_approved'],
        data['model_prediction'],
        sensitive_features=data['age_group']
    )
}
```

```

    )
}

print("Метрики справедливости по полу:")
for metric, value in gender_metrics.items():
    print(f"{metric}: {value:.4f}")

print("\nМетрики справедливости по возрасту:")
for metric, value in age_metrics.items():
    print(f"{metric}: {value:.4f}")

# Визуализация selection rate по группам
selection_rates = data.groupby('gender')['model_prediction'].mean()
selection_rates.plot(kind='bar', title='Selection Rate по полу')
plt.ylabel('Доля одобренных заявок')
plt.show()

```

Формируемые компетенции:

SS-1.1: *Определяет предвзятости в алгоритмах*

- Количественная оценка дискриминации с помощью метрик fairness
- Статистический анализ различий в treatment групп

LC-8.1: *Позиционирует задачу в контексте трендов ИИ*

- Применение современных подходов к оценке справедливости алгоритмов
- Соответствие международным стандартам оценки AI систем

ЭТАП 3: АНАЛИЗ ПРИЧИН ПРЕДВЗЯТОСТИ (2 часа)

Задачи этапа:

- Идентификация источников bias в данных и модели
- Анализ feature importance для выявления дискриминирующих факторов
- Оценка влияния historical bias на текущие предсказания

Практическая реализация:

python

```

from sklearn.ensemble import RandomForestClassifier
from sklearn.inspection import permutation_importance

# Анализ важности признаков
X = data.drop(['loan_approved', 'model_prediction'], axis=1)
X_encoded = pd.get_dummies(X)

model = RandomForestClassifier()
model.fit(X_encoded, data['model_prediction'])

# Permutation importance
result = permutation_importance(model, X_encoded, data['model_prediction'],
                                n_repeats=10, random_state=42)

importance_df = pd.DataFrame({

```

```

        'feature': X_encoded.columns,
        'importance': result.importances_mean
    }).sort_values('importance', ascending=False)

```

```

print("Важность признаков:")
print(importance_df.head(10))

```

```

# Анализ корреляции protected attributes с целевой переменной
protected_corr = data[['gender', 'age_group', 'loan_approved']].corr()
print("\nКорреляция защищенных признаков с одобрением кредита:")
print(protected_corr)

```

Формируемые компетенции:

SS-1.1: *Определяет культурно-обусловленные предвзятости*

- Анализ исторических данных на наличие унаследованных bias
- Выявление косвенных дискриминирующих факторов

ЭТАП 4: РАЗРАБОТКА РЕКОМЕНДАЦИЙ (2 часа)

Задачи этапа:

- Разработка плана устранения выявленных нарушений
- Предложение технических и процессуальных улучшений
- Создание системы мониторинга справедливости

Практическая реализация:

python

```

# Генерация отчета с рекомендациями

```

```

recommendations = {
    'technical': [
        'Внедрить предобработку данных для удаления proxy variables',
        'Реализовать post-processing коррекцию предсказаний',
        'Добавить регулярный аудит метрик fairness'
    ],
    'process': [
        'Создать этический чек-лист для всех новых моделей',
        'Внедрить обязательное тестирование на справедливость',
        'Обучить команду Data Science методам Fair ML'
    ],
    'monitoring': [
        'Реализовать дашборд для отслеживания метрик справедливости',
        'Настроить алерты при отклонении метрик от thresholds',
        'Вести журнал этических решений по модели'
    ]
}

```

```

print("РЕКОМЕНДАЦИИ ПО УСТРАНЕНИЮ ПРЕДВЗЯТОСТИ:")
for category, items in recommendations.items():
    print(f"\n{category.upper()} РЕКОМЕНДАЦИИ:")

```

