

Чек-лист безопасного внедрения long-horizon ИИ-агента

PDF-файл для [downloads/checklist.pdf](#): краткая памятка для команд, которые дают ИИ-агенту доступ к инструментам, коду, данным или инфраструктуре. Чек-лист основан только на практических выводах статьи и не добавляет внешних фактов.

Права и среда выполнения

- Запускайте агента сначала в sandbox или отдельном экспериментальном окружении.
- Используйте read-only режим по умолчанию для внутренних данных, репозиториях и инфраструктурных ресурсов.
- Выдавайте scoped credentials вместо широких постоянных доступов.
- Применяйте короткоживущие токены там, где агенту действительно нужен доступ к внешним системам.
- Не давайте production-права без отдельного явного подтверждения.

Операции с риском

- Требуйте human-in-the-loop для удаления файлов, изменения баз данных, деплоя и публикации кода во внешний репозиторий.
- Разделяйте разрешения на чтение, запись, запуск тестов, сетевые запросы и публикацию результатов.
- Останавливайте сессию или запрашивайте подтверждение, если агент пытается обойти ограничение среды.
- Показывайте пользователю, зачем агент просит расширить доступ и какие последствия возможны.

Наблюдаемость и проверки

- Логируйте всю траекторию, а не только финальный ответ.
- Фиксируйте команды, tool calls, файлы, API, запросы прав, ошибки и изменения плана.
- Превращайте каждый реальный сбой в incident-derived evaluation перед расширением доступа.
- Не считайте высокий benchmark по long-running workflows доказательством безопасной автономии.
- Расширяйте полномочия staged rollout: offline-задачи, read-only режим, ограниченные действия с подтверждением, затем более широкие сценарии.