NIST запустил инициативу, которая не позволит машинам навредить человечеству.

Национальный институт стандартов и технологий (NIST – это Национальный институт стандартов и технологий, подразделение Министерства торговли США. Ранее известный как Национальное бюро стандартов, NIST продвигает и поддерживает стандарты измерений. У него также есть активные программы для поощрения и помощи промышленности и науки в разработке и использовании этих стандартов." data-html="true" data-original-title="NIST" >NIST) объявил о запуске новой программы Assessing Risks and Impacts of AI (ARIA), направленной на оценку безопасности и надёжности искусственного интеллекта. Программа поможет организациям и частным лицам понять, насколько безопасна и эффективна будет та или иная технология ИИ после её внедрения. Проект реализуется совместно с Институтом безопасности ИИ.

Зачем нужна ARIA?

ARIA появилась как ответ на исполнительный указ президента Байдена о безопасности ИИ, подписанный в 2021 году. Проект направлен на создание защищенной и справедливой экосистемы для разработки и применения продвинутых нейросетей. Основная цель указа — убедиться, что новые технологии приносят пользу обществу, а не создают угрозы.

Указ был разработан в результате множества консультаций с экспертами в области технологий, права, а также представителями бизнеса и гражданского общества. В нем прописаны основные принципы, которым должны следовать разработчики и пользователи ИИ: прозрачность, безопасность, защита данных и справедливость.

«Чтобы полностью понять влияние ИИ на наше общество, мы должны тестировать его в реальных условиях — и именно этим занимается программа ARIA», — заявила министр торговли США Джина Раймондо . Она отметила, что инициатива является частью усилий Министерства торговли по выполнению указа Байдена.

Участники ARIA будут оценивать, как технологии работают в различных реальных сценариях, а не только в лабораторных условиях. Это позволит выявить возможные негативные последствия на ранних этапах. Лори Е. Локасио, директор NIST, подчеркнула, что ARIA поддержит работу Института безопасности ИИ США и расширит взаимодействие NIST с научным сообществом.

Что уже сделано?

В январе 2023 года NIST выпустил документ под названием AI Risk Management Framework (Рамки управления рисками ИИ), который стал основой для программы ARIA. В документе прописано, как использовать количественные и качественные методы для анализа и мониторинга опасностей современного мира. ARIA разработает новые методологии и метрики для оценки того, насколько хорошо системы ИИ работают в различных социальных контекстах.

Результаты исследований ARIA помогут NIST и Институту безопасности ИИ США создавать надёжные системы, гарантируя пользователям полную защиту и конфиденциальность.