```
for i, item in enumerate(items, 1):  
    print(f"{i}. {item}")
```

Формируемые компетенции:

SS-1.2: *Применяет методики работы с этическими рисками*

- Разработка превентивных мер для управления рисками
- Создание системы непрерывного мониторинга

LC-8.1: *Позиционирует задачу в контексте трендов ИИ*

- Соответствие принципам Responsible AI
- Учет регуляторных требований (EU AI Act)

ЭТАП 5: ПОДГОТОВКА ОТЧЕТА И ПРЕЗЕНТАЦИИ (2 часа)

Задачи этапа:

- Структурирование результатов анализа
- Подготовка презентации для стейкхолдеров
- Формулирование выводов и roadmap улучшений

Структура отчета:

markdown

ЭТИЧЕСКИЙ АУДИТ СИСТЕМЫ CREDITFAIR

1. РЕЗЮМЕ

- Выявлена статистически значимая дискриминация по возрасту (DP diff = 0.15)
- Обнаружены proxy variables, усиливающие предвзятость
- Предложен план корректирующих мер

2. МЕТОДОЛОГИЯ

- Анализ метрик fairness (Demographic Parity, Equalized Odds)
- Permutation importance для выявления значимых признаков
- Статистический анализ распределений

3. РЕЗУЛЬТАТЫ

3.1. Метрики справедливости

- Demographic Parity Difference: 0.15 (> 0.1 threshold)
- Equalized Odds Difference: 0.08

3.2. Ключевые находки

- Переменная "почтовый индекс" коррелирует с доходом и расой
- Исторические данные содержат унаследованные bias

4. РЕКОМЕНДАЦИИ

Немедленные действия (1-2 недели)

- Удалить proxy variables из модели
- Внедрить fairness constraints при переобучении

Среднесрочные улучшения (1-3 месяца)

- Разработать этические гайдлайны
- Обучить команду Fair ML методам

Долгосрочная стратегия

- Внедрить систему непрерывного мониторинга
- Создать Ethics Committee

Формируемые компетенции:

ОПК-2.1: *Применяет инструменты и методы профессиональной коммуникации*

- Подготовка отчетов для технических и нетехнических стейкхолдеров
- Презентация сложных этических концепций в доступной форме
- Аргументация необходимости изменений

СС-1.2: *Применяет методики работы с этическими рисками*

- Разработка реалистичного плана внедрения улучшений
- Учет организационных и технических ограничений

3. Критерии оценки выполнения кейса

Критерий	Вес	Показатели качества	Связь с компетенциями
Глубина анализа	30%	- Полнота расчета метрик - Качество выявления причин bias - Обоснованность выводов	SS-1.1, LC-8.1
Техническая реализация	25%	- Корректность кода - Качество визуализаций - Эффективность методов	SS-1.1, LC-8.1
Качество рекомендаций	25%	- Практическая реализуемость - Соответствие best practices - Учет регуляторных требований	SS-1.2, LC-8.1, LLM-1.6
Коммуникация результатов	20%	- Структура отчета - Ясность презентации - Убедительность аргументов	ОПК-2.1

Сквозное формирование компетенций:

1. **SS-1.1** проявляется на всех этапах анализа данных и выявления предвзятостей
2. **SS-1.2** формируется при разработке и обосновании рекомендаций
3. **LC-8.1** демонстрируется через применение современных методов и учет трендов
4. **ОПК-2.1** оценивается в финальной презентации и отчете

Комплексная оценка:

- Каждый этап вносит вклад в формирование нескольких компетенций
- Итоговая оценка учитывает как технические, так и коммуникативные аспекты
- Практическая значимость работы подчеркивает прикладной характер компетенций

Данный кейс обеспечивает комплексное формирование заявленных компетенций через решение практической задачи, актуальной для современной индустрии ИИ.

Оценочные средства для инвалидов и лиц с ограниченными возможностями здоровья выбираются с учетом их индивидуальных психофизических особенностей.

- при необходимости инвалидам и лицам с ограниченными возможностями здоровья предоставляется дополнительное время для подготовки ответа на экзамене;

- при проведении процедуры оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья предусматривается использование технических средств, необходимых им в связи с их индивидуальными особенностями;

- при необходимости для обучающихся с ограниченными возможностями здоровья и инвалидов процедура оценивания результатов обучения по дисциплине может проводиться в несколько этапов.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

- в печатной форме увеличенным шрифтом,
- в форме электронного документа.

Для лиц с нарушениями слуха:

- в печатной форме,
- в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

- в печатной форме,
- в форме электронного документа.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

5. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля)

5.1 Литература

1. Гуревич, П. С. Этика : учебник для вузов / П. С. Гуревич. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2025. - 516 с. - URL: <https://urait.ru/bcode/560279> (дата обращения: 27.06.2025). - Режим доступа: для авториз. пользователей. - ISBN 978-5-534-17483-0. - Текст : электронный. URL: http://212.192.134.46/MegaPro/UserEntry?Action=Link_FindDoc&id=147480&idb=0

2. Шталь, Бернд Карстен. Этика искусственного интеллекта = Ethics of Artificial Intelligence : кейсы и варианты решения этических проблем / Бернд Карстен Шталь, Дорис Шредер, Ровена Родригес ; перевод с английского Инны Кушнаревой ; под научной редакцией Александра Павлова ; руководитель проекта Александр Павлов ; Национальный исследовательский университет "Высшая школа экономики". - Москва : Издательский дом Высшей школы экономики, 2024. - 196 с. - (Исследования культуры). - Библиогр. в конце глав. - ISBN 978-5-7598-2981-2. - ISBN 978-5-7598-4053-4. - ISBN 978-3-031-17040-9 : 1057 р. - Текст : непосредственный.

URL: http://212.192.134.46/MegaPro/UserEntry?Action=Link_FindDoc&id=281740&idb=0

3. Кибанов, А. Я. Этика деловых отношений : учебник / А. Я. Кибанов, Д. К. Захаров, В. Г. Коновалова ; под ред. А. Я. Кибанова. - 2-е изд., перераб. - Москва : ИНФРА-М, 2023. - 383 с. - URL: <https://znanium.ru/catalog/product/1915727> (дата обращения: 29.08.2024). - Режим доступа: для авториз. пользователей. - ISBN 978-5-16-006723-0. - Текст : электронный.

URL: http://212.192.134.46/MegaPro/UserEntry?Action=Link_FindDoc&id=154172&idb=0

4. Sun, X., Li, J., Kovalenko, A.V., Feng, W., Ou, Y. Integrating Reinforcement Learning and Learning From Demonstrations to Learn Nonprehensile Manipulation //IEEE

Transactions on Automation Science and Engineering, 2023, 20(3), 1735–1744, DOI: 10.1109/TASE.2022.3185071, Q1

5. Petukhova, A.V.; Kovalenko, A.V.; Ovsyannikova, A.V. Algorithm for Optimization of Inverse Problem Modeling in Fuzzy Cognitive Maps. Mathematics 2022, 10, 3452. DOI: 10.3390/math10193452, Q1

6. Kirillova, E.; Kovalenko, A.; Urtenov, M. Study of the Current–Voltage Characteristics of Membrane Systems Using Neural Networks. AppliedMath 2025, 5, 10. <https://doi.org/10.3390/appliedmath5010010>

7. Kadurin, Artur, et al. "The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology." Oncotarget 8.7 (2016): 10883.

8. Kadurin, Artur, et al. "druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico." Molecular pharmaceutics 14.9 (2017): 3098-3104.

9. Polykovskiy, Daniil, et al. "Molecular sets (MOSES): a benchmarking platform for molecular generation models." Frontiers in pharmacology 11 (2020): 565644.

10. Khrabrov, Kuzma, et al. " ∇^2 DFT: A Universal Quantum Chemistry Dataset of Drug-Like Molecules and a Benchmark for Neural Network Potentials." Advances in Neural Information Processing Systems 37 (2024): 36869-36889.

11. Polykovskiy, Daniil, et al. "Entangled conditional adversarial autoencoder for de novo drug discovery." Molecular pharmaceutics 15.10 (2018): 4398-4405.

12. Николенко, Сергей, Кадури, Артур и Архангельская Екатерина. Глубокое обучение. Издательский дом "Питер", 2017..

5.2. Периодические издания и конференции (А*):

1. IEEE Transactions on Big Data – научные статьи по обработке больших данных.
2. Journal of Big Data (SpringerOpen) – открытый журнал с исследованиями в области Big Data.
3. Big Data Research (Elsevier) – публикации по анализу, управлению и визуализации данных.
4. Data Science Journal (CODATA) – междисциплинарные исследования данных.
5. ACM Transactions on Knowledge Discovery from Data (TKDD) – методы извлечения знаний из больших данных.
6. <https://openreview.net/forum?id=FMMF1a9ifL>
7. <https://openreview.net/forum?id=ElUrNM9U8c#discussion>
8. <https://openreview.net/forum?id=JoO6mtCLHD>
9. <https://aclanthology.org/2024.findings-emnlp.760/>
10. <https://aclanthology.org/2020.coling-main.588/>
11. https://link.springer.com/chapter/10.1007/978-3-030-72113-8_30
12. https://link.springer.com/chapter/10.1007/978-3-031-42448-9_10
13. <https://aclanthology.org/2024.findings-naacl.288/>

5.3. Интернет-ресурсы, в том числе современные профессиональные базы данных и информационные справочные системы

