

Самореферентная надежность: недостающее свойство безопасности в адаптивном регулировании клинического ИИ

Источник: Frontiers in AI — Medicine

Оригинал: <https://www.frontiersin.org/articles/10.3389/frai.2026.1973339>

безопасность ИИ

валидация моделей

клинические испытания

постмаркетинговый надзор

регулирование ИИ

Адаптивное управление клиническим искусственным интеллектом (ИИ) переходит от статических процессов предварительного одобрения к непрерывному надзору после вывода на рынок, что создает условия, при которых ИИ-сгенерированная самооценка может способствовать обеспечению безопасности после развертывания. Существующие регуляторные подходы, такие как План контроля предопределенных изменений (Predetermined Change Control Plan) и принципы надлежащей практики машинного обучения (Good Machine Learning Practice), охватывают мониторинг, модификацию и дрейф производительности, но не определяют явно надежность ИИ-сгенерированной самооценки как отдельной цели валидации. Новые исследования клинической оценки на основе ИИ показывают существенную вариабельность согласованности между моделями, оценщиками, задачами и критериями оценки. Хотя эти данные в значительной степени основаны на использовании ИИ в качестве оценщика, а не на прямой самооценке, они демонстрируют, почему надежность оценочных суждений не должна выводиться из общей производительности модели.

Мы определяем надежность самооценки (Self-Evaluative Reliability, SER) как согласованность и надежность, с которыми система ИИ оценивает качество, неопределенность или ограничения собственных результатов в рамках predetermined условий оценки. Мы предлагаем рассматривать SER как отдельное свойство безопасности, которое должно проходить независимую валидацию, когда самооценочные суждения участвуют в мониторинге, эскалации, адаптации или продолжении развертывания. Вместо введения новой регуляторной рамки SER расширяет устоявшиеся принципы воспроизводимости и надежности на оценки, сгенерированные ИИ. Установление этой валидации до того, как самооценка станет критически важной для функционирования системы, может укрепить безопасность и прозрачность адаптивного управления клиническим ИИ.

Перевод выполнен: 03.10.2026 | ai4med.ru

Машинный перевод. Рекомендуем сверять с оригиналом при клиническом использовании.