Приложение к ОПОП «Безопасность и этика искусственного интеллекта»

СПбГЭТУ «ЛЭТИ»
ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ

МИНОБРНАУКИ РОССИИ

федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И.Ульянова (Ленина)»
(СПбГЭТУ «ЛЭТИ»)

РАБОЧАЯ ПРОГРАММА

дисциплины

«ДОВЕРЕННЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ»

для подготовки магистров

по направлению

09.04.01 «Информатика и вычислительная техника»

по программе

«Безопасность и этика искусственного интеллекта»

Санкт-Петербург

2021

ЛИСТ СОГЛАСОВАНИЯ

Разработчики:

доцент, к.т.н. Сулавко А.Е.

профессор, д.т.н. Ложников П.С.

Рабочая программа рассмотрена и одобрена на заседании кафедры ВТ 02.09.2021, протокол № 6

Рабочая программа рассмотрена и одобрена учебно-методической комиссией ФКТИ, 09.09.2021, протокол № 6

Согласовано в ИС ИОТ

Начальник ОМОЛА Загороднюк О.В.

1 СТРУКТУРА ДИСЦИПЛИНЫ

Обеспечивающий факультет	ФКТИ
Обеспечивающая кафедра	ИС
Общая трудоемкость (ЗЕТ)	2
Курс	1
Семестр	2
Виды занятий	
Лекции (академ. часов)	17
Практические занятия (академ. часов)	17
Все контактные часы (академ. часов)	34
Самостоятельная работа, включая часы на контроль (академ. часов)	38
Всего (академ. часов)	72
Вид промежуточной аттестации	
Зачет (курс)	1

2 АННОТАЦИЯ ДИСЦИПЛИНЫ

«ДОВЕРЕННЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ»

Содержание дисциплины включает в себя изучение проблем обеспечения доверия к искусственному интеллекту (ИИ) и подходов к их решению, а также свойств, качеств доверенного искусственного интеллекта и понятий, непосредственно связанных с доверием к ИИ, таких как управление рисками ИИ, робастность ИИ, объяснимость ИИ, функциональная безопасность ИИ. Практические занятия ориентированы на проведение аналитической работы, развитие способностей к проведению аналитико-синтетических исследований научных публикаций и технических статей на примере тематики обеспечения доверия к ИИ.

SUBJECT SUMMARY

«TRUSTWORTHINESS IN ARTIFICIAL INTELLIGENCE»

The content of the discipline includes the study of the problems of ensuring trust in artificial intelligence (AI) and approaches to their solution, as well as the properties, qualities of trusted artificial intelligence and concepts directly related to trust in AI, such as AI risk management, AI robustness, explainability of AI. , functional safety of AI. Practical exercises are focused on conducting analytical work, developing the ability to conduct analytical and synthetic research of scientific publications and technical articles on the example of the topic of ensuring confidence in AI.

3 ОБЩИЕ ПОЛОЖЕНИЯ

3.1 Цели и задачи дисциплины

1. Цель дисциплины:

-изучение проблем обеспечения доверия к искусственному интеллекту (ИИ) и подходов к их решению;

-формирование умений и навыков составления требований к системам доверенного искусственного интеллекта на основе анализа профессиональной информации из данной области;

-приобретение знаний, необходимых для исследования и разработок архитектур систем доверенного искусственного интеллекта для различных предметных областей на основе комплексов методов и инструментальных средств систем искусственного интеллекта.

2. Задачи дисциплины:

-изучение проблемы обеспечения доверия к искусственному интеллекту (ИИ) и подходов к их решению;

-выработка практических навыков проведения аналитической работы у студентов, развитие способности к проведению аналитико-синтетических исследований научных публикаций и технических статей на примере тематики обеспечения доверия к ИИ;

-получение представления об основных свойствах, качествах доверенного искусственного интеллекта и понятиях, непосредственно связанных с доверием к ИИ, таких как управление рисками ИИ, робастность ИИ, объяснимость ИИ, функциональная безопасность ИИ.