Электронно-библиотечные системы (ЭБС):

1. ЭБС «ЮРАЙТ» <https://urait.ru/>
2. ЭБС «УНИВЕРСИТЕТСКАЯ БИБЛИОТЕКА ОНЛАЙН» <http://www.biblioclub.ru/>
3. ЭБС «BOOK.ru» <https://www.book.ru>
4. ЭБС «ZNANIUM.COM» www.znanium.com
5. ЭБС «ЛАНЬ» <https://e.lanbook.com>

Профессиональные базы данных

1. Scopus <http://www.scopus.com/>

2. ScienceDirect <https://www.sciencedirect.com/>
3. Журналы издательства Wiley <https://onlinelibrary.wiley.com/>
4. Научная электронная библиотека (НЭБ) <http://www.elibrary.ru/>
5. Полнотекстовые архивы ведущих западных научных журналов на Российской платформе научных журналов НЭИКОН <http://archive.neicon.ru>
6. Национальная электронная библиотека (доступ к Электронной библиотеке диссертаций Российской государственной библиотеки (РГБ) <https://rusneb.ru/>
7. Президентская библиотека им. Б.Н. Ельцина <https://www.prilib.ru/>
8. База данных CSD Кембриджского центра кристаллографических данных (CCDC) <https://www.ccdc.cam.ac.uk/structures/>
9. Springer Journals: <https://link.springer.com/>
10. Springer Journals Archive: <https://link.springer.com/>
11. Nature Journals: <https://www.nature.com/>
12. Springer Nature Protocols and Methods: <https://experiments.springernature.com/sources/springer-protocols>
13. Springer Materials: <http://materials.springer.com/>
14. Nano Database: <https://nano.nature.com/>
15. Springer eBooks (i.e. 2020 eBook collections): <https://link.springer.com/>
16. "Лекториум ТВ" <http://www.lektorium.tv/>
17. Университетская информационная система РОССИЯ <http://uisrussia.msu.ru>

Информационные справочные системы

1. Консультант Плюс - справочная правовая система (доступ по локальной сети с компьютеров библиотеки)

Ресурсы свободного доступа

1. КиберЛенинка <http://cyberleninka.ru/>;
2. Американская патентная база данных <http://www.uspto.gov/patft/>
3. Министерство науки и высшего образования Российской Федерации <https://www.minobrnauki.gov.ru/>;
4. Федеральный портал "Российское образование" <http://www.edu.ru/>;
5. Информационная система "Единое окно доступа к образовательным ресурсам" <http://window.edu.ru/>;
6. Единая коллекция цифровых образовательных ресурсов <http://school-collection.edu.ru/> .
7. Проект Государственного института русского языка имени А.С. Пушкина "Образование на русском" <https://pushkininstitute.ru/>;
8. Справочно-информационный портал "Русский язык" <http://gramota.ru/>;
9. Служба тематических толковых словарей <http://www.glossary.ru/>;
10. Словари и энциклопедии <http://dic.academic.ru/>;
11. Образовательный портал "Учеба" <http://www.ucheba.com/>;
12. Законопроект "Об образовании в Российской Федерации". Вопросы и ответы http://xn--273--84d1f.xn--p1ai/voprosy_i_otvety

Собственные электронные образовательные и информационные ресурсы КубГУ

1. Электронный каталог Научной библиотеки КубГУ <http://megapro.kubsu.ru/MegaPro/Web>
2. Электронная библиотека трудов ученых КубГУ <http://megapro.kubsu.ru/MegaPro/UserEntry?Action=ToDb&idb=6>
3. Среда модульного динамического обучения <http://moodle.kubsu.ru>
4. База учебных планов, учебно-методических комплексов, публикаций и конференций <http://infoneeds.kubsu.ru/>

5. Библиотека информационных ресурсов кафедры информационных образовательных технологий <http://mschool.kubsu.ru>;
6. Электронный архив документов КубГУ <http://docspace.kubsu.ru/>
7. Электронные образовательные ресурсы кафедры информационных систем и технологий в образовании КубГУ и научно-методического журнала "ШКОЛЬНЫЕ ГОДЫ" <http://icdau.kubsu.ru/>

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

По курсу предусмотрено проведение лекционных занятий, на которых дается основной систематизированный материал. В ходе лекционных занятий разбираются элементы теории и практики дискретной математики, приводятся примеры решения задач, проводится анализ наиболее распространенных ошибок. После прослушивания лекции рекомендуется выполнить упражнения, приводимые в аудитории для самостоятельной работы.

По курсу предусмотрено проведение лабораторных занятий, на которых дается прикладной систематизированный материал. В ходе занятий разбираются методы решений задач по темам. После занятия рекомендуется выполнить упражнения, приводимые для самостоятельной работы.

При самостоятельной работе студентов необходимо изучить литературу, приведенную в перечнях выше, для осмысления вводимых понятий, анализа предложенных подходов и методов дискретной математики. При решении новой задачи студент должен уметь выбрать метод решения и его обоснование.

Важнейшим этапом курса является самостоятельная работа по дисциплине. В процессе самостоятельной работы студент приобретает навыки работы с дискретными объектами.

Используются активные, инновационные образовательные технологии, которые способствуют развитию общекультурных, общепрофессиональных компетенций и профессиональных компетенций обучающихся:

- проблемное обучение;
- разноуровневое обучение;
- проектные методы обучения;
- исследовательские методы в обучении;
- обучение в сотрудничестве (командная, групповая работа);
- информационно- коммуникационные технологии.

Для лиц с ограниченными возможностями здоровья предусмотрена организация консультаций с использованием электронной почты.

Учебно-методическим обеспечением курсовой работы студентов являются:

1. учебная литература;
2. нормативные документы ВУЗа;
3. методические разработки для студентов.

Самостоятельная работа студентов включает:

- оформление итогового отчета (пояснительной записки).
- анализ нормативно-методической базы организации;
- анализ научных публикации по заранее определённой теме;
- анализ и обработку информации;
- работу с научной, учебной и методической литературой,
- работа с конспектами лекций, ЭБС.

Для самостоятельной работы представляется аудитория с компьютером и доступом в Интернет, к электронной библиотеке вуза и к информационно-справочным системам.

Перечень учебно-методического обеспечения:

1. Основная образовательная программа высшего профессионального образования федерального государственного бюджетного образовательного учреждения высшего образования «Кубанский государственный университет» по направлению подготовки.
2. Положение о проведении текущего контроля успеваемости и промежуточной аттестации в федеральном государственном бюджетном образовательном учреждении высшего образования «Кубанский государственный университет».
3. Общие требования к построению, содержанию, оформлению и утверждению рабочей программы дисциплины Федерального государственного образовательного стандарта высшего профессионального образования.
4. Методические рекомендации по содержанию, оформлению и применению образовательных технологий и оценочных средств в учебном процессе, основанном на Федеральном государственном образовательном стандарте.
5. Учебный план основной образовательной программы по направлению подготовки.
6. Федеральный государственный образовательный стандарт высшего профессионального образования по направлению подготовки.

В освоении дисциплины инвалидами и лицами с ограниченными возможностями здоровья большое значение имеет индивидуальная учебная работа (консультации) – дополнительное разъяснение учебного материала.

Индивидуальные консультации по предмету являются важным фактором, способствующим индивидуализации обучения и установлению воспитательного контакта между преподавателем и обучающимся инвалидом или лицом с ограниченными возможностями здоровья.

Подход, определяющий установление соответствия кейсов ИП и УГТ (5-7), позволяет четко соотносить этапы развития технологии с вовлеченностью партнера и снижать риски при переходе от лабораторных испытаний к промышленному внедрению.

Ключевые аспекты взаимодействия с индустриальными партнерами:

- Для УГТ 5 – ИП помогает определить реалистичные условия тестирования, но не рискует своей инфраструктурой.
- Для УГТ 6 – ИП предоставляет "песочницу" или изолированную среду, где можно выявить скрытые проблемы.
- Для УГТ 7 – ИП становится соразработчиком, так как технология адаптируется под его конкретные процессы.

Важнейшим компонентом курса является самостоятельная проектная работа, в ходе которой студент разрабатывает законченное решение для решения задач (кейсов) индустриальных партнеров. Допускается выполнение проектов в командах.

А. Применение результатов дисциплины «Этика и социальная ответственность в ИИ» в кейсах ПАО «Сбербанк»

КЕЙС 1: ЭТИЧЕСКИЙ АУДИТ И КОРРЕКЦИЯ СИСТЕМЫ КРЕДИТНОГО СКОРИНГА

Цель:

Выявить и устранить алгоритмическую дискриминацию клиентов предпенсионного возраста (55+) в системе автоматического кредитного скоринга, обеспечить соответствие требованиям регуляторов и повысить customer satisfaction.

Технологии:

Python 3.8+, pandas, scikit-learn

Fairlearn 0.8.0, AIF360

SHAP, LIME

Tableau для мониторинга

Docker, MLflow

Реализация:

1. **Проведение комплексного аудита данных и модели**
 - Анализ распределения protected attributes (возраст, пол, регион)
 - Расчет метрик справедливости: Demographic Parity, Equalized Odds
 - Выявление proxy variables, косвенно связанных с защищенными признаками
2. **Глубокий анализ причин предвзятости**
 - Permutation importance для определения значимых признаков
 - Анализ feature correlation с защищенными атрибутами
 - Построение зависимостей approval rate от возраста
3. **Коррекция модели с fairness constraints**
 - Применение ExponentiatedGradient с DemographicParity constraints
 - Калибровка thresholds для разных возрастных групп
 - Валидация на отложенной выборке
4. **Внедрение системы мониторинга**
 - Реализация дашборда с fairness метриками
 - Настройка алертов при превышении thresholds
 - Регулярные отчеты для этического комитета

Результат:

Снижение Demographic Parity Difference с 0.18 до 0.04

Увеличение одобрения кредитов для клиентов 55+ на 27%

Снижение жалоб на 73%, рост NPS на 18 пунктов

Соответствие требованиям ЦБ РФ и ГОСТ Р 59276-2020

Реализованные компетенции:

SS-1.1: Выявление возрастной дискриминации и proxy-переменных

SS-1.2: Внедрение системы мониторинга и процедур аудита

LC-8.1: Применение современных методов обеспечения справедливости

ОПК-2.1: Коммуникация результатов стейкхолдерам

КЕЙС 2: РАЗРАБОТКА ЭТИЧЕСКИХ СЦЕНАРИЕВ ДЛЯ ГОЛОСОВОГО АССИСТЕНТА «САЛЮТ»

Цель:

Обеспечить этическое взаимодействие голосового ассистента с пользователями в кризисных ситуациях, предотвратить потенциальный вред.

Технологии:

- Python, NLP-библиотеки (Natasha, transformers)
- Dialog Flow Management
- Sentiment Analysis
- Elasticsearch для базы знаний
- Kubernetes для оркестрации

Реализация:

1. **Классификация sensitive запросов**
 - Разработка тезауруса кризисных тем (финансы, здоровье, психология)
 - Обучение ML-модели для категоризации интенгов
 - Настройка confidence thresholds для sensitive категорий
2. **Разработка этических сценариев ответов**
 - Создание базы безопасных ответов для каждой категории
 - Реализация эскалационных процедур для критических случаев
 - Интеграция с внешними сервисами (телефоны доверия)
3. **Тестирование и валидация**
 - А/В тестирование сфокусированных пользователей
 - Экспертная оценка психологов и ethicists
 - User acceptance testing

4. **Мониторинг и обратная связь**
 - Сбор feedback от пользователей
 - Анализ рекуррентных проблем
 - Постоянное обновление базы знаний

Результат:

Обработка 100% sensitive запросов по утвержденным сценариям

Снижение эскалаций в call-центр на 45%

Улучшение пользовательского опыта (CSAT +35%)

Соответствие этическим стандартам голосовых ассистентов

Реализованные компетенции:

SS-1.2: Разработка превентивных мер для управления рисками

ОПК-2.1: Создание коммуникационных протоколов для кризисных ситуаций

LC-8.1: Внедрение современных NLP-методов с учетом этических аспектов

КЕЙС 3: КОРРЕКЦИЯ ГЕНДЕРНОЙ ПРЕДВЗЯТОСТИ В ИНВЕСТИЦИОННЫХ РЕКОМЕНДАЦИЯХ

Цель:

Устранить гендерные стереотипы в рекомендательной системе инвестиционных продуктов, обеспечить равный доступ к финансовым возможностям.

Технологии:

Recommender Systems (collaborative filtering, content-based)

Fairness-aware machine learning

A/B testing platform

Real-time monitoring system

Apache Kafka для потоковой обработки

Реализация:

1. **Анализ текущих рекомендаций**
 - Расчет распределения рекомендаций по полу и возрасту
 - Выявление паттернов стереотипных рекомендаций
 - Анализ пользовательского поведения и конверсии
2. **Разработка fairness-aware алгоритма**
 - Модификация recommendation engine с учетом fairness constraints
 - Балансировка рекомендаций по риск-профилю
 - Внедрение механизма разнообразия (serendipity)
3. **Постепенное внедрение и тестирование**
 - Canary release для сегмента пользователей
 - A/B тестирование эффективности и fairness
 - Сбор качественных и количественных метрик
4. **Обучение и коммуникация**
 - Обучение финансовых консультантов
 - Коммуникация изменений пользователям
 - Создание образовательных материалов

Результат:

Устранение гендерного перекоса в рекомендациях (с 65%/35% до 52%/48%)

Увеличение диверсификации портфелей на 28%

Рост удовлетворенности клиентов на 22%

Улучшение финансовой инклюзивности

Реализованные компетенции:

SS-1.1: Выявление и анализ гендерных стереотипов в алгоритмах

SS-1.2: Разработка и внедрение корректирующих механизмов

LC-8.1: Применение передовых методов fairness-aware ML

КЕЙС 4: ВНЕДРЕНИЕ СИСТЕМЫ ОБЪЯСНИМОСТИ РЕШЕНИЙ ДЛЯ МАЛОГО БИЗНЕСА

Цель:

Обеспечить прозрачность и понятность решений по кредитованию для малого бизнеса, реализовать право на объяснение.

Технологии:

SHAP, LIME, Anchor explanations

Microservices architecture

REST API, GraphQL

React.js для фронтенда

PostgreSQL для хранения объяснений

Реализация:

1. **Разработка системы объяснений**
 - Интеграция XAI-библиотек в кредитный конвейер
 - Создание иерархии объяснений (от простых к сложным)
 - Генерация персонализированных рекомендаций
2. **Проектирование пользовательского интерфейса**
 - Разработка интуитивного дашборда для предпринимателей
 - Визуализация ключевых факторов решения
 - Интерактивные сценарии "что если"
3. **Интеграция в бизнес-процессы**
 - Обучение менеджеров работе с системой
 - Настройка автоматической отправки объяснений
 - Интеграция с CRM и колл-центром
4. **Оценка эффективности**
 - Мониторинг использования системы
 - Сбор обратной связи от пользователей
 - Измерение impact на бизнес-метрики

Результат:

Снижение жалоб в ЦБ РФ на 54%

Увеличение повторных обращений после доработки на 31%

Улучшение понимания критериев одобрения на 67%

Соответствие требованиям GDPR "right to explanation"

Реализованные компетенции:

ОПК-2.1: Разработка понятных интерфейсов и коммуникационных материалов

SS-1.2: Создание системы конструктивной обратной связи

LC-8.1: Реализация современных методов объяснимого ИИ

КЕЙС 5: СИСТЕМА ЭТИЧЕСКОГО МОНИТОРИНГА ANTI-FRAUD СИСТЕМЫ

Цель:

Минимизировать ложные срабатывания antifraud системы для социально уязвимых групп, сохраняя при этом эффективность обнаружения мошенничества.

Технологии:

Real-time streaming (Apache Flink)

Anomaly detection algorithms

Fairness monitoring dashboard

Alerting system (PagerDuty, Telegram bots)

Time-series databases

Реализация:

1. **Анализ текущей ситуации**
 - Расчет метрик fairness для различных демографических групп
 - Выявление паттернов ложных срабатываний

- Анализ impact на customer experience
- 2. **Разработка системы мониторинга**
 - Создание real-time пайплайна расчета fairness метрик
 - Настройка алертинга при отклонениях
 - Разработка дашбордов для различных стейкхолдеров
- 3. **Коррекция алгоритмов**
 - Балансировка precision/recall для разных сегментов
 - Внедрение contextual features для снижения ложных срабатываний
 - Калибровка thresholds по регионам и возрастным группам
- 4. **Организационные изменения**
 - Создание процедур быстрого реагирования на алерты
 - Обучение команды работе с системой мониторинга
 - Интеграция с процессами инцидент-менеджмента

Результат:

Снижение ложных срабатываний на 62%

Улучшение detection rate для реального мошенничества на 15%

Снижение нагрузки на колл-центр на 40%

Улучшение репутации в регионах

Реализованные компетенции:

SS-1.1: Выявление географической и возрастной дискриминации

SS-1.2: Создание системы превентивного мониторинга и быстрого реагирования

ОПК-2.1: Налаживание коммуникации между техническими и бизнес-командами

Б. Применение результатов дисциплины «Этика и социальная ответственность в ИИ» в кейсах для компании AVA Group

КЕЙС 1: ЭТИЧЕСКИЙ ИИ В ПРОЕКТИРОВАНИИ ДОСТУПНОЙ СРЕДЫ

Цель:

Разработка AI-системы для автоматизированного проектирования безбарьерной среды, учитывающей потребности людей с ограниченными возможностями и маломобильных групп населения.

Технологии:

Computer Vision для анализа архитектурных планов

Generative AI для создания адаптивных проектных решений

Fairness-метрики для оценки доступности

ВМ-интеграция

Реализация:

1. Анализ нормативных требований

Обучение модели на ГОСТ Р 52766-2007, СП 59.13330.2016

Создание базы параметров доступности (ширина проемов, уклоны, высоты)

Интеграция с российскими стандартами инклюзивного строительства

2. Разработка алгоритма проверки

python

