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
« 29 » августа 2025 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

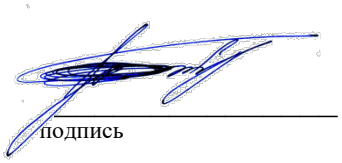
Краснодар 2025

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

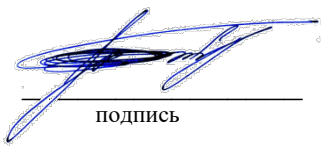
подпись

Рабочая программа дисциплины утверждена на заседании центра искусственного интеллекта
протокол № 01 «28» августа 2025 г.
Руководитель центра ИИ Коваленко А.В.



подпись

Утверждена на заседании учебно-методической комиссии факультета компьютерных технологий и прикладной математики
протокол № 01 «28» августа 2025 г.
Председатель УМК факультета Коваленко А.В.



подпись

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

Луценко Евгений Вениаминович, доктор экономических наук, кандидат технических наук, профессор кафедры компьютерных технологий и систем Федерального государственного бюджетное образовательное учреждение 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 Машинное обучение, Глубокое обучение, Нейросетевые технологии, Обработка данных на Python, Аналитика данных ,Технологии обработки больших данных, Технологии компьютерного зрения, Технологии обработки языка, Генеративный искусственный интеллект, Промпт-инжиниринг.

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	**Знает:** передовые методики анализа долгосрочных социальных последствий внедрения ИС, методы управления конфликтами требований (например, безопасность vs. приватность). **Умеет:** прогнозировать системные этические последствия и возникающие дилеммы на ранних этапах проектирования ИС; разрабатывать и аргументировать стратегию балансировки противоречивых этических требований; интегрировать этический анализ в общий процесс проектирования ИС. **Владеет:** навыками фасилитации междисциплинарного обсуждения требований к ИС и разработки организационных механизмов для управления этическими аспектами на протяжении всего жизненного цикла системы.
ML-1 Способен применять знания об истории развития и трендах современного ИИ для формулирования корректных постановок задач и поиска перспективных способов решения проблем с помощью ИИ	
ML-1.1 Позиционирует собственную задачу в заданной области знания с точки зрения трендов современного искусственного интеллекта	Анализирует современные тренды в области регулирования и этики ИИ (EU AI Act, ГОСТы) при постановке задач разработки ИИ-систем, обосновывая актуальность и корректность их постановки с учетом социального контекста.
SS-1 Способен осуществлять свою трудовую деятельность с учетом определения корректной роли ИИ в различных процессах, критического анализа последствий применения ИИ-технологий, этических принципов	
SS-1.1 Определяет ценностные предпосылки, когнитивные искажения, культурно-обусловленные предвзятости в данных, алгоритмах, постановке задач для ИИ.	Проводит аудит датасетов и моделей на наличие предвзятостей, выявляет и анализирует этические риски, связанные с постановкой задачи и контекстом применения ИИ-системы.
SS-1.2 Применяет методики работы с этическими и социальными рисками, возникающими на разных стадиях жизненного цикла ИИ.	Разрабатывает этические гайдлайны и рекомендации по минимизации рисков для конкретного ИИ-проекта, используя известные frameworks (например, Microsoft RAI, OECD AI Principles).

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	СРС
1	**2**	**3**	**4**	**5**	**6**	**7**
1.	Введение в этику ИИ. Основные концепции и вызовы.	10	2	2		6
2.	Предвзятость и справедливость алгоритмов (Algorithmic Fairness).	12	2	4		6
3.	Конфиденциальность, прозрачность и подотчетность ИИ-систем.	12	2	4		6
4.	Регулирование ИИ: правовые нормы и отраслевые стандарты.	10	2	2		6
5.	Социальные последствия ИИ и управление рисками.	8	2	2		4
6.	Инструменты и практики внедрения ответственного ИИ в проекты.	12	4	4		4
7.	Презентация и защита проектов.	5,8	2	-		3,8
ИТОГО по разделам дисциплины		**69,8**	**16**	**18**		**35,8**
Контроль самостоятельной работы (КСР)		2				
Промежуточная аттестация (ИКР)		0,2				

№	Наименование разделов (тем)	Количество часов				
		Всего	Аудиторная работа			Внеаудиторная работа
			Л	ПЗ	ЛР	СРС
1	2	3	4	5	6	7
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

