ТЕМА	ЭЛЕМЕНТ	БАЗОВЫЙ	ПРОДВИНУТЫЙ	ЭКСПЕРТНЫЙ	МАТЕРИАЛЫ
Понятие искусственного интеллекта	Знания	Общее понятие искусственного интеллекта. Базовые принципы работы и обучения нейронных сетей. История возникновения термина искусственный интеллект: - "Искусственный интеллект" история возникновения термина в странах мира и в России. - Современное понимание ИИ в различных юрисдикциях и в различных видах документов.	Особенности технологии искусственного интеллекта, связанные с возможностью возникновения этически и социально значимых проблем в контексте ИИ: непрозрачность алгоритмов, высокий уровень автономии систем ИИ, зависимость качества работы ИИ-сервисов от качества, репрезентативности, сбалансированности и разметки данных, используемых для обучения.	Принципы работы и обучения нейронных сетей. Глубокое обучение, обучение с подкреплением. Объяснимость и интерпретируемость моделей. Понимание, какие задачи решает ИИ, каковы основные области применения технологии (распознавание речи и образов, обработка естественного языка, рекомендации и пр.).	ПРИКАЗ МИНИСТЕРСТВА ЭКОНОМИЧЕСКОГО РАЗВИТИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ ОТ 29.06.2021 № 392 "ОБ УТВЕРЖДЕНИИ КРИТЕРИЕВ ОПРЕДЕЛЕНИЯ ПРИНАДЛЕЖНОСТИ ПРОЕКТОВ К ПРОЕКТАМ В СФЕРЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА" HTTP://PUBLICATION.PRAVO.GOV.RU/DOCUMENT/VIEW/0001202107290008 НАЦИОНАЛЬНАЯ СТРАТЕГИЯ: УКАЗ ПРЕЗИДЕНТА РОССИЙСКОЙ ФЕДЕРАЦИИ ОТ 15.02.2024 Г. № 124 О ВНЕСЕНИИ ИЗМЕНЕНИЙ В УКАЗ ПРЕЗИДЕНТА РОССИЙСКОЙ ФЕДЕРАЦИИ ОТ 10 ОКТЯБРЯ 2019 Г. № 490 «О РАЗВИТИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В РОССИЙСКОЙ ФЕДЕРАЦИИ» И В НАЦИОНАЛЬНУЮ СТРАТЕГИЮ, УТВЕРЖДЕННУЮ ЭТИМ УКАЗОМ HTTP://KREMLIN.RU/ACTS/BANK/50326 ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: СОВРЕМЕННЫЙ ПОДХОД / С. РАССЕЛ, П. НОРВИГ ; [ПЕР. С АНГЛ. И РЕД. К. А. ПТИЦИНА]. – 2-Е ИЗД. – МОСКВА : ВИЛЬЯМС, 2005. – 1408 С. : ИЛ. – ПАРАЛ. ТИТ. Л. НА АНГЛ. ЯЗ. – ПРИЛ.: С. 1288–1301 МАРКОВ С. ОХОТА НА ЭЛЕКТРООВЕЦ. БОЛЬШАЯ КНИГА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА. МОСКВА, 2024: HTTPS://MARKOFF.SCIENCE/#BOOK + ДОКУМЕНТЫ ПО СТАНДАРТИЗАЦИИ, КОНВЕНЦИЯ ЮНЕСКО ПО ЭТИКЕ
	Практическое задание	Сравните дефиницию термина ИИ в различных юрисдикциях. Ответьте на вопрос - существует ли отличие в понимании указанного термина?	Представьте, что Вы участвуете в разработке системы ИИ для автоматизированного принятия решений (например, для решения задачи по подбору персонала, выдаче кредита и т.п.). Какие данные требуются для обучения этой системы ИИ? Какими могут быть социально значимые следствия недостаточной репрезентативности или несбалансированности данных?	Выберите в списке технологий те, что не относятся к технологиям ИИ и поясните выбор: 1. Автоматизация. 2. Глубокое обучение. 3. Компьютерное зрение. 4. Экспертные системы.	
	Вопрос для контрольной/экзамена	Дайте определение ИИ и приведите различия в их понимании в различных странах.	Чем ИИ отличается от автоматизации или обычной ИИ-системы?	Объясните основные принципы работы и обучения нейронных сетей, выделяя различия между сверточными, рекуррентными нейросетями и трансформерами. Опишите, в чем заключается глубокое обучение и обучение с подкреплением, а также обсудите важность объяснимости и интерпретируемости моделей. Приведите примеры задач, которые может решать ИИ, и основные области применения этих технологий, такие как распознавание речи, обработка естественного языка и рекомендательные системы.	