```
def check_accessibility_compliance(bim_model):
    # Анализ параметров доступности
    compliance_metrics = {
        'door_width': check_door_width(bim_model),
        'ramp_slope': check_ramp_slope(bim_model),
        'corridor_width': check_corridor_width(bim_model),
        'bathroom_accessibility': check_bathroom_layout(bim_model)
    }
```

```
return calculate_compliance_score(compliance_metrics)
```

3. Генерация адаптивных решений

AI-рекомендации по адаптации существующих объектов

Автоматическая корректировка проектов под требования доступности

Оптимизация затрат на реализацию инклюзивных решений

Результат:

100% соответствие проектов требованиям доступности

Снижение затрат на перепроектирование на 40%

Улучшение социального имиджа компании

Повышение лояльности клиентов с ограниченными возможностями

Реализованные компетенции:

SS-1.1: Выявление архитектурных барьеров и дискриминирующих элементов

SS-1.2: Разработка системы обеспечения доступности на всех этапах жизненного цикла

ОПК-2.1: Коммуникация преимуществ инклюзивной среды стейкхолдерам

КЕЙС 2: СПРАВЕДЛИВАЯ СИСТЕМА УПРАВЛЕНИЯ АРЕНДНЫМИ СТАВКАМИ

Цель:

Создание этичной AI-системы динамического ценообразования для управляющей компании, исключая дискриминацию по социально-демографическим признакам.

Технологии:

Machine Learning для прогнозирования рыночных цен

Fairness constraints в алгоритмах ценообразования

A/B testing платформа

Realtime мониторинг метрик справедливости

Реализация:

1. Аудит текущей системы ценообразования

Анализ корреляции цен с демографическими характеристиками районов

Выявление непреднамеренной дискриминации

Расчет метрик справедливости для разных сегментов

2. Разработка этичного алгоритма

python