№	Наименование раздела (темы)	Содержание раздела (темы)	Форма текущего контроля
1	2	3	
1	Введение в этику ИИ. Основные концепции и вызовы.	История и эволюция ИИ-этики. Ключевые понятия: моральный агент, ответственность, справедливость, добросовестность. Актуальные вызовы: автономное оружие, слежка, манипуляция поведением. Связь с направлениями "ИИ и аналитика данных".	Опрос, тест
2	Предвзятость и справедливость алгоритмов (Algorithmic Fairness).	Источники предвзятости (данные, алгоритмы, интерпретация). Математические определения справедливости (demographic parity, equalized odds). Кейсы: COMPAS, Amazon HR AI. Методы выявления и устранения bias (pre-, in-, post-processing).	Опрос
3	Конфиденциальность, прозрачность и подотчетность ИИ-систем.	Проблема "черного ящика". Методы объяснимого ИИ (XAI). Право на объяснение (GDPR). Дифференциальная конфиденциальность (differential privacy). Безопасность ИИ-систем (adversarial attacks).	Тест
4	Регулирование ИИ: правовые нормы и отраслевые стандарты.	Обзор международных инициатив (EU AI Act, OECD Principles). Российское законодательство в сфере ИИ (Стратегия ИИ, ГОСТы). Корпоративные стандарты (Microsoft RAI, Google's AI Principles).	Опрос
5	Социальные последствия ИИ и управление рисками.	ИИ и будущее труда. Автоматизация и безработица. Цифровое неравенство. Долгосрочные риски и дебаты о AGI. Роль публичной дискуссии.	Опрос
6	Инструменты и практики внедрения ответственного ИИ в проекты.	Модель жизненного цикла ответственного ИИ. Инструменты для аудита (Fairlearn, AIF360, IBM AI Fairness 360). Разработка этических гайдлайнов. Проведение этических ревью.	Подготовка проекта
7	Презентация и защита проектов.	Разбор лучших практик подготовки и проведения презентаций. Критерии оценки проектов.	Защита проекта

2.3.2 Занятия семинарского типа / лабораторные занятия

№	Наименование раздела (темы)	Наименование лабораторных работ	Форма текущего контроля
1	2	3	4
1	Введение в этику ИИ.	**Практическая работа 1:** Разбор кейса Cambridge Analytica. Дискуссия о роли данных и алгоритмов в манипуляции общественным мнением.	Отчет, участие в дискуссии
2	Предвзятость и справедливость алгоритмов.	**Практическая работа 2:** "Этический аудит датасета". Анализ открытого датасета (Adult Census) на предмет репрезентативности и скрытых предвзятостей.	Отчет
		Практическая работа 3: "Сравнение моделей с коррекцией bias". Использование библиотеки Fairlearn для обучения и сравнения "справедливой" и "обычной" моделей.	Отчет, код
3	Конфиденциальность, прозрачность и подотчетность.	**Практическая работа 4:** "Анализ объяснимости модели". Применение методов SHAP/LIME для интерпретации предсказаний ML-модели.	Отчет
4	Регулирование ИИ.	**Практическая работа 5:** "Симуляция этического ревью". Ролевая игра: оценка гипотетического ИИ-стартапа с точки зрения соответствия EU AI Act.	Протокол заседания, презентация
5	Социальные последствия ИИ.	**Практическая работа 6:** "Дебаты по этическим дилеммам ИИ". Командные дебаты на темы: "Распознавание лиц в публичных пространствах", "Автономный транспорт и проблема вагонетки".	Участие в дебатах
6	Инструменты и практики.	**Практическая работа 7:** "Разработка этического чек-листа для проекта". Создание и защита чек-листа для сквозного этического контроля в гипотетическом проекте.	Чек-лист, презентация
7	Презентация и защита проектов.	**Практическая работа 8:** Презентация и защита итоговых групповых проектов по этическому аудиту ИИ-системы.	Защита проекта, отчет

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 ее части)	Наименование оценочного средства	
			Текущий контроль	Промежуточная аттестация
1	Введение в этику ИИ. Основные концепции и вызовы	ОПК-2.1, ML-1.1	Практическая работа 1	Контрольные вопросы (1, 2, 26)
2	Предвзятость и справедливость алгоритмов	SS-1.1, SS-1.2	Практические работы 2, 3	Контрольные вопросы (3-8, 27)
3	Конфиденциальность, прозрачность и подотчетность	SS-1.1, SS-1.2	Практическая работа 4	Контрольные вопросы (9-14, 28)
4	Регулирование ИИ: правовые нормы и стандарты	ML-1.1, ОПК-2.1	Практическая работа 5	Контрольные вопросы (15-19, 29)
5	Социальные последствия ИИ и управление рисками	SS-1.1, ML-1.1	Практическая работа 6	Контрольные вопросы (20-23, 30)
6	Инструменты и практики внедрения ответственного ИИ	SS-1.2, ОПК-2.1	Практическая работа 7	Контрольные вопросы (24, 25, 31)
7	Презентация и защита проектов	ОПК-2.1, SS-1.2	Защита итогового проекта	Защита проекта