Механизмы регулирования в сфере искусственного интеллекта	Знания	I. Особенности различных типов регулирования ИИ. Принцип взаимодополнения трех типов регулирования в сфере ИИ: 1. Этика в сфере ИИ как инструмент саморегулирования и источник будущих правовых норм в сфере ИИ. 2. Правовое регулирование ИИ: возможности и ограничения. 3. Нормативно-техническое регулирование: стандартизация и сертификация в сфере ИИ. II. Вертикальный и горизонтальный подходы к регулированию ИИ: определения и цели.	Специфика регулирования ИИ в РФ. Ключевые стратегические документы: 1. Общие положения Национальной стратегии РФ в области ИИ; 2. Концепция регулирования ИИ - общее направление развития регулирования технологии; 3. ФП "Искусственный интеллект" нац. программы "Цифровая экономика РФ" - общий вектор развития; Нормативно-правовое регулирование в сфере ИИ: - ЭПР- экспериментальные правовые режимы в РФ: сферы применения, цели введения, особенности функционирования. - Правовые основы интеллектуальной собственности при использовании ИИ: Условия использования Yandex GPT, пользовательские соглашения Fusion Brain и Шедеврум - Регулирование данных - Регулирование рекомендательных сервисов - Регулирование беспилотных автомобилей - Регулирование ИИ в здравоохранении - Ключевые национальные стандарты в сфере ИИ Этическое регулирование: - Базовые принципы Кодекса этики в сфере ИИ; - Декларация об ответственной разработке и использовании сервисов на основе генеративного искусственного интеллекта генеративном ИИ; - Декларация об ответственном экспорте технологий ИИ и ПО на их основе. - Белая книга этики в сфере ИИ.	Индексы развития ИИ в разных странах - регуляторная политика и этика ИИ: 1. AI Index report Stanford 2. The Global AI Index (Tortois) 3. Этические индексы Канады и Китая Специфика международного этического регулирования ИИ по уровням: 1. Международное сообщество: основные идеи Рекомендации ОЭСР по ИИ, Рекомендации ЮНЕСКО об этических аспектах ИИ, Глобального цифрового договора ООН. 2. Национальные принципы этики ИИ: ЕС, США, Китай, Сингапур, Англия, Южная Корея. 3. Отдельные организации: Google, Misrosoft, Baidu, SAP + российские 4. НКО и частные группы исследователей: Open AI, Montreal Declaration for Responsible AI Development. Глобальный атлас регулирования ИИ: специфика национального регулирования ИИ в зарубежных странах лидерах по разработке и / или внедрению ИИ: ЕС, Китай, США, Сингапур. Специфика разработки технических стандартов, затрагивающих этические аспекты технологии ИИ.	НАЦИОНАЛЬНАЯ СТРАТЕГИЯ: УКАЗ ПРЕЗИДЕНТА РОССИЙСКОЙ ФЕДЕРАЦИИ ОТ 15.02.2024 Г. № 124 О ВНЕСЕНИИ ИЗМЕНЕНИЙ В УКАЗ ПРЕЗИДЕНТА РОССИЙСКОЙ ФЕДЕРАЦИИ ОТ 10 ОКТЯБРЯ 2019 Г. № 490 «О РАЗВИТИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В РОССИЙСКОЙ ФЕДЕРАЦИИ» И В НАЦИОНАЛЬНУЮ СТРАТЕГИЮ, УТВЕРЖДЕННУЮ ЭТИМ УКАЗОМ KREMLIN.RU КОДЕКС ЭТИКИ В СФЕРЕ ИИ РФ ETHICS.A-AI.RU ДЕКЛАРАЦИЯ ОБ ОТВЕТСТВЕННОЙ РАЗРАБОТКЕ И ИСПОЛЬЗОВАНИИ СЕРВИСОВ НА ОСНОВЕ ГЕНЕРАТИВНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ГЕНЕРАТИВНОМ ИИ ETHICS.A-AI.RU ЭПР В РФ WWW.ECONOMY.GOV.RU TC164.RU WWW.ISO.ORG WWW.UN.ORG AI.GOV.RU AI.GOV.RU AI.GOOGLE WWW.MICROSOFT.COM WWW.SBER-BANK.BY
	Практическое задание	1. Сформулируйте преимущества и ограничения каждого их 3 типов регулирования для развития технологии ИИ.	Проанализируйте конкретный работающий сервис, основанный на ИИ, на предмет его соответствия Кодексу этики в сфере ИИ в РФ. В случае если проверка некоторых положений Кодекса оказалась неосуществимой из-за недостатка информации из открытых источников, оцените, насколько это повлияет на Ваш пользовательский опыт.	1. Глобальный опыт регулирования: создайте таблицу, обобщающую мировой опыт регулирования ИИ, с выделением вертикальных и горизонтальных подходов и перспектив развития технологии. 2. Спроектируйте, какими могут быть следующие шаги в регулировании технологии с учетом текущих технических возможностей и вызовов ИИ.	