3. Знание основных источников патентной и научно-технической информации, площадки для размещения научных работ и докладов конференций в области доверенного искусственного интеллекта; основных концепций искусственного

интеллекта, связанные с робастностью, оценкой и управлением рисками, функциональной безопасностью, объяснимостью, обнаружением аномалий и компьютерных атак, защитой данных; основных международных и национальных стандартов в области доверенного искусственного интеллекта.

4. Умение готовить научные доклады и презентации для конференций и семинаров по тематике доверенного искусственного интеллекта; формулировать, анализировать, декомпозировать задачи, связанные с обеспечением доверия к искусственному интеллекту с использованием применяемых в этой области принципов, методов и подходов; интерпретировать требования заказчика и применять концепции искусственного интеллекта при формулировании технического задания.

5. Владение навыками проведения аналитико-синтетического исследования патентов и научно-технических публикаций в области обеспечения доверия к искусственному интеллекту; профессиональной информацией в области патентов, стандартов и научно-технических достижений в области доверенного искусственного интеллекта при составлении требований к разработке таких систем, а также стандартов и технических спецификаций; базовыми навыками составления технических отчетов и обзоров в области доверенного искусственного интеллекта с использованием концепций робастности, управления рисками, функциональной безопасности, объяснимости, обнаружения аномалий. Исследовать и разрабатывать архитектуры систем искусственного интеллекта для различных предметных областей.

3.2 Место дисциплины в структуре ОПОП

Дисциплина изучается на основе ранее освоенных дисциплин учебного плана:

1. «Машинное обучение»

2. «Введение в нейронные сети»

и обеспечивает изучение последующих дисциплин:

1. «Защищенное исполнение искусственного интеллекта»

2. «Этика и правовые проблемы искусственного интеллекта»

3.3 Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

В результате освоения образовательной программы обучающийся должен достичь следующие результаты обучения по дисциплине:

Код компетенции/ индикатора компетенции	Наименование компетенции/индикатора компетенции
ПК-23	Способен исследовать и разрабатывать архитектуры систем искусственного интеллекта для различных предметных областей на основе комплексов методов и инструментальных средств систем искусственного интеллекта
ПК-23.1	*Исследует и разрабатывает архитектуры систем искусственного интеллекта для различных предметных областей*

4 СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1 Содержание разделов дисциплины

4.1.1 Наименование тем и часы на все виды нагрузки

№ п/п	Наименование темы дисциплины	Лек, ач	Пр, ач	СР, ач
1	Введение	1		
2	Проблемы доверия искусственному интеллекту	2	2	4
3	Управление рисками искусственного интеллекта	2	2	4
4	Объяснимость моделей искусственного интеллекта и машинного обучения	2	2	4
5	Робастность искусственного интеллекта и нейронных сетей	2	2	4
6	Функциональная безопасность искусственного интеллекта	2	2	4
7	Компьютерные атаки на искусственный интеллект	2	2	4
8	Защита данных	2	3	6
9	Обнаружение аномалий	1	2	4
10	Заключение	1		4
	Итого, ач	17	17	38
	Из них ач на контроль	0	0	0
	Общая трудоемкость освоения, ач/зе	72/2		

4.1.2 Содержание

№ п/п	Наименование темы дисциплины	Содержание
1	Введение	Предмет дисциплины и ее задачи. Содержание и форма проведения занятий.
2	Проблемы доверия искусственному интеллекту	Уровни доверия. Компоненты и свойства доверенного ИИ. Аппаратные платформы для доверенного ИИ. Неисправности оборудования. Понятие рисков ИИ. Заинтересованные стороны, их активы и ценности. Ответственность, подотчетность и управляемость ИИ. Безопасность на объектах критической информационной инфраструктуры. Уязвимости ИИ и новые угрозы безопасности. Дрейф модели и предвзятость ИИ. Непредсказуемость ИИ. Непрозрачность ИИ. Проблемы, связанные со спецификацией, внедрением и использованием систем ИИ. Меры по смягчению последствий.Жизненный цикл ИИ. Тестирование ИИ. Развертывание модели. Решения MLOps, ModelOps