Показатели, критерии и шкала оценки сформированных компетенций

№ п/п	Код и наименование индикатора	Результаты обучения	Наименование оценочного средства	
			Текущий контроль	Промежуточная аттестация
Соответствие освоения компетенций планируемым результатам обучения и критериям их оценивания (оценка: **удовлетворительно /зачтено**)				
на пороговом уровне:				
1	**ОПК-2.1** Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС	**Знает:** основные этические принципы ИИ (справедливость, прозрачность, подотчетность) и типовые этические риски в ИС (предвзятость алгоритмов, нарушение приватности). **Умеет:** выявлять явные этические проблемы в поставленной задаче (например, использование защищенных признаков в модели) и формулировать базовые требования к ИС на основе предоставленного шаблона этического чек-листа. **Владеет:** навыками систематизации очевидных этических требований к ИС по заданному алгоритму.	Практические работы, проект	Защита проекта, вопросы 26-31

2	**ML-1.1** Позиционирует задачу в контексте трендов ИИ	**Знать:** основные тренды развития ИИ **Уметь:** находить информацию о регуляторных требованиях **Владеть:** базовой терминологией в области этики ИИ	Практические работы 1,5	Вопросы 1,2,15-19
3	**SS-1.1** Определяет предвзятости в данных и алгоритмах	**Знать:** основные типы алгоритмических предвзятостей **Уметь:** выявлять явные случаи bias в датасетах **Владеть:** методами первичного анализа данных	Практические работы 2,3	Вопросы 3-8
4	**SS-1.2** Применяет методики работы с этическими рисками	**Знать:** основные этапы жизненного цикла ИИ **Уметь:** применять готовые чек-листы для оценки рисков **Владеть:** шаблонами для этического аудита	Практические работы 4,7	Вопросы 9-14,24,25
Соответствие освоения компетенций планируемым результатам обучения и критериям их оценивания (оценка: **хорошо /зачтено**)				
<u>на базовом уровне</u>:				
1	**ОПК-2.1** Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС	**Знает:** методы анализа стейкхолдеров и их интересов применительно к этическим аспектам ИС, а также frameworks ответственного ИИ (OECD AI Principles, Microsoft RAI). **Умеет:** применять системный подход для выявления скрытых этических рисков и ценностных предпосылок в постановке задачи; формулировать комплексные этические требования к ИС, учитывающие жизненный цикл системы и интересы различных групп стейкхолдеров. **Владеет:** навыками проведения декомпозиции предметной области для выявления точек возникновения этических рисков и методиками разработки этических требований к ИС.	Практические работы, проект	Защита проекта, вопросы 26-31
2	**ML-1.1** Позиционирует задачу в контексте трендов ИИ	**Знать:** особенности национального и международного регулирования ИИ **Уметь:** анализировать соответствие проекта регуляторным требованиям **Владеть:** методами исследования нормативной базы	Практические работы 1,5	Вопросы 1,2,15-19
3	**SS-1.1** Определяет предвзятости в данных и алгоритмах	**Знать:** математические метрики справедливости алгоритмов **Уметь:** проводить комплексный аудит датасетов и моделей	Практические работы 2,3	Вопросы 3-8

		Владеть: инструментами анализа fairness		
4	**SS-1.2** Применяет методики работы с этическими рисками	**Знать:** frameworks ответственного ИИ (OECD, EU AI Act) **Уметь:** разрабатывать этические гайдлайны для проектов **Владеть:** методами интеграции этики в ML-пайплайны	Практические работы 4,7	Вопросы 9-14,24,25
Соответствие освоения компетенций планируемым результатам обучения и критериям их оценивания (оценка: **отлично /зачтено**)				
<u>на продвинутом уровне:</u>				
1	**9.1**	**Знает:** передовые методики анализа долгосрочных социальных последствий внедрения ИС, методы управления конфликтами требований (например, безопасность vs. приватность). **Умеет:** прогнозировать системные этические последствия и возникающие дилеммы на ранних этапах проектирования ИС; разрабатывать и аргументировать стратегию балансировки противоречивых этических требований; интегрировать этический анализ в общий процесс проектирования ИС. **Владеет:** навыками фасилитации междисциплинарного обсуждения требований к ИС и разработки организационных механизмов для управления этическими аспектами на протяжении всего жизненного цикла системы.	Практические работы, проект	Защита проекта, вопросы 26-31
2	**ML-1.1** Позиционирует задачу в контексте трендов ИИ	**Знать:** перспективные направления развития этики ИИ **Уметь:** прогнозировать этические вызовы новых технологий **Владеть:** методологией анализа долгосрочных последствий	Практические работы 1,5	Вопросы 1,2,15-19
3	**SS-1.1** Определяет предвзятости в данных и алгоритмах	**Знать:** передовые методы обнаружения и устранения bias **Уметь:** разрабатывать custom-метрики для специфических кейсов **Владеть:** экспертизой в области algorithmic fairness	Практические работы 2,3	Вопросы 3-8
4	**SS-1.2** Применяет методики работы с этическими рисками	**Знать:** методы управления рисками throughout жизненного цикла ИИ **Уметь:** создавать системы мониторинга этических аспектов **Владеть:** навыками построения	Практические работы 4,7	Вопросы 9-14,24,25

