**Доверенный искусственный интеллект**

**Введение**

А.Е. Самотуга, П.С. Ложников

Конспект лекции

# Введение

## 1 Предмет дисциплины и ее задачи

Цель дисциплины изучение проблем обеспечения доверия к искусственному интеллекту (ИИ) и подходов к их решению и формирование знаний навыков составления требований к системам доверенного искусственного интеллекта на основе анализа профессиональной информации из данной области.

**Задачи дисциплины:**

* Изучение проблемы обеспечения доверия к искусственному интеллекту (ИИ) и подходы к их решению. Выработка практические навыки 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 представление 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.
* Знание основных источников патентной и научно­технической 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.
* 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
* Владение навыками проведения аналитико­синтетического исследования патентов и научно­технических публикаций в области обеспечения доверия к искусственному интеллекту; профессиональной информацией в области патентов, стандартов и научно­технических 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.

## Содержание и форма проведения занятий

Дисциплина «Доверенный искусственный интеллект» включает в себя следующие разделы (модули):

1. Проблемы доверия искусственному интеллекту;
2. Управление рисками искусственного интеллекта;
3. Объяснимость моделей искусственного интеллекта и машинного обучения;
4. Робастность искусственного интеллекта и нейронных сетей;
5. Функциональная безопасность искусственного интеллекта;
6. Компьютерные атаки на искусственный интеллект;
7. Защита данных;
8. Обнаружение аномалий;
9. Заключение.

## 3 Проблема обеспечения доверия к искусственному интеллекту

В ходе обучения было усвоена ключевая особенность искусственного интеллекта - способность имитировать когнитивные функции человека выполнять практически значимые задачи обработки данных и получать результаты, как минимум сопоставимые с результатами интеллектуальной деятельности человека. При таком подходе, на первый взгляд, не должно быть значимых преград для повсеместного внедрения ИИ в любые практические области деятельности человека. Но остается проблема обеспечения доверия ИИ, которая все еще непреодолимо стоит на пути интенсификации процесса цифровой трансформации экономики с активным задействованием ИИ.

В случае применения искусственного интеллекта на предприятиях и в отраслях, где требуются надежные, объяснимые, понятные подходы к принятию решений, важно использовать доверенные системы ИИ. То есть системы, к которым потребитель и регулирующие организации могут проявлять доверие. Тогда важно определиться с тем, что мы будем понимать под доверием системам искусственного интеллекта. Здесь мы можем опираться на ГОСТ Р 59276-2020. Согласно документу, доверие к системе ИИ обеспечивается подтверждением соответствия представительного набора существенных характеристик системы ИИ требованиям: разработчиком системы ИИ, потребителем системы ИИ, регулятором. В качестве существенных характеристик выделяют: функциональные возможности, надежность, эффективность, практичность, сопровождаемость, мобильность.

**Список источников**

1. ГОСТ Р 59276-2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения от 23 декабря 2020 - docs.cntd.ru [Электронный ресурс]. URL:<https://docs.cntd.ru/document/1200177291> (дата обращения: 12.04.2022)