ЛЕКЦИЯ 15
ЭТИКА, БЕЗОПАСНОСТЬ И БУДУЩЕЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Цель лекции

Сформировать понимание этических, правовых и безопасностных аспектов разработки и применения ИИ; рассмотреть справедливость, прозрачность, конфиденциальность, авторское право и ответственность, безопасность генеративных и агентных систем, понятие AGI и влияние ИИ на рынок труда.

Этика и безопасность как часть ИИ

Качество интеллектуальной системы определяется не только точностью модели. Реальное применение затрагивает людей, данные, права и организационные процессы, поэтому необходимо оценивать безопасность, справедливость, прозрачность, конфиденциальность и ответственность.

Responsible AI

Responsible AI определяется как подход к разработке и использованию ИИ с учётом безопасности, справедливости, прозрачности, конфиденциальности и ответственности.

Этические проблемы

Этические вопросы возникают, когда автоматизированное решение влияет на человека или общество: дискриминация, непрозрачность критериев, чрезмерная автоматизация, неправомерная обработка данных и вводящий в заблуждение контент.

Смещения в данных

Модель может воспроизводить систематические искажения обучающих данных. Причинами являются несбалансированная выборка, исторические практики, ошибки разметки и недостаточное представительство отдельных случаев.

Справедливость

Справедливость требует анализа качества не только в среднем, но и для значимых групп и сценариев. Высокая общая accuracy может скрывать серьёзные ошибки для меньшего класса.

Прозрачность

Пользователь должен понимать назначение системы, область применения, основные ограничения и роль автоматизации. Прозрачность не обязательно означает раскрытие всех технических деталей.

Объяснимость

Объяснимость связана с возможностью понять факторы, повлиявшие на результат. Требуемый уровень зависит от риска и предметной области.

Ответственность

ИИ может поддерживать решение, но ответственность организации и человека не исчезает. Необходимо определить, кто утверждает модель, контролирует данные, принимает итоговое решение и рассматривает ошибки.

Персональные данные

При разработке ИИ необходимо минимизировать объём данных, ограничивать доступ, защищать хранение и не передавать чувствительную информацию внешним сервисам без соответствующего разрешения и мер защиты.

Авторское право

Авторское право выделеяется как отдельный аспект. Генеративный ИИ создаёт вопросы о происхождении обучающих материалов, допустимости использования результата и обязанностях пользователя. Конкретные выводы зависят от применимого законодательства и лицензий.

Академическая добросовестность

ИИ может помогать структурировать и редактировать материал, но не должен использоваться для фабрикации данных, ссылок или научных результатов. Библиографические сведения необходимо проверять.

Достоверность контента

Генеративная модель способна создавать убедительный текст, изображение, аудио или видео, не соответствующие реальности. Правдоподобие не является доказательством достоверности.

ИИ как инструмент и объект риска

Двойственная роль ИИ: он помогает автоматизировать защиту, но генеративные модели могут масштабировать социальную инженерию и вводящий в заблуждение контент, а сами AI-системы могут становиться объектом атак.

Безопасность данных

Необходимо контролировать происхождение, целостность, актуальность и права доступа к данным. Ошибки или подмена данных могут приводить к систематически неверному поведению модели.

Безопасность модели

Модель следует тестировать не только на средней точности, но и на редких и потенциально опасных сценариях. Для генеративных систем оцениваются нежелательные ответы, утечки и устойчивость к манипулятивным входам.

Безопасность промптов и контекста

Prompt Injection показывает, что внешний текст может пытаться изменить поведение системы. Доверенные инструкции и недоверенные данные необходимо разделять.

Безопасность AI-агентов

Агент способен выполнять действия, поэтому цена ошибки выше. Необходимы минимальные привилегии, журналирование, условия остановки и подтверждение критических операций.

Human oversight

Человек должен иметь возможность проверить существенные факты, остановить автоматизированный процесс и пересмотреть ошибочное решение, особенно в высокорисковых сценариях.

Управление рисками

Управление рисками включает идентификацию возможного ущерба, оценку вероятности и последствий, внедрение мер контроля, тестирование и мониторинг после внедрения.

Мониторинг после внедрения

Качество модели может меняться из-за сдвига данных и условий эксплуатации. Поэтому отслеживаются метрики, ошибки, жалобы пользователей, изменения данных и инциденты безопасности.

AGI

Artificial General Intelligence обычно обозначает гипотетические системы с широкими интеллектуальными возможностями, не ограниченными одной узкой задачей. Единого общепринятого технического критерия достижения AGI нет.

Узкий ИИ и AGI

Современные прикладные системы решают определённые задачи или наборы задач и оцениваются конкретными метриками. AGI остаётся концепцией и предметом научной и общественной дискуссии.

Прогнозы о будущем ИИ

Долгосрочные прогнозы различаются из-за неопределённости темпов развития моделей, стоимости вычислений, регулирования, данных и общественного принятия. Их следует рассматривать как сценарии, а не гарантированные события.

ИИ и рынок труда

Корректнее анализировать не только исчезновение профессий, но и автоматизацию отдельных задач и изменение содержания профессиональной деятельности.