		культуры ответственного ИИ в организациях		

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

Типовые контрольные задания

Практическая работа 1: "Анализ кейса этического сбоя ИИ"
Задание: Проанализируйте кейс Cambridge Analytica или Amazon HR AI. Выявите основные этические нарушения, предложите меры по их предотвращению. Подготовьте краткий отчет (1-2 стр.) с рекомендациями.

Практическая работа 2: "Этический аудит датасета"
Задание: Проведите анализ датасета Adult Census Income на предмет репрезентативности и потенциальных предвзятостей. Используйте библиотеки Pandas и Fairlearn для расчета базовых метрик справедливости.

Практическая работа 3: "Сравнение моделей с коррекцией bias"
Задание: Обучите две модели классификации - базовую и с применением методов обеспечения справедливости. Сравните их метрики accuracy и fairness.

Дискуссия (Практическая работа 6)
Формат: Структурированные дебаты или обсуждение в малых группах
Цель: Развитие навыков аргументации, критического мышления и этического анализа
Пример темы: "Следует ли полностью запретить использование распознавания лиц в публичных пространствах?"
Как проводится:
1. **Подготовка (15 минут):**
Группа делится на 2 команды ("за" и "против")
Участники изучают материалы и готовят аргументы
2. **Проведение (30 минут):**
Презентация позиций (по 5 минут на команду)
Открытая дискуссия и опровержения
Выработка консенсусного решения
3. **Рефлексия (15 минут):**
Анализ качества аргументов
Оценка изменения позиций участников
Формулирование выводов
Критерии оценки дискуссии:
- Качество подготовки и глубина аргументов
- Логичность и структурированность высказываний
- Умение слушать и учитывать контраргументы
- Способность к конструктивному диалогу и компромиссу

Протокол заседания (Практическая работа 5)
Что это?
Формат: Ролевая игра - заседание этического комитета
Цель: Отработка процедуры этического ревью реального проекта

Контекст: Оценка гипотетического ИИ-стартапа на соответствие этическим стандартам

Структура протокола:

ПРОТОКОЛ №1
заседания Этического комитета по проекту "FaceCheck HR"
(система анализа эмоций кандидатов при приеме на работу)

Присутствовали:
Председатель: Иванов А.С.
Члены комитета: Петрова И.К., Сидоров В.М.
Разработчик проекта: Джонсон Л.
Эксперт по защите данных: Козлова Е.П.

Повестка дня:
1. Оценка соответствия проекта принципам EU AI Act
2. Анализ рисков дискриминации и нарушения приватности
3. Выработка рекомендаций для доработки проекта

Ход заседания:
1. **Выступление разработчика** (10 мин) - представление функционала системы
2. **Вопросы членов комитета** (15 мин) - уточнение технических и этических аспектов
3. **Обсуждение рисков** (20 мин) - идентификация потенциальных проблем
4. **Формулирование рекомендаций** (15 мин)

Выявленные риски:
- Возможность дискриминации по культурным особенностям выражения эмоций
- Отсутствие механизма информированного согласия
- Непрозрачность алгоритма принятия решений

Рекомендации:
1. Внедрить мультикультурную калибровку алгоритма
2. Разработать процедуру явного информированного согласия
3. Добавить функцию объяснения рекомендаций системы

Решение:
Проект рекомендован к доработке с учетом высказанных замечаний.
Срок представления исправленной версии - 30 дней.

Подписи: _____

Чем это полезно для студентов:
- Отработка реального рабочего процесса этического комитета
- Развитие навыков документального оформления решений
- Понимание процедуры compliance-проверок
- Умение формулировать конструктивные рекомендации

Критерии оценки протокола:
- Полнота отражения дискуссии
- Корректность идентификации рисков
- Практическая реализуемость рекомендаций
- Соблюдение формата деловой документации

Эти форматы обеспечивают практическое освоение компетенций **ОПК-2.1** (профессиональная коммуникация) и **SS-1.2** (управление этическими рисками) в условиях, максимально приближенных к реальной профессиональной деятельности.