	Вопрос для контрольной/экзамена	Каковы отличия правового регулирования от этического? В чем заключается преимущество этического регулирования в сфере ИИ по сравнению с правовым и нормативно-техническим?	Опишите основные компоненты отечественной системы регулирования в сфере ИИ. Как компоненты связаны между собой?	Какой ключевой фактор, по вашему мнению, необходимо учитывать при оценке соответствия этических и правовых норм уровню развития технологии ИИ? Объясните, почему этот фактор важен и как он может повлиять на разработку более эффективного регулирования.	
Ключевые этические принципы, ценности и риски в сфере ИИ	Знания	Основные этические принципы в сфере ИИ: справедливость, прозрачность, конфиденциальность, объяснимость, надежность, ответственность, человекоориентированность, автономия принятия решений человеком. Понимание ключевых принципов этики, включая справедливость (например, обеспечение отсутствия предвзятости в алгоритмах), прозрачность (например, разъяснение пользователям, как принимаются решения на основе ИИ), и ответственность (например, установление механизмов подотчетности для разработчиков и пользователей ИИ). Ключевые этические риски в сфере ИИ: алгоритмически обусловленная предвзятость, недостаточный уровень предсказуемости и прозрачности ИИ, дезинформация, "галлюцинирование", риск нарушения конфиденциальности, возможность дестабилизация рынка труда, утечки данных, риски, связанные с распределением ответственности, определением владельца авторских прав, экологические риски, риски опасного использования ИИ. Умение анализировать социальные, экономические и экологические последствия применения ИИ-технологий с использованием методов оценки воздействия, таких как анализ сценариев или социально-экономическое моделирование.	Концепция риск-ориентированного подхода в рамках этики в сфере ИИ. Понимание влияния этических принципов на принятие решений в реальных ситуациях разработки и использования ИИ на примере конкретных кейсов. Знание подходов для поиска компромиссных решений при столкновении различных этических требований в контексте ИИ (например, между конфиденциальностью и безопасностью) и умение смоделировать ситуацию.	Особенности разработки этических стандартов и политик в сфере ИИ с учетом особенностей ИИ-сервиса, а также положений национального регулирования сферы ИИ; специфика разработки нормативных рекомендаций и стандартов, применимых в различных юрисдикциях, с учетом особенностей ИИ-сервиса и международных аспектов регулирования ИИ. Система управления этическими рисками на уровне организации и шире - общества, экосистемы (в том числе представление о методах оценки и снижения рисков, связанных с долгосрочными и системными последствиями использования ИИ, включая потенциальное воздействие на различные социальные группы).	КОДЕКС ЭТИКИ В СФЕРЕ ИИ РФ ETHICS.A-AI.RU ДЕКЛАРАЦИЯ ОБ ОТВЕТСТВЕННОЙ РАЗРАБОТКЕ И ИСПОЛЬЗОВАНИИ СЕРВИСОВ НА ОСНОВЕ ГЕНЕРАТИВНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ГЕНЕРАТИВНОМ ИИ ETHICS.A-AI.RU AI RISK REPOSITORY AIRISK.MIT.EDU (MIT) ЭПР В РФ WWW.ECONOMY.GOV.RU NORMATIVNOE_REGULIROVANIE_CIFROVOY_SREDY/EKSPERIMENTALNYE_PRAVOVYE_REZHIMY/ TC164.RU ISO/IEC JTC 1/SC 42 ARTIFICIAL INTELLIGENCE ISO.ORG WWW.UN.ORG STRATEGY.EC.EUROPA.EU WWW.NIST.GOV DEEPMIND.GOOGLE
	Практическое задание	Приведите пример существующего (существовавшего ранее) сервиса на основе ИИ. 1. Проанализируйте, какие этические принципы утверждаются или нарушаются при разработке и использовании данного сервиса (примерный перечень ценностей включает в себя: человекоориентированность, автономия человека, непредвзятость, справедливость, отсутствие дискриминации, неприкосновенность частной жизни, защита от опасного использования). 2. Перечислите потенциальные риски, которые могут быть связаны с использованием сервиса. Поясните и обоснуйте свой выбор.	Как ответственность за ошибки или негативные последствия работы ИИ должна быть распределена между разработчиками, пользователями и владельцами системы? Какие этические принципы помогут правильно определить границы этой ответственности?	Крупная компания разрабатывает ИИ-сервис (можно взять, например, образовательный сервис, сервис по подбору персонала, сервис генерации изображений и.т.