KiloClaw борется с «теневым ИИ» с помощью управления автономными агентами

Источник: AI News Healthcare

Оригинал: https://www.artificialintelligence-news.com/news/kiloclaw-targets-shadow-ai-autonomous-agent-governance/

shadow AI | автономные агенты | безопасность ИИ | корпоративное ПО | управление данными

С запуском **KiloClaw** предприятия получили инструмент для обеспечения управления автономными агентами и контроля «теневого ИИ» (**shadow AI**).

В то время как в течение последнего года компании занимались защитой больших языковых моделей (**LLM**) и формализацией соглашений с поставщиками, разработчики и специалисты по работе с данными начали действовать самостоятельно. Сотрудники обходят официальные процедуры закупок, развертывая автономных агентов на личной инфраструктуре для автоматизации своих ежедневных рабочих процессов.

Эта практика, известная как «Приноси свой собственный ИИ» (**Bring Your Own AI** или **BYOAI**), подвергает конфиденциальные корпоративные данные риску попадания в нерегулируемые внешние среды. Чтобы устранить эту уязвимость, поставщик программного обеспечения Kilo запустил **KiloClaw for Organizations** — платформу корпоративного уровня, созданную для ограничения децентрализованного развертывания агентов и восстановления архитектурного надзора.

Kilo нацелена на проблему отсутствия видимости процессов развертывания агентов. Когда инженеры настраивают автономных агентов для анализа логов ошибок или финансовые аналитики развертывают локальные скрипты

для сверки электронных таблиц, они отдают приоритет немедленной эффективности, а не протоколам безопасности. Эти агенты регулярно получают доступ к корпоративным каналам **Slack**, доскам **Jira** и частным репозиториям кода через личные ключи **API**.

Поскольку эти соединения происходят вне официальной сферы контроля ИТ-отделов, они создают «слепые зоны» для эксфильтрации (вывода) данных и утечки интеллектуальной собственности. **KiloClaw** предоставляет централизованную панель управления для групп безопасности, позволяя выявлять, отслеживать и ограничивать действия этих автономных субъектов, не препятствуя при этом росту их продуктивности.

Невидимая инфраструктура концепции «Принеси своего собственного агента»

Текущий сдвиг напоминает эпоху «Принеси свое собственное устройство» (**Bring Your Own Device** или **BYOD**) начала 2010-х годов, когда сотрудники использовали личные смартфоны для корпоративной электронной почты, что вынудило ИТ-отделы внедрить системы управления мобильными устройствами.

Эквивалент в сфере ИИ несет в себе более высокие риски. Скомпрометированный телефон может раскрыть статический почтовый ящик, но неконтролируемый автономный агент обладает правами на активное выполнение операций. Он читает, записывает, изменяет и удаляет данные на интегрированных платформах со скоростью, которую человек не может воспроизвести.

Эти автономные скрипты также часто полагаются на внешние вычислительные мощности. Сотрудник может запускать агента локально, в то время как агент отправляет корпоративные данные на сторонние серверы логического вывода (**inference servers**) для обработки запросов. Если эти провайдеры используют полученные данные для обучения будущих моделей, предприятие теряет контроль над своей интеллектуальной собственностью.

KiloClaw, со своей стороны, устанавливает защитную границу вокруг этих процессов. Вместо того чтобы игнорировать внешние развертывания, платформа включает их в реестр, где специалисты по комплаенсу могут проводить аудит поведения и потоков данных.

Управление идентификацией и доступом для автономных ИИ-агентов

Управление автономными системами требует иной технической архитектуры, чем управление человеческим персоналом. Традиционные системы управления идентификацией и доступом (**Identity and Access Management** или **IAM**) предназначены для учетных данных человека или статического взаимодействия «приложение-приложение».

Однако автономные агенты динамичны. Агенты последовательно связывают задачи, формулируя новые запросы на основе результатов предыдущих действий. Агент может запросить доступ к базе данных планирования ресурсов предприятия (**ERP**) на середине выполнения задачи, и стандартное защитное программное обеспечение с трудом сможет определить, является ли это враждебным поведением или легитимной операцией.

KiloClaw рассматривает агентов как отдельные сущности, требующие ограничительных, ограниченных по времени областей полномочий. Вместо того чтобы позволять разработчикам вставлять постоянные высокоуровневые ключи **API** в экспериментальные модели, **KiloClaw** выдает краткосрочные, узко определенные токены доступа.

Если агент, предназначенный для обобщения еженедельных маркетинговых писем, попытается скачать базу данных клиентов, платформа обнаружит нарушение области полномочий и аннулирует доступ. Такое сдерживание ограничивает «радиус поражения» внутри корпоративной сети в случае, если модель с открытым исходным кодом поведет себя непредсказуемо.

Как такие инструменты, как KiloClaw, балансируют между скоростью и комплаенсом

Введение тотального запрета на кастомные инструменты автоматизации редко приносит успех; это заставляет подобные действия уходить в тень, побуждая инженеров маскировать трафик и скрывать рабочие процессы. Платформы, подобные **KiloClaw**, стремятся создать санкционированную среду, где сотрудники могут безопасно регистрировать свои инструменты.

Чтобы эта структура управления работала, ИТ-руководителям необходимо уделять приоритетное внимание интеграции. **KiloClaw** напрямую подключается к конвейерам непрерывной интеграции и развертывания (**CI/**

CD), которые уже используют команды разработчиков. Автоматизируя проверки безопасности и предоставление разрешений, группы безопасности устраняют трение, которое заставляет сотрудников обходить правила.

Предприятия могут устанавливать базовые шаблоны, детализирующие, какие данные внешние модели могут обрабатывать, позволяя работникам развертывать агентов в рамках заранее одобренных границ. Это поддерживает соблюдение нормативных требований без ущерба для автоматизации рабочих процессов.

Развитие инструментов управления «теневым ИИ» указывает на новый этап алгоритмического регулирования. Первоначальная реакция корпораций на генеративные модели была сосредоточена на политиках допустимого использования текстовых чат-ботов. Теперь фокус смещается в сторону оркестрации, сдерживания и подотчетности систем между собой. Регуляторы по всему миру также изучают, как компании контролируют автоматизированные системы, превращая проверяемый надзор в юридическое обязательство.

По мере размножения цифровых агентов в корпоративных сетях концепция «межсетевого экрана для агентов» (**Agent Firewall**) становится стандартной статьей ИТ-бюджета. Платформы, которые выстраивают связи между намерениями человека, машинным исполнением и корпоративными данными, составят основу будущих операций по обеспечению безопасности.

Выход **KiloClaw** в сегмент организационного управления подчеркивает меняющуюся реальность для высшего руководства: непосредственная угроза включает в себя добросовестных сотрудников, передающих ключи от сети нерегулируемым машинам. Установление структурного авторитета над этими нечеловеческими субъектами необходимо для безопасного использования их потенциала.