Соответствие практических работ и индикаторов компетенций

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

Индикаторы компетенции	Соответствующие лабораторные работы	Обоснование
ОПК-2.1 Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС	**ПР 1:** Анализ кейса этического сбоя ИИ **ПР 5:** Симуляция этического ревью (ролевая игра) **ПР 6:** Дебаты по этическим дилеммам ИИ **ПР 7:** Разработка этического чек-листа	В рамках дисциплины системный подход применяется для анализа этических аспектов ИИ-систем: выявления стейкхолдеров и их интересов, определения этических требований на всех этапах жизненного цикла ИС, анализа компромиссов между различными этическими принципами и интеграции этических considerations в процесс проектирования ответственных ИИ-решений.
ML-1.1 Позиционирует собственную задачу в заданной области знания с точки зрения трендов современного искусственного интеллекта	**ПР 1:** Анализ кейса этического сбоя ИИ **ПР 4:** Анализ объяснимости модели (XAI) **ПР 5:** Симуляция этического ревью	Работы требуют анализа современных трендов регулирования ИИ (EU AI Act), применения методов объяснимого ИИ (XAI) и оценки соответствия актуальным этическим стандартам
SS-1.1 Определяет ценностные предпосылки, когнитивные искажения, культурно-обусловленные предвзятости в данных, алгоритмах, постановке задач для ИИ	**ПР 2:** Этический аудит датасета **ПР 3:** Сравнение моделей с коррекцией bias **ПР 4:** Анализ объяснимости модели (XAI)	Работы напрямую нацелены на выявление и анализ предвзятостей в данных и алгоритмах, использование инструментов Fairness и XAI для диагностики дискриминации
SS-1.2 Применяет методики работы с этическими и социальными рисками, возникающими на разных стадиях жизненного цикла ИИ	**ПР 5:** Симуляция этического ревью **ПР 6:** Дебаты по этическим дилеммам **ПР 7:** Разработка этического чек-листа для проекта	Работы требуют применения frameworks ответственного ИИ (OECD, EU AI Act), разработки превентивных мер и методик управления рисками throughout жизненного цикла ИИ-систем

Тест для текущего контроля по темам 1-2
(Введение в этику ИИ, Предвзятость и справедливость)
Время выполнения: 30 минут
Формат: 20 вопросов (закрытые, на соответствие, одиночный/множественный выбор)
Проходной балл: 15/20 правильных ответов

Примеры тестовых заданий:

1. Основным предметом изучения этики ИИ является:
a) Максимизация прибыли компаний-разработчиков
b) Техническая оптимизация алгоритмов
c) **Социальные последствия и моральные аспекты применения ИИ**
d) Скорость обработки данных

2. Алгоритмическая предвзятость (bias) может возникать на этапе:
a) Сбора данных
b) Проектирования алгоритма
c) Интерпретации результатов
d) **Всех перечисленных этапов**

3. Метрика Demographic Parity требует, чтобы:
a) Точность прогноза была одинаковой для всех групп
b) **Доля положительных решений была одинаковой для всех групп**
c) Ложные срабатывания распределялись равномерно
d) Все группы были представлены в данных поровну

4. Кейс с системой COMPAS (2016) продемонстрировал проблему:
a) Низкой производительности алгоритмов
b) **Расовой дискриминации в оценке рецидива**
c) Нарушения конфиденциальности данных
d) Неправильной визуализации результатов

5. Основная цель EU AI Act (2021):
a) Запретить все системы искусственного интеллекта
b) **Установить категории риска и требования для ИИ-систем**
c) Унифицировать языки программирования для ИИ
d) Снизить стоимость разработки ИИ
(Полная версия теста включает 20 вопросов с вариантами ответов)

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

Методические рекомендации, определяющие процедуры оценивания ответа на зачете:

Шкала оценивания зачета:

«Зачтено»:

- Полные, структурированные ответы на основные вопросы (1-25)
- Глубокий анализ кейсов и практических ситуаций (26-31)
- Умение связать теорию с практическими примерами
- Знание современной нормативной базы и инструментов
- Способность аргументировать этическую позицию

«Не зачтено»:

- Отсутствие понимания ключевых концепций этики ИИ
- Неспособность проанализировать простейшие кейсы
- Грубые ошибки в применении терминологии
- Незнание базовых регуляторных требований

Методические рекомендации по оцениванию практических работ

Критерии оценки практических работ:

Компонент оценки	Пороговый уровень (3/5)	Базовый уровень (4/5)	Продвинутый уровень (5/5)
Теоретическая подготовка (20%)	Базовое понимание концепций	Глубокое понимание с ссылками на источники	Критический анализ, сравнение подходов
Практическая реализация (40%)	Работоспособное решение	Оптимизированный код с комментариями	Инновационное решение с обработкой edge-cases
Анализ результатов (30%)	Простые выводы по результатам	Сравнительный анализ с интерпретацией	Комплексный анализ с рекомендациями
Оформление (10%)	Базовая структура отчета	Профессиональное оформление	