п.), который планируется выводить на международный рынок. Подготовьте стратегию управления этическими рисками и нормативные рекомендации, учитывающие как национальные, так и международные аспекты регулирования в сфере ИИ. 1. Опишите специфику разработки этических стандартов для вашего ИИ-сервиса. Укажите, какие особенности сервиса необходимо учитывать. 2. Предложите нормативные рекомендации, которые могут быть применимы в разных юрисдикциях. 3. Разработайте стратегию управления этическими рисками, учитывая долгосрочные и системные последствия использования ИИ-сервиса. Включите методы оценки и снижения рисков, а также опишите, как ваше решение защитит интересы различных социальных групп.	
	Вопрос для контрольной/экзамена	Объясните, как принципы справедливости, прозрачности и объяснимости применяются при разработке и внедрении систем искусственного интеллекта.	Что подразумевается под риск-ориентированным подходом в этике в сфере ИИ? Как этот подход может помочь в минимизации потенциальных негативных последствий? Приведите пример того, как такой подход можно применить на практике при разработке и внедрении системы ИИ в социальной сфере.	Опишите ключевые аспекты разработки этических стандартов для ИИ-сервиса, предназначенного для использования в международном контексте.	
Способы контроля возникновения этических рисков в сфере ИИ	Знания	Технологические методы интеграции этических ценностей и принципов в технологии ИИ: обучение с подкреплением на основе обратной связи от людей (RLFH), работа ИИ-тренеров, RAG-решения, оценка возможностей моделей по решению этических задач (бенчмаркинг) методы формирования тестовых наборов для создания ИИ-решений, соответствующих ожиданиям общества.	Подход комплексной имитации реальных атак (red teaming) для выявления этически значимых уязвимостей систем ИИ включает следующие методы: Организация внутренних испытаний для разработчиков. Проведение тщательно спланированных внутренних тестов с целью выявления и устранения этических уязвимостей в ИИ-системах. Это может включать симуляцию сценариев, в которых ИИ может действовать предвзято, нарушать конфиденциальность данных или принимать решения, ведущие к непредвиденным последствиям. Модели проведения хакатонов. Создание и проведение хакатонов, где участники сосредоточены на обнаружении и демонстрации потенциальных этических рисков и слабых мест в системах ИИ. Эти мероприятия позволяют привлекать разнообразные команды экспертов, включая разработчиков, юристов и специалистов по этике. Целенаправленный джейлбрейкинг. Испытания, направленные на взлом или обход защитных механизмов ИИ-систем, чтобы оценить их устойчивость к злоупотреблениям и выявить сценарии, в которых алгоритмы могут быть использованы во вред. Такой подход помогает разработчикам понять, как ИИ может быть манипулирован или использован с нарушением этических норм.	Концепция организации внешнего независимого этического комитета, консультирующего по этически значимым аспектам проектирования, разработки, тестирования, внедрения и использования искусственного интеллекта. Подходы к формированию внутренней политики по повышению уровня соответствия ИИ-решений этическим стандартам для разработчика ИИ-решений. Подходы к разработке дополнительных соглашений к договору между заинтересованными сторонами с целью урегулирования этически значимых аспектов внедрения и/или эксплуатации ИИ-сервисов: - Введение этических обязательств; - Определение ответственности; - Механизмы мониторинга и отчетности; - Процедуры управления рискам; - Информированное согласие пользователей; - Урегулирование спорных вопросов.	ARXIV.ORG ARXIV.ORG ARXIV.ORG SYSBLOK.RU AWS.AMAZON.COM WWW.SBERBANK.RU WWW.ANTHROPIC.COM AR5IV.LABS.ARXIV.ORG WWW.IMD.ORG
	Практическое задание	Составьте перечень существующих бенчмарков, позволяющих оценить ответы модели ИИ на предмет соответствия нормам этики ответы модели ИИ.	Проведите тестирование одной из моделей и найдите этические уязвимости. Опишите возможные способы устранения выявленных уязвимостей.	Разработайте пример пункта дополнительного соглашения к договору между заказчиком и разработчиком, который регулировал бы этические аспекты (например, обязательства по обеспечению прозрачности, механизм разрешения этических споров).	