№ п/п	Наименование темы дисциплины	Содержание
3	Управление рисками искусственного интеллекта	Принципы управления рисками. Обязательства по управлению рисками. Распределение ролей, полномочий и ответственности по управлению рисками. Процесс управления рисками. Критерии рисков ИИ. Оценка рисков ИИ. Идентификация рисков ИИ. Анализ рисков ИИ. Классификация рисков ИИ.
4	Объяснимость моделей искусственного интеллекта и машинного обучения	Прозрачность и объяснимость ИИ. Цели обеспечения объяснимости ИИ. Объяснимость данных. Оценка вклада признаков в результат анализа. Объяснимость причинно-следственных связей решений ИИ. Определение соответствия моделей или данных этическим нормам. Методы и подходы к реализации объяснимого ИИ и метрики объяснимости (деревья решений, GA2M, TCAV, LIME, SHAPи другое). Объяснимость данных. Объяснимость на разных этапах жизненного цикла ИИ. Программные продукты для создания объяснимого ИИ или повышения объяснимости ИИ.
5	Робастность искусственного интеллекта и нейронных сетей	Концепция робастности. Типичный алгоритм для оценки робастности. Метрики робастности. Методы оценки робастности (статистические, формальные, эмпирические). Оценка робастности искусственных нейронных сетей. Устойчивость обучения. Повышение устойчивости обучения.
6	Функциональная безопасность искусственного интеллекта	Понятие функциональной безопасности ИИ. Управление безопасностью ИИ. Свойства и связанные с ними факторы риска безопасности технологий ИИ. Трёхкомпонентная структура системы ИИ. Технологические элементы для создания и исполнения модели ИИ. Уровень автоматизации и контроля ИИ. Дрейф данных. Дрейф концепции. Безопасность ИИ. Проблемы с системным оборудованием. Меры контроля и смягчения последствий.
7	Компьютерные атаки на искусственный интеллект	Классификация компьютерных атак на ИИ. Извлечение данных (входных, выходных). Извлечение параметров модели для понимания алгоритма ее работы. Извлечение и интерпретация знаний ИИ. Манипуляции с моделями. Манипуляции входными и обучающими данными. Состязательные атаки. Зондирование моделей.Отказ в обслуживании ИИ. Методы противодействия атакам на ИИ. Программные продукты для детектирования и отражения атак на ИИ.
8	Защита данных	Защита данных на этапе хранения, обучения и исполнения. Отравление данных. Дифференциальная конфиденциальность. Гомоморфное шифрование. Классическое шифрование. Федеративное обучение. Генерация синтетических наборов данных. Метрики приватности. Программные продукты для защиты конфиденциальности и целостности данных ИИ.

№ п/п	Наименование темы дисциплины	Содержание
9	Обнаружение аномалий	Определение дрейфа данных и других проблем. Утечка данных. Обнаружение заражения данных. Соблюдение нормативных требований к потреблению данных модели. Программные продукты для обнаружения аномалий в данных.
10	Заключение	Итоги курса. Перспективы развития отрасли доверенного ИИ.

4.2 Перечень лабораторных работ

Лабораторные работы не предусмотрены.

4.3 Перечень практических занятий

Наименование практических занятий	Количество ауд. часов
1. Проблемы доверия искусственному интеллекту	2
2. Управление рисками искусственного интеллекта	2
3. Объяснимость моделей искусственного интеллекта и машинного обучения	2
4. Робастность искусственного интеллекта и нейронных сетей	2
5. Функциональная безопасность искусственного интеллекта	2
6. Компьютерные атаки наискусственный интеллект	2
7. Защита данных. Генерация синтетических наборов данных	3
8. Обнаружение аномалий	2
Итого	17

4.4 Курсовое проектирование

Курсовая работа (проект) не предусмотрены.

4.5 Реферат

Исходные данные и требования: Обучающиеся готовят рефераты на темы, выданные преподавателем, работа выполняется индивидуально.

Каждый реферат представляет собой краткое аналитико-синтетическое исследование документального потока (научные статьи, интернет-источники, обзорные статьи и другие публикации).