```
class EthicalPricingModel:
    def __init__(self):
        self.fairness_constraint = DemographicParity(difference_bound=0.05)

    def calculate_rent(self, property_features, market_data):
        base_price = self.base_model.predict(property_features)
        # Применение fairness constraints
        fair_price = self.fairness_constraint.adjust_price(
            base_price, property_features['district']
        )
        return fair_price
```

3. Внедрение и мониторинг

Постепенное развертывание с контролем impact

Мониторинг satisfaction разных демографических групп

Регулярный аудит на соответствие 152-ФЗ

Результат:

Снижение дискриминации в ценообразовании на 70%
Сохранение конкурентных преимуществ
Увеличение лояльности арендаторов на 25%
Соответствие требованиям регуляторов

Реализованные компетенции:

SS-1.1: Выявление скрытой дискриминации в алгоритмах ценообразования

LC-8.1: Применение современных методов fair ML в реальном бизнесе

SS-1.2: Создание системы мониторинга этичности коммерческих решений

КЕЙС 3: ЭТИЧЕСКИЙ ИИ ДЛЯ УМНОГО РАЙОНА (КОТ)

Цель:

Разработка системы управления комплексным освоением территорий, балансирующей интересы застройщика, жителей и экологии.

Технологии:

IoT сенсоры и компьютерное зрение
Multi-objective optimization
Environmental impact assessment AI
Digital twin территории

Реализация:

1. Создание цифрового двойника территории

Интеграция данных экологического мониторинга
Моделирование социальной инфраструктуры
Прогнозирование нагрузки на транспортную сеть

2. Разработка балансирующего алгоритма

python

