Проблемы управления агентным ИИ в рамках Закона ЕС об ИИ в 2026 году

Источник: AI News Healthcare

Дата публикации: 2026-08

Оригинал: https://www.artificialintelligence-news.com/news/agentic-ais-governance-challenges-under-the-eu-ai-act-in-2026/

EU AI Act | агентный ИИ | информационная безопасность | регулирование | управление рисками

ИИ-агенты открывают перспективы автоматического перемещения данных между системами и инициирования решений, однако в некоторых случаях они могут действовать без четкой записи о том, что, когда и почему они предприняли.

Это создает потенциальную проблему управления (governance), за которую в конечном итоге несут ответственность ИТ-руководители. Если организация не может отследить действия агента и не имеет надлежащего контроля над его полномочиями, руководители не смогут доказать регуляторам, что система работает безопасно или даже законно.

Этот вопрос станет более актуальным с августа этого года, когда вступит в силу **EU AI Act** (Закон ЕС об искусственном интеллекте). Согласно тексту Закона, предусмотрены существенные штрафы за нарушения в управлении ИИ, особенно когда он используется в высокорискованных областях, таких как обработка персональных данных или проведение финансовых операций.

Что ИТ-руководителям необходимо учитывать в ЕС

Для снижения высокого уровня риска можно предпринять несколько шагов, среди которых наиболее важными являются: идентификация агента, ведение подробных журналов (логов), проверка политик, человеческий надзор, возможность быстрого отзыва полномочий, наличие документации от поставщиков и формирование доказательной базы для представления регуляторам.

Существует несколько вариантов, которые лица, принимающие решения, могут рассмотреть для создания записей о деятельности агентных систем. Например, **Python SDK** (набор инструментов для разработки программного обеспечения) **Asqav** может криптографически подписывать каждое действие агента и связывать все записи в неизменяемую цепочку хешей — метод, который больше ассоциируется с технологией блокчейн. Если кто-то или что-то изменит или удалит запись, проверка цепочки завершится ошибкой.

Для групп управления использование подробной, централизованной и, возможно, зашифрованной системы учета всех агентных ИИ является мерой, обеспечивающей данные, выходящие далеко за рамки разрозненных текстовых логов, создаваемых отдельными программными платформами. Независимо от технических деталей того, как создаются и хранятся записи, ИТ-руководителям необходимо точно видеть, где, когда и как агентные экземпляры действуют во всем предприятии.

Многие организации допускают ошибку на этом первом этапе любого документирования автоматизированной деятельности под управлением ИИ. Необходимо вести реестр каждого работающего агента с его уникальной идентификацией, а также записи о его возможностях и предоставленных разрешениях. Этот «список агентских активов» тесно переплетается с требованиями статьи 9 **EU AI Act**, которая гласит:

- **Статья 9**: Для высокорискованных областей управление рисками ИИ должно быть непрерывным, основанным на доказательствах процессом, встроенным в каждый этап развертывания (разработка, подготовка, производство), и находиться под постоянным контролем.

Кроме того, лицам, принимающим решения, необходимо знать о статье 13 Закона:

- Высокорискованные системы ИИ должны быть спроектированы таким образом, чтобы те, кто их развертывает, могли понимать выходные данные системы. Таким образом, система ИИ от стороннего

поставщика должна быть интерпретируемой для пользователей (а не представлять собой непрозрачный массив кода) и должна сопровождаться достаточной документацией для обеспечения её безопасного и законного использования.

Это требование означает, что выбор модели и методов её развертывания являются одновременно техническими и нормативными вопросами.

Применение тормозов

Важно, чтобы любое развертывание агентов предусматривало возможность отзыва операционной роли ИИ, предпочтительно в течение нескольких секунд. Способность быстро отозвать полномочия должна быть частью процессов реагирования на чрезвычайные ситуации. Варианты отзыва должны включать немедленное аннулирование привилегий, немедленное прекращение доступа к **API** (интерфейсу прикладного программирования) и очистку очереди задач.

Наличие человеческого надзора в сочетании с предоставлением достаточного контекста для принятия людьми обоснованных решений означает, что операторы должны иметь возможность отклонить любое предлагаемое действие. Считается недостаточным, если лицо, проверяющее решение, видит только промпт (подсказку) или показатель уверенности (confidence score). Эффективный надзор требует информации о контексте, полномочиях каждого агента и достаточного времени для вмешательства с целью предотвращения ошибок.

Соображения по многоагентным системам

Хотя каждое действие агента должно регистрироваться автоматически и сохраняться, многоагентные процессы особенно сложно отслеживать, так как сбои могут происходить в цепочках агентов. Поэтому важно тестировать политики безопасности во время разработки любой системы, которая намерена использовать нескольких агентов.

Наконец, регулирующие органы могут в любое время потребовать логи и техническую документацию, и они определенно понадобятся им после любого инцидента, о котором им станет известно.

Заключение

Вопрос, который должны рассмотреть ИТ-руководители, планирующие использовать ИИ с конфиденциальными данными или в высокорискованных средах, заключается в том, может ли каждый аспект технологии быть идентифицирован, ограничен политикой, проверен, прерван и объяснен. Если ответ неясен, система управления еще не выстроена.