Объем реферата 5-10 страниц. Реферат должен быть оформлен в виде статьи и содержать введение, основную часть с подзаголовками и выводы по исследуемой проблеме. Дополнительно студент может подготовить публикацию на конференцию или в журнал ВАК/РИНЦ/Scopus/WebofScience.

Критерии оценки рефератов:

1. Полнота представления материала (должны быть освещены основные аспекты исследуемой проблемы, количество источников не менее 15, из них не менее 30% – публикации в зарубежных научно-технических журналах).

2. Глубина проработки и ясность изложения материала (формулирование мыслей, содержательные выводы).

3. Структура изложения (соответствие структурных элементов заданию).

Примерные темы:

№ п/п	Название темы	Перевод темы
1	Принципы обеспечения доверия к искусственному интеллекту при разработке интеллектуальных систем в области бизнес-аналитики	
2	Управление рисками искусственного интеллекта на объектах критической информационной инфраструктуры	
3	Объяснимость моделей искусственного интеллекта, осуществляющих мониторинг оборудования при добычи нефти и газа с целью предсказания предаварийных ситуаций	
4	Методы обеспечения робастности обучения нейронных сетей	
5	Функциональная безопасность искусственного интеллекта в нефтегазовой отрасли	
6	Классификация компьютерных атак на искусственный интеллект в банковской сфере	
7	Защита персональных данных пользователей в банковской сфере, при их обработке методами машинного обучения	
8	Обнаружение аномалий в данных на этапе обучения искусственного интеллекта	

№ п/п	Название темы	Перевод темы
9	Информационная безопасность моделей машинного обучения: актуальные проблемы и перспективы их решения	
10	Автоматическое машинное обучение на малых выборках в системах информационной безопасности	

4.6 Индивидуальное домашнее задание

Индивидуальное домашнее задание не предусмотрено.

4.7 Доклад

Доклад не предусмотрен.

4.8 Кейс

Кейс не предусмотрен.

4.9 Организация и учебно-методическое обеспечение самостоятельной работы

Изучение дисциплины сопровождается самостоятельной работой студентов с рекомендованными преподавателем литературными источниками и информационными ресурсами сети Интернет. Планирование времени для изучения дисциплины осуществляется на весь период обучения, предусматривая при этом регулярное повторение пройденного материала. Обучающимся, в рамках внеаудиторной самостоятельной работы, необходимо регулярно дополнять сведениями из литературных источников материал, законспектированный на лекциях. При этом на основе изучения рекомендованной литературы целесообразно составить конспект основных положений, терминов и определений, необходимых для освоения разделов учебной дисциплины.

Особое место уделяется консультированию, как одной из форм обучения

и контроля самостоятельной работы. Консультирование предполагает особым образом организованное взаимодействие между преподавателем и студентами,при этом предполагается, что консультант либо знает готовое решение, которое он может предписать консультируемому, либо он владеет способами деятельности, которые указывают путь решения проблемы.

Текущая СРС	**Примерная трудоемкость, ач**
Работа с лекционным материалом, с учебной литературой	6
Опережающая самостоятельная работа (изучение нового материала до его изложения на занятиях)	0
Самостоятельное изучение разделов дисциплины	0
Выполнение домашних заданий, домашних контрольных работ	0
Подготовка к лабораторным работам, к практическим и семинарским занятиям	7
Подготовка к контрольным работам, коллоквиумам	0
Выполнение расчетно-графических работ	0
Выполнение курсового проекта или курсовой работы	0
Поиск, изучение и презентация информации по заданной проблеме, анализ научных публикаций по заданной теме	15
Работа над междисциплинарным проектом	0
Анализ данных по заданной теме, выполнение расчетов, составление схем и моделей, на основе собранных данных	0
Подготовка к зачету, дифференцированному зачету, экзамену	10
ИТОГО СРС	38

5 Учебно-методическое обеспечение дисциплины

5.1 Перечень основной и дополнительной литературы, необходимой для освоения дисциплины

