

**Доверенный искусственный интеллект**

**Заключение**

А.Е. Самотуга, П.С. Ложников

**ЗАКЛЮЧЕНИЕ**

## Итоги курса

В рамках курса мы рассмотрели компоненты, требования к доверенному искусственному интеллекту, что позволило получить точное понимание данного направления ИИ. Познакомились с особенностями источников рисков и процесса оценки рисков для ИИ. Обсудили важность робастности ИИ, современные способы атак на ИИ и угрозы ИИ.

Рассматриваемый материал позволил сделать ряд выводов:

Доверенный ИИ является более обширной областью знаний, чем защищенный ИИ. Он включает не только защиту от атак, устранение угроз и защиты данных ИИ, но такие важные направления, как оценка рисков, обнаружение аномалий, мониторинг, объяснимость ИИ;

Законодательство в области доверенного, защищенного ИИ на данный момент находиться в стадии разработки, большинство стандартов находятся на стадии проекта, следовательно до сих пор прорабатываются;

Потребность в доверенном ИИ связана с тем, что для сохранения, а тем более, для увеличения темпов цифровой трансформации и внедрения ИИ в бизнесс-процессы предприятий, нужны решения, которые будут в достаточной степени и с определенной вероятностью защищены, которые позволят сохранять в тайне сведения ограниченного доступа (конфиденциальные данные, персональные данные), риски от внедрения которых можно оценить, и которые, наконец, могут объяснить причину тех или иных результатов модели ИИ в понятной форме;

Кроме разработки нормативных документов в этой области, есть острая необходимость отечественных решений, особенно с учетом критичности данного направления, сложной международной обстановки, необходимости импортозамещения и в то же время, почти полным отсутствием отечественных решений.

## Перспективы развития отрасли доверенного ИИ

Ключевые перспективы, которые можно выделить, включают в себя рост рынка доверенного ИИ и появления все большего числа отечественных решений, развитие российского законодательства, стандартов, активную поддержку со стороны государства как подготовки кадров для данной отрасли, так и предприятий, разрабатывающих такого рода продукты.

Особое внимание будет уделяться как различным подходам к защищенности ИИ, так и к его объяснимости. Наиболее перспективные направления представляют собой разработка облачных систем для онлайн обучения, позволяющие создавать некоторые шаблонные модели, активно задействовать механизм переиспользования отдельных составляющих моделей, выполнять мониторинг поведения модели после ее внедрения и отделить ИИ от процесса принятия критических решений.