Методические рекомендации, определяющие процедуры оценивания тестов

Критерии оценивания и шкала оценок

Система подсчета баллов:

- Каждый правильный ответ = 1 балл
- Максимальный балл = 20
- **Неправильные ответы и пропуски = 0 баллов**

Шкала перевода в оценки:

Количество правильных ответов	Оценка	Уровень освоения
18-20	**Отлично**	Продвинутый уровень
15-17	**Хорошо**	Базовый уровень
11-14	**Удовлетворительно**	Пороговый уровень
0-10	**Неудовлетворительно**	Недостаточный уровень

3. Анализ результатов и обратная связь

Статистический анализ теста:

- Расчет среднего балла по группе
- Определение вопросов с наименьшей успеваемостью
- Анализ распределения результатов

Интерпретация результатов:

- **18-20 баллов:** Глубокое понимание концепций, способность применять знания в практических ситуациях
- **15-17 баллов:** Уверенное знание основных принципов, понимание ключевых проблем этики ИИ

- **11-14 баллов:** Минимальное понимание необходимых концепций, требуется дополнительная работа
- **0-10 баллов:** Критический пробел в знаниях, необходимо повторение материала

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

1. Общие сведения

Условия применения практических работ:

1) Программное обеспечение:

- Python 3.8+ с библиотеками: pandas, numpy, matplotlib, seaborn
- Специализированные библиотеки этического ИИ: fairlearn, aif360, shap, lime
- Jupyter Notebook / Google Colab для интерактивной разработки
- Средства визуализации: plotly, altair для интерактивных дашбордов

2) Аппаратное обеспечение:

- Стандартные вычислительные мощности (CPU) достаточны для большинства задач
- Доступ к интернету для работы с облачными средами и датасетами
- Рекомендуемый объем ОЗУ: 8+ ГБ для работы с большими датасетами
- **3) Облачная инфраструктура:**
- Google Colab Pro для доступа к GPU при необходимости
- Kaggle Notebooks для работы с тестовыми датасетами
- GitHub для версионного контроля и совместной работы

2. Цели, задачи и ожидаемые результаты

2.1. Обоснование необходимости практических работ

Практические работы в дисциплине необходимы поскольку:

a) Прикладной характер компетенций:

- Этические принципы требуют практического применения в реальных сценариях
- Навыки выявления bias не могут быть сформированы только теоретически
- Умение проводить этический аудит требует работы с реальными данными и кейсами

b) Формирование профессионального мышления:

- Развитие критического мышления для анализа этических дилемм
- Отработка навыков принятия решений в условиях неопределенности
- Формирование системного подхода к управлению рисками

c) Подготовка к реальным профессиональным задачам:

- Опыт работы с инструментами этического аудита (Fairlearn, AIF360)
- Навыки коммуникации этических аспектов с техническими и нетехническими стейкхолдерами
- Понимание процедур compliance в соответствии с регуляторными требованиями

2.2. Задачи практических работ

Технические задачи:

- Освоение инструментов анализа справедливости алгоритмов
- Применение методов выявления и минимизации предвзятости
- Разработка систем мониторинга этических аспектов ИИ-систем

Аналитические задачи:

- Проведение комплексного этического аудита ИИ-проектов
- Анализ нормативных требований и стандартов
- Оценка социальных последствий внедрения ИИ-решений

Коммуникативные задачи:

- Подготовка отчетов и презентаций по этическим аспектам

- Участие в дискуссиях и этических ревью
- Разработка документации для различных стейкхолдеров

2.3. Ожидаемые результаты

После выполнения цикла практических работ студенты смогут:

На техническом уровне:

Проводить аудит датасетов на наличие предвзятостей

Применять метрики fairness для оценки алгоритмов

Использовать методы объяснимого ИИ (XAI) для интерпретации решений

На аналитическом уровне:

Выявлять и классифицировать этические риски ИИ-проектов

Оценивать соответствие нормативным требованиям

Разрабатывать стратегии управления этическими рисками

На коммуникативном уровне:

- Эффективно представлять результаты этического аудита
- Участвовать в междисциплинарных обсуждениях этических аспектов
- Разрабатывать этические гайдлайны и политики

2.4. Необходимые ресурсы для реализации

Методические материалы:

Пошаговые инструкции для каждой практической работы

Примеры кода и шаблоны отчетов

Банк кейсов для анализа этических дилемм

Техническая поддержка:

Настроенные среды разработки с необходимыми библиотеками

Доступ к тестовым датасетам и вычислительным ресурсам

Консультационная поддержка по техническим вопросам

Организационное обеспечение:

Четкое расписание и дедлайны для каждой работы

Система оценки с прозрачными критериями

Механизм обратной связи и доработки

3. Порядок реализации практических работ

3.1. Организационная структура проведения работ

Этап 1: Подготовительный (1 неделя)