№ п/п	Название, библиографическое описание	К-во экз. в библ.
Основная литература		
1	Николенко С. Глубокое обучение [Электронный ресурс] / С. Николенко, А. Кадурин, Е. Архангельская, 2019. -480 с.	неогр.
2	Загорулько, Юрий Алексеевич. Искусственный интеллект. Инженерия знаний [Электронный ресурс] : Учебное пособие для вузов / Загорулько Ю. А., Загорулько Г. Б., 2020. -93 с	неогр.
Дополнительная литература		
1	Андрей Бурков Машинное обучение без лишних слов [Электронный ресурс] / Бурков Андрей, 2020. -192 с.	неогр.
2	Бринк Х. Машинное обучение [Электронный ресурс] / Х. Бринк, Д. Ричардс, М. Феверолф, 2017. -336 с.	неогр.
3	Чио К. Машинное обучение и безопасность [Электронный ресурс] : руководство / К. Чио, Д. Фримэн, 2020. -388 с.	неогр.
4	Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных [Электронный ресурс], 2015. -400 с.	неогр.

5.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет», используемых при освоении дисциплины

№ п/п	Электронный адрес
1	Ахметов Б.С., Иванов А.И., Фунтиков В.А., Безяев А.В., Малыгина Е.А. Технология использования больших нейронных сетей для преобразования нечетких биометрических данных в код ключа доступа: Монография. / Алматы: ТОО «Издательство LEM», 2014 – 144 с. http://lib.tarsu.kz/rus/all.doc/Elektron_res/Axmetov_Texnologia%20neironnix%20setei.pdf
2	Этика и «цифра»: от проблем к решениям. Аналитический доклад – URL: https://ethics.cdto.center/2021/5-3-doverennyj-ii-v-regulirovanii-i-standartah
3	Иванов А.И., Сулавко А.Е. Использование сетей корреляционных нейронов с многоуровневым квантованием: защита от извлечения знаний из параметров решающего правила // Пенза – 2020 г. Издательство «ПГУ» – 48 с. Тираж 300 экз. ISBN 978-5-907364-02-8 – URL: https://tsib.pnzgu.ru/files/tsib.pnzgu.ru/ivanov_sulavko_preprint_2020_ispolzsetkorrelneyron.p

№ п/п	Электронный адрес
4	Иванов, А. И., Золотарева Т. А. Искусственный интеллект в защищенном исполнении: синтез статистико-нейросетевых автоматов многокритериальной проверки гипотезы независимости малых выборок биометрических данных : препринт. – Пенза : Изд-во ПГУ, 2020. – 104 с. – URL: https://tsib.pnzgu.ru/files/tsib.pnzgu.ru/ivanov_zolotareva_preprint_2020_iskusstvintellekt.pdf

5.3 Адрес сайта курса

Адрес сайта курса: https://vec.etu.ru/moodle/course/view.php?id=7824

6 Критерии оценивания и оценочные материалы

6.1 Критерии оценивания

Для дисциплины «Доверенный искусственный интеллект» формой промежуточной аттестации является зачет.

Зачет

Зачет по дисциплине выставляется по результатам текущего контроля, при условии, что студент в целом овладел курсом, усвоил некоторые разделы на уровне определений и формулировок, умеет решать задачи и применять полученные знания. Зачет по дисциплине не выставляется, если студент не освоил курс и испытывает трудности при ответе на ключевые вопросы дисциплины.

Особенности допуска

Допуском к зачету является подготовка, предоставление и защита реферата, успешное выполнение тестовых заданий на практических занятиях.