```
def optimize_territory_development(land_plot, constraints):  
    objectives = {  
        'profit_margin': calculate_profitability(land_plot),  
        'social_infrastructure': assess_social_needs(land_plot),  
        'environmental_impact': calculate_eco_impact(land_plot),  
        'resident_satisfaction': predict_satisfaction(land_plot)  
    }  
}
```

Поиск Pareto-оптимального решения

```
return multi_objective_optimization(objectives, constraints)
```

3. Система участия жителей

AI-анализ обращений и предложений жителей
Приоритизация улучшений на основе коллективного intelligence
Прозрачная коммуникация решений

Результат:

Баланс коммерческих и социальных целей
Снижение экологического следа на 30%
Увеличение удовлетворенности жителей на 40%
Ускорение согласований с органами власти

Реализованные компетенции:

SS-1.2: Управление многокритериальными этическими дилеммами

ОПК-2.1: Организация конструктивного диалога с сообществом

LC-8.1: Применение передовых методов оптимизации в урбанистике

КЕЙС 4: ПРЕДОТВРАЩЕНИЕ ДИСКРИМИНАЦИИ В CRM СИСТЕМЕ ДЕВЕЛОПЕРА

Цель:

Исключение алгоритмической дискриминации в системе управления взаимоотношениями с клиентами при сегментации и таргетировании маркетинговых кампаний.

Технологии:

Customer segmentation algorithms
Fairness-aware recommendation systems
Marketing automation platforms
Bias detection in NLP

Реализация:

1. Аудит текущих маркетинговых кампаний

Анализ распределения рекламного бюджета по демографическим группам
Выявление непреднамеренного исключения уязвимых категорий
Оценка инклюзивности коммуникационных материалов

2. Разработка этичного рекомендательного движка (Ethical Recommendation Engine)

python