Автоматизация задач

ИИ способен автоматизировать поиск, первичную классификацию, обработку типовых данных и подготовку черновиков. Одновременно возрастает роль постановки задачи, проверки, ответственности и предметной экспертизы.

Новые компетенции

Специалистам становятся необходимы AI literacy, работа с данными, критическая проверка результатов, информационная безопасность, предметная экспертиза и понимание этических последствий.

ИИ в образовании

ИИ расширяет возможности объяснения и индивидуальной поддержки, но создаёт вопросы самостоятельности, авторства, качества и конфиденциальности. Учебные задания должны оценивать понимание, а не только готовый текст.

ИИ в науке

ИИ может помогать с анализом данных, программированием и структурированием текста. Исследователь остаётся ответственным за достоверность данных, корректность методов, библиографию и выводы.

ИИ в информационной безопасности

ИИ применяется для обнаружения аномалий, анализа трафика, вредоносного ПО, фишинга и журналов. Одновременно необходимо защищать сами AI-системы и учитывать злоупотребление генеративными возможностями.

Баланс инноваций и контроля

Полный отказ от ИИ лишает организацию полезной автоматизации, а неконтролируемое внедрение создаёт риски. Практический подход — выбирать допустимые сценарии, классифицировать данные, тестировать, постепенно расширять автономность и проводить мониторинг.

Принципы Responsible AI

Принцип	**Контрольный вопрос**
Безопасность	Не создаёт ли система неприемлемый риск?
Справедливость	Не возникает ли систематическая дискриминация?
Прозрачность	Понятны ли назначение и ограничения?
Конфиденциальность	Правомерно ли обрабатываются данные?
Ответственность	Кто отвечает за итоговое решение?

Риски на жизненном цикле

Этап	**Риск**	**Пример контроля**
Данные	Смещение, утечка, подмена	Проверка качества и доступа
Обучение	Переобучение и нежелательные закономерности	Валидация
Оценка	Неподходящие метрики	Несколько метрик и анализ ошибок
Внедрение	Неправильная автоматизация	Human-in-the-loop
Эксплуатация	Сдвиг данных	Мониторинг

Основные выводы

Этика и безопасность являются частью жизненного цикла ИИ.

Responsible AI включает безопасность, справедливость, прозрачность, конфиденциальность и ответственность.

Генеративные и агентные системы требуют усиленного контроля из-за способности создавать контент и выполнять действия.

AGI остаётся концепцией без единого общепринятого критерия.

Влияние ИИ на рынок труда следует рассматривать через автоматизацию задач и изменение компетенций.

Ответственное внедрение требует мониторинга и человеческого контроля.

Контрольные вопросы

1. Что такое Responsible AI и какие принципы он включает?
2. Как смещения в данных влияют на справедливость решений?
3. Чем прозрачность отличается от объяснимости?
4. Какие требования важны при работе с персональными данными?
5. Какие вопросы авторского права и академической добросовестности возникают при генеративном ИИ?
6. Почему ИИ является одновременно инструментом защиты и объектом риска?
7. Какие меры необходимы для безопасного использования AI-агентов?

8. Что такое AGI и почему нет единого критерия его достижения?

9. Как ИИ может изменить рынок труда и профессиональные компетенции?

10. Какие этапы включает ответственное внедрение и мониторинг ИИ?

Список источников

1. Бессмертный, И. А. Системы искусственного интеллекта: методическое пособие / И. А. Бессмертный, А. Е. Авдюшина, А. В. Кугаевских, Ю. А. Королева, Н. Г. Рущенко. – Санкт-Петербург: Университет ИТМО, 2024. – 281 с.

2. Соснило, А. И. Атлас искусственного интеллекта для бизнеса и власти: учебное пособие / А. И. Соснило. – Санкт-Петербург: Университет ИТМО, 2022. – 98 с.

3. Сидоркина, И. Г. Системы искусственного интеллекта: учебное пособие / И. Г. Сидоркина. – Москва: КНОРУС, 2018. – 248 с.

4. Баланов, А. Н. Искусственный интеллект. Понимание, применение и перспективы: учебник для вузов / А. Н. Баланов. – Санкт-Петербург: Лань, 2024. – 312 с.

5. Калинин, М. О. Основы искусственного интеллекта. Безопасность искусственного интеллекта: учебное пособие / М. О. Калинин, В. М. Крундышев. – Санкт-Петербург, 2024. – 220 с.

6. Андреянов, Н. В. Методы искусственного интеллекта: учебное пособие / Н. В. Андреянов, Т. С. Евдокимова, А. Д. Павлов [и др.]. – Казань: КНИТУ-КАИ, 2024. – 392 с.

7. Сасскинд, Д. Будущее без работы. Технологии, автоматизация и стоит ли их бояться / Д. Сасскинд ; пер. с англ. под ред. А. Дунаева. – Москва: Индивидуум, 2021. – 352 с.

8. Основы искусственного интеллекта: учебное пособие. – Гл. 15: Этика, безопасность и будущее ИИ.