Генеральный директор Anthropic Дарио Амодеи заявил, что его компания намерена к 2027 году научиться выявлять большинство проблем в работе передовых моделей искусственного интеллекта (ИИ), сообщает TechCrunch. В эссе «Срочность интерпретируемости» Амодеи подчеркнул, что современные ИИ-системы — это «чёрный ящик», внутренние процессы которого остаются загадкой даже для разработчиков.

Интерпретируемость, то есть способность понимать, как ИИ принимает решения, становится критически важной, поскольку эти системы всё больше влияют на экономику, технологии и безопасность.

Анодеи сравнивает ИИ-модели с организмами, которые «выращивают», а не создают, из-за чего их поведение сложно предсказать. Он выразил обеспокоенность тем, что без понимания их работы мощные ИИ могут стать опасными, особенно если достигнут уровня искусственного общего интеллекта (AGI) — систем, сравнимых по интеллекту с человеком. Anthropic уже добилась прогресса в отслеживании процессов принятия решений ИИ, но Амодеи призывает OpenAI и Google DeepMind усилить исследования в этой области.

Компания также выступает за «лёгкие» государственные регуляции, включая требования к прозрачности безопасности ИИ, и предлагает ограничить экспорт чипов в Китай, чтобы замедлить глобальную гонку ИИ.