6.2 Оценочные материалы для проведения текущего контроля и промежуточной аттестации обучающихся по дисциплине

Весь комплект контрольно-измерительных материалов для проверки сформированности компетенции (индикатора компетенции) размещен в закрытой части по адресу, указанному в п. 5.3

6.3 График текущего контроля успеваемости

Неделя	Темы занятий	Вид контроля
1	Проблемы доверия искусственному интеллекту	
2		Тест
3	Управление рисками искусственного интеллекта	
4		Тест
5	Объяснимость моделей искусственного интеллекта и машинного обучения	
6		Тест
7	Робастность искусственного интеллекта и нейронных сетей	
8		Тест
9	Функциональная безопасность искусственного интеллекта	
10		Тест
11	Компьютерные атаки на искусственный интеллект	
12		Тест
13	Защита данных Обнаружение аномалий	
14		
15		
16		
17		Реферат

6.4 Методика текущего контроля

на лекционных занятиях

Текущий контроль включает в себя:

- контроль посещаемости (не менее 80% занятий);

- проведение дискуссий в конце каждой лекции, активное участие в которых может учитываться преподавателем, как один из способов текущего контроля на лекционных занятиях.

на практических (семинарских) занятиях

Текущий контроль включает в себя:

- контроль посещаемости (не менее 80 % занятий);

- устный опрос по теме практического занятия;

- выполнение одного теста по 1-7 теме. Тест состоит из 10 тестовых за-

даний (критерии оценки: «удовлетворительно» необходимо дать правильные ответы не менее 60% тестовых вопросов, «хорошо» - не менее 75% или более, «отлично» - более 90%);

- защита рефератов на 13 - 17 неделях по мере подготовки их обучающимися. Критерии оценки реферата: «зачтено» - реферат /доклад соответствует целям и задачам поставленной темы, отражена новизна проблематики, авторская позиция. Обучающийся представил презентацию, показал умение работать с литературой, систематизировать и структурировать материал, в докладе продемонстрировано умение обобщать, сопоставлять различные точки зрения по рассматриваемому вопросу, аргументировать основные положения и выводы. Оформление отчетного материала соответствует заданным преподавателем требованиям

«Не зачтено» - обучающийся не знает, не ориентируется в материале, представленном в реферате.

В ходе проведения практических занятий целесообразно привлечение студентов к как можно более активному участию в дискуссиях, решении задач, обсуждениях и т. д. При этом активность студентов также может учитываться преподавателем, как один из способов текущего контроля на практических занятиях. В соответствии с графиком текущего контроля успеваемости студенты проходят тестирование,по результатам которого выставляется оценка по пятибалльной системе. Студенты, получившие за тест оценку не менее ”удовлетворительно”, а также предоставившие и защитившие реферат, допускаются к зачету.

самостоятельной работы студентов

Контроль самостоятельной работы студентов осуществляется на лекционных и практических занятиях студентов по методикам, описанным выше.

7 Описание информационных технологий и материально-технической базы

Тип занятий	Тип помещения	Требования к помещению	Требования к программному обеспечению
Лекция	Лекционная аудитория	1) Количество посадочных мест – в соответствии с контингентом, 2) рабочее место преподавателя, персональный компьютер IBM, совместимый Pentium или выше, проектор, экран/интерактивная панель, меловая/маркерная доска.	1) Windows 7 и выше; 2) Microsoft Office 2007 и выше.
Практические занятия	Аудитория	1) Количество посадочных мест, оборудованных компьютерами IBM совместимыми Pentium или выше, – в соответствии с контингентом, 2) рабочее место преподавателя, персональный компьютер IBM совместимый Pentium или выше, проектор, экран/интерактивная панель, меловая/маркерная доска.	1) Windows 7 и выше; 2) Microsoft Office 2007 и выше.
Самостоятельная работа	Помещение для самостоятельной работа	Оснащено компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета.	1) Windows XP и выше; 2) Microsoft Office 2007 и выше

8 Адаптация рабочей программы для лиц с ОВЗ

Адаптированная программа разрабатывается при наличии заявления со стороны обучающегося (родителей, законных представителей) и медицинских показаний (рекомендациями психолого-медико-педагогической комиссии). Для инвалидов адаптированная образовательная программа разрабатывается в соответствии с индивидуальной программой реабилитации.

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

№ п/п	Дата	Изменение	Дата и номер протокола заседания УМК	Автор	Начальник ОМОЛА
1	23.12.2021	Внесены изменения в компетентностную модель образовательной программы, на основании письма Минобрнауки России от 21.12.2021 № МН-5/22720	23.12.2021 №9		