Знакомство с теоретической базой

Установка и настройка программного обеспечения

Формирование рабочих групп (для групповых проектов)

Этап 2: Выполнение (2 недели)

Поэтапное выполнение заданий практической работы

Промежуточные консультации с преподавателем

Самостоятельная работа и групповые обсуждения

Этап 3: Защита и оценка (1 неделя)

Презентация результатов

Обсуждение и ответы на вопросы

Сдача итогового отчета

3.2. Детализированное описание практических работ

Практическая работа 1: "Анализ кейса этического сбоя ИИ"

Цель: Развитие навыков анализа реальных этических инцидентов и формирования рекомендаций

Пошаговая инструкция:

- **Шаг 1: Выбор и анализ кейса (2 часа)**
- Выберите один из предложенных кейсов (COMPAS, Amazon HR AI, Cambridge Analytica)
- Проведите анализ по схеме:

Контекст и описание системы
Выявленные этические нарушения
Причины и последствия инцидента
Действия заинтересованных сторон

Шаг 2: Разработка рекомендаций (2 часа)

- Предложите меры по предотвращению подобных инцидентов
- Разработайте план исправления последствий
- Сформулируйте превентивные меры для будущих проектов

Шаг 3: Оформление отчета (1 час)

- Подготовьте отчет в формате аналитической записки (1-2 страницы)
- Включите разделы: резюме, анализ, рекомендации, выводы

Инструменты и материалы:

- Банк кейсов с описанием реальных инцидентов
- Шаблон аналитической записки
- Критерии оценки анализа

Практическая работа 2: "Этический аудит датасета"

Цель: Освоение методов выявления предвзятостей в тренировочных данных

Пошаговая инструкция:

Шаг 1: Подготовка данных (1 час)

```
python
# Загрузка и первичный анализ датасета
import pandas as pd
import fairlearn
from fairlearn.metrics import demographic_parity_difference

data = pd.read_csv('adult_census.csv')
print(f"Размер датасета: {data.shape}")
print(f"Колонки: {data.columns.tolist()}")
```

Шаг 2: Анализ репрезентативности (2 часа)

- Анализ распределения protected attributes (пол, возраст, раса)
- Проверка баланса классов в целевой переменной
- Визуализация распределений по группам

Шаг 3: Расчет метрик справедливости (2 часа)

```
python
# Расчет базовых метрик fairness
from fairlearn.metrics import (
    demographic_parity_ratio,
    equalized_odds_difference
)

# Пример расчета Demographic Parity
dp_diff = demographic_parity_difference(
    y_true, y_pred, sensitive_features=gender
)
```

Индивидуальные задания:

1. Рассчитайте не менее 3 метрик fairness для разных protected attributes
2. Визуализируйте распределения с помощью matplotlib/seaborn
3. Сформулируйте выводы о наличии предвзятостей

Критерии проверки:

- Корректность расчета метрик
- Качество визуализаций
- Глубина анализа результатов

3.3. Методика проведения групповых работ

Для работ 5-7 применяется следующий подход:

Формирование групп:

Размер группы: 3-4 человека
Распределение ролей: модератор, аналитик, документалист, презентатор
Четкое определение зон ответственности

Процесс работы:

Регулярные групповые обсуждения (оффлайн/онлайн)
Ведение протоколов встреч и принятия решений
Система peer-review внутри группы

Оценка групповой работы:

Индивидуальный вклад каждого участника
Качество групповой динамики и коммуникации
Итоговый результат группы

4. Система контроля и оценки

4.1. Критерии оценки практических работ

Общие критерии для всех работ:

Компонент	**Пороговый уровень (3/5)**	**Базовый уровень (4/5)**	**Продвинутый уровень (5/5)**
Теоретическая подготовка (20%)	Базовое понимание концепций	Глубокое понимание с ссылками на источники	Критический анализ, сравнение подходов
Практическая реализация (40%)	Работоспособное решение	Оптимизированное решение с комментариями	Инновационное решение с обработкой edge-cases
Анализ результатов (30%)	Простые выводы по результатам	Сравнительный анализ с интерпретацией	Комплексный анализ с рекомендациями
Оформление (10%)	Базовая структура отчета	Профессиональное оформление с графиками	Публикация на GitHub, интерактивные дашборды

4.2. Шкала оценивания

Количественная шкала:

0-49 баллов: Неудовлетворительно
50-69 баллов: Удовлетворительно
70-84 баллов: Хорошо
85-100 баллов: Отлично

Качественные дескрипторы:

Отлично: Работа демонстрирует глубокое понимание этических аспектов, творческий подход и способность к самостоятельному анализу

Хорошо: Работа показывает уверенное владение материалом и умение применять изученные методы