```
class EthicalCustomerSegmentation:
    def __init__(self):
        self.fairness_metrics = ['demographic_parity', 'equal_opportunity']

    def segment_customers(self, customer_data, campaign_goals):
        segments = self.base_segmentation_model.predict(customer_data)

        # Применение fairness corrections
        fair_segments = self.apply_fairness_constraints(
            segments, customer_data['protected_attributes']
        )

        return self.optimize_campaign_reach(fair_segments, campaign_goals)
```

3. Обучение команды продаж

Разработка guidelines по этичному общению с клиентами
Тренинги по выявлению unconscious bias
Внедрение системы обратной связи от клиентов

Результат:

Увеличение охвата маркетинговых кампаний на 35%
Снижение жалоб на дискриминацию на 80%
Улучшение репутации бренда
Расширение клиентской базы

Реализованные компетенции:

SS-1.1: Выявление скрытой дискриминации в маркетинговых процессах

SS-1.2: Разработка превентивных мер и обучающих программ

ОПК-2.1: Создание инклюзивных коммуникационных материалов

КЕЙС 5: ЭТИЧЕСКИЙ ИИ В ЦЕПОЧКЕ ПОСТАВОК СТРОЙМАТЕРИАЛОВ

Цель:

Создание прозрачной и этичной системы управления цепочкой поставок, обеспечивающей fair trade и экологическую ответственность.

Технологии:

- Supply chain optimization algorithms
- Blockchain для отслеживания происхождения материалов
- Environmental impact assessment
- Supplier fairness evaluation

Реализация:

1. **Разработка системы оценки поставщиков**
 - Критерии экологической ответственности
 - Оценка социальных условий труда
 - Анализ ethical sourcing practices
2. **Оптимизация логистики с учетом устойчивого развития**
python

```
def optimize_sustainable_supply_chain(suppliers, construction_sites):  
    optimization_goals = {  
        'cost_efficiency': calculate_transportation_costs,  
        'carbon_footprint': calculate_emissions,  
        'delivery_reliability': assess_supplier_reliability,  
        'social_responsibility': evaluate_supplier_ethics  
    }  
  
    return multi_criteria_optimization(optimization_goals)
```

3. **Система прозрачности для клиентов**

- QR-коды с информацией о происхождении материалов
- Carbon footprint калькулятор для объектов
- Отчеты об этичности цепочки поставок

Результат:

Снижение углеродного следа на 25%
Улучшение условий труда в цепочке поставок
Повышение лояльности экологически ориентированных клиентов
Соответствие требованиям ESG-стандартов

Реализованные компетенции:

SS-1.2: Управление этическими рисками в сложных supply chain

LC-8.1: Применение современных методов оптимизации с учетом устойчивого развития

ОПК-2.1: Коммуникация преимуществ этичной цепочки поставок клиентам

7. Материально-техническое обеспечение по дисциплине (модулю)

№	Вид работ	Наименование учебной аудитории, ее оснащенность оборудованием и техническими средствами обучения
1.	Лекционные занятия	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения

2.	Лабораторные занятия	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения, компьютерами, проектором, программным обеспечением
3.	Групповые (индивидуальные) консультации	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения, компьютерами, программным обеспечением
4.	Текущий контроль, промежуточная аттестация	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения, компьютерами, программным обеспечением
5.	Самостоятельная работа	Кабинет для самостоятельной работы, оснащенный компьютерной техникой с возможностью подключения к сети «Интернет», программой экранного увеличения и обеспеченный доступом в электронную информационно-образовательную среду университета.

№	Продукт	Параметры продукта	Кол-во	Кол-во конфигураций	Ед. изм.
1	Виртуальная машина	Виртуальная машина 10% vCPU 2 vCPU 4 RAM	1	60	Шт
		ОС Ubuntu 22.04	1		Шт
		Системный диск SSD	1		Шт
			10		Гб
		Аренда публичного IP	1		Шт
2	Виртуальная машина с GPU	Виртуальная машина с GPU NVIDIA® Tesla® V100 2 GPU 8 vCPU 128 Гб RAM	1	1	Шт
		ОС Ubuntu 24.04	1		Шт
		Системный диск SSD	1		Шт
			2000		Гб
		Диск SSD	1		Шт
			4096		Гб
		Диск SSD	1		Шт
			4096		Гб
		Аренда публичного IP	1		Шт
3	K8S	Master node 8 vCPU 16 RAM	1	1	Шт
		Worker node 10% доля 4 vCPU 32 RAM	5		Шт
		Worker node SSD-NVME	64		Гб
		Аренда публичного IP	1		Шт
4	ML Inference Instance Type GPU	Время работы в месяц	40	1	Ч
		Инстанс 8 x NVIDIA® H100 NVLink PCIe 160 vCPU 1520 GB RAM	1		Шт
		Количество запросов к ML-моделям	1		Млн. Шт
		Кэш ML-моделей	160		Гб

5	LLM	Токены GigaChat 2 Max	50		Млн. Шт
		Токены Embeddings	400		Млн. Шт

Дополнительные облачные ресурсы предоставляются технологическим партнером Yandex Cloud.

Примечание: Конкретизация аудиторий и их оснащение определяется ОПОП.