	Вопрос для контрольной/экзамена	Какие этические ценности помогает внедрить в ИИ обучение с подкреплением на основе обратной связи от людей (RLFH)?	Составьте перечень задач, которые решаются с помощью метода red teaming	Какие преимущества связаны с созданием внешнего независимого этического комитета для компании-разработчика ИИ?	
Основы отраслевого этического регулирования	Знания	Понимание ключевых этических принципов и стандартов, применимых к конкретным отраслям (например, здравоохранение, транспорт, финансы). Способность идентифицировать основные этические вызовы, связанные с использованием ИИ в отраслевых приложениях. Умение объяснить важность соблюдения этических норм и стандартов в профессиональной практике.	Знание основных нормативных актов и кодексов, регулирующих использование ИИ в этих областях. Глубокое понимание отраслевых требований (образование, медицина, правосудие, пользовательские сервисы) к этике ИИ, включая международные и национальные стандарты, а также специфику регулирования в разных юрисдикциях. Знание механизмов контроля и обеспечения соблюдения этических норм. Способность анализировать и оценивать этические риски, связанные с внедрением ИИ в конкретной отрасли, а также предлагать меры по их минимизации. Умение разрабатывать и обосновывать политику этичного использования ИИ для конкретных проектов.	Особенности существующего нормативного-технического, правового и этического регулирования социально значимых отраслей использования ИИ. Специфика разработки этических и технических стандартов, основ правового регулирования для социально значимых отраслей использования ИИ. национальных и международных законодательных актах, регулирующих использование ИИ, а также способность адаптировать и применять эти нормы в конкретной отрасли различий в регулировании ИИ в разных странах и умение разрабатывать универсальные или адаптируемые этические стандарты и рекомендации, которые учитывают культурные, социальные и юридические особенности различных юрисдикций. существующих законодательных и нормативных актах, регулирующих использование ИИ в разных странах и отраслях. Умение адаптировать эти стандарты к специфическим условиям разных юрисдикций и отраслей	AI.GOV.RU ETHICS.CDTO.CENTER WWW.WHO.INT WWW.ICMR.GOV.IN UNESDOC.UNESCO.ORG WWW.UNESCO.ORG WWW.OECD-ILIBRARY.ORG RUSSELLGROUP.AC.UK WWW.CBR.RU TC164.RU
	Практическое задание	Описать, какие этические принципы наиболее важны для конкретной отрасли и почему. Привести пример, как нарушение этих норм может повлиять на пользователей и общество.	Провести анализ кейса, связанного с нарушением этических норм при применении ИИ в отрасли (например, предвзятость в медицинских данных), и предложить способы улучшения этического регулирования.	Представьте, что вы являетесь консультантом по этике и правовому регулированию ИИ. Вам поручили подготовить рекомендации для внедрения системы искусственного интеллекта в социально значимую сферу, например, в образование или здравоохранение и т.п.. 1. Определите ключевые этические и технические стандарты, которые необходимо учитывать. 2. Опишите существующие правовые аспекты регулирования , которые обеспечат безопасное и ответственное использование технологии. Обозначьте, если выявили, пробелы в правовом регулировании. 3. Разработайте краткий план рекомендаций по учету этических и правовых аспектов при внедрении системы ИИ, включая меры по защите прав и интересов пользователей.	
	Вопрос для контрольной/экзамена	Привести пример наиболее важных и применимых этических принципов применения ИИ в одной из социально значимых отраслей.	Какие основные этические риски связаны с использованием технологий искусственного интеллекта в отрасли (например, в здравоохранении, образовании, транспорте и т.п.)? К каким потенциальным последствиям могут привести ошибки или смещение данных, используемых для обучения систем ИИ в транспорте или системах распознававания медицинских изображений и т.п.? Какие меры могут быть приняты для их минимизации?	Какие ключевые этические и правовые принципы необходимо учитывать при разработке и внедрении систем искусственного интеллекта, например, в образование (выбирается любая социально значимая сфера). Почему соблюдение этих принципов важно для защиты интересов человека и общества?	