Удовлетворительно: Работа соответствует минимальным требованиям, но содержит существенные недочеты

Неудовлетворительно: Работа не демонстрирует понимания ключевых концепций или содержит критические ошибки

4.3. Процедура защиты работ

Подготовка к защите:
Подготовка презентации (5-7 слайдов)
Репетиция выступления (3-5 минут)
Подготовка к ответам на вопросы
Процесс защиты:
Краткая презентация основных результатов
Ответы на вопросы преподавателя и студентов
Обсуждение практической значимости работы
Критерии оценки защиты:
Качество презентации и коммуникации
Глубина понимания темы
Умение аргументировать и защищать свою позицию

5. Адаптация для студентов с овз

5.1. Для студентов с нарушениями зрения:
Альтернативные форматы материалов (аудиоописания, тактильные графики)
Увеличенное время выполнения работ (дополнительные 50%)
Использование screen readers и голосового ввода

5.2. Для студентов с нарушениями слуха:
Визуальные дублирования устных инструкций
Письменные материалы всех обсуждений
Использование систем визуального оповещения о времени

5.3. Для студентов с нарушениями опорно-двигательного аппарата:
- Адаптация интерфейсов программного обеспечения
- Альтернативные способы ввода данных
- Увеличенное время на выполнение задач

5.4. Общие адаптации:
- Индивидуальный график консультаций
- Гибкие дедлайны с учетом особенностей здоровья
- Альтернативные форматы сдачи работ

6. Методические рекомендации для преподавателей

6.1. Организация учебного процесса:
- Четкое планирование времени для каждой практической работы
- Регулярный мониторинг прогресса студентов
- Своевременное предоставление обратной связи

6.2. Мотивация студентов:
- Связь практических работ с реальными профессиональными задачами
- Демонстрация практической значимости получаемых навыков
- Создание атмосферы открытого обсуждения и взаимного обучения

6.3. Обеспечение качества:
- Постоянное обновление материалов в соответствии с развитием области
- Регулярный сбор обратной связи от студентов
- Анализ результатов и корректировка методики преподавания

Данные методические указания обеспечивают системный подход к организации практических работ и способствуют эффективному формированию заявленных компетенций у студентов.

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

```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:

SS-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

```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

ML-1.1: *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

ML-1.1: *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: Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС

SS-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, ML-1.1
Техническая реализация	25%	- Корректность кода - Качество визуализаций - Эффективность методов	SS-1.1, ML-1.1
Качество рекомендаций	25%	- Практическая реализуемость - Соответствие best practices - Учет регуляторных требований	SS-1.2, ML-1.1
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. **ML-1.1** демонстрируется через применение современных методов и учет трендов
4. **ОПК-2.1** Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС

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 этнических 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. "$\nabla^ 2$ 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, Кадурин, Артур и Архангельская Екатерина. Глубокое обучение. Издательский дом" Питер", 2017..

5.2. Периодические издания и конференции (A*):

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.prlib.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
ML-1.1: Применение современных методов обеспечения справедливости
ОПК-2.1: Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС КЕЙС 2: РАЗРАБОТКА ЭТИЧЕСКИХ СЦЕНАРИЕВ ДЛЯ ГОЛОСОВОГО АССИСТЕНТА «САЛЮТ»

Цель:

Обеспечить этичное 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 интентов
   - Настройка confidence thresholds для sensitive категорий
2. **Разработка этичных 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**
   - A/B тестирование сфокурированных 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

ML-1.1: 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

ML-1.1: 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: Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС
SS-1.2: Создание системы конструктивной обратной связи
ML-1.1: 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: Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС

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
BIM-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

```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: Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС

КЕЙС 2: СПРАВЕДЛИВАЯ СИСТЕМА УПРАВЛЕНИЯ АРЕНДНЫМИ СТАВКАМИ

Цель:

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

Технологии:

Machine Learning для прогнозирования рыночных цен

Fairness constraints в алгоритмах ценообразования

A/B testing платформа

Pealtime 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

```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:

Сничение 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
ML-1.1: 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

```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: Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС
ML-1.1: 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: Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС

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
ML-1.1: Применение современных методов оптимизации с учетом устойчивого развития
ОПК-2.1: Способен применять системный подход к анализу предметной (проблемной) области, выявлению требований к ИС

7. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю)

7.1 Перечень информационно-коммуникационных технологий

1. Электронная почта mail.ru, yandex.ru
2. Yandex Browser
3. Система управления обучением Moodle – сдача работ

7.2 Перечень лицензионного и свободно распространяемого программного обеспечения

1. OpenOffice
2. GIT
3. Yandex Browser
4. Mozilla Firefox
5. Google Chrome
6. Python + Jupyter + Google Colab
7. SymPy/SageMath
8. Octave (аналог MATLAB)

8. 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.