



**Учебный план дополнительной профессиональной
программы повышения квалификации:
«Безопасность систем искусственного интеллекта:
угрозы, защита и соответствие требованиям»**

№ п/п	Наименование темы	Описание темы	Всего, час	В том числе			
				Лекции	Практ.	Сам. работа	Итог. аттест.
1.	ИИ-системы: поверхность атаки и требования	Демонстрация: письмо со спрятанной инструкцией заставляет почтового агента переслать вложения наружу; фильтр опасных слов не помогает. ИИ-система как объект защиты: модель, данные, инструменты, память, права, журналы. Классы угроз и международные ориентиры (OWASP Top 10 для приложений на языковых моделях). Раздел об ИИ приказа ФСТЭК № 117 и методический документ ФСТЭК: системы с ИИ как отдельный объект оценки. От эффектных атак курс сразу переходит к модели угроз: активы, границы доверия, поверхность атаки, полномочия, цепочка поставок (модели, плагины, серверы инструментов), память, базы знаний, учётные записи, журналы. Учебная лаборатория близка к корпоративной: пользователь → агент → база знаний → инструменты → база данных и API → журналы. Демонстрация: Письмо со спрятанной инструкцией заставляет почтового агента переслать вложения на внешний адрес. Никакого взлома, только текст. Инвентаризация ИИ в своей организации, включая «теневой» ИИ (личные чаты	4	3	-	1	-



		сотрудников).					
2.	Требования на практике, утечки и отравление данных	Отравление базы знаний: подложенный документ с ложным регламентом, на который агент уверенно ссылается; поиск источника по журналу. Демонстрация: Агент поддержки «случайно» раскрывает клиенту данные другого клиента. В базу знаний подложен ложный регламент, и агент уверенно на него ссылается. Требования как проверяемые пункты по цепочке «требование → что проверить → какое свидетельство собрать → какой тест провести → что записать в отчёт»; воспроизведение утечки данных через ответ модели (инструкция «не раскрывай» не работает при хитром вопросе). Результат: перечень проверок и протокол атаки «утечка через ответ». Воспроизведение утечки в учебной системе, ее поиск по журналу и трассе, поиск причины и источника ложного ответа. Задание: составить перечень применимых требований и проверок для ИИ-системы своей организации.	4	1	2	1	-
3.	Злоупотребление ресурсами и атаки на инструменты	Сторонние серверы инструментов (например, по протоколу MCR) как новый участник доверия; описание инструмента может содержать внедрённую инструкцию. Демонстрация: Агент с правом записи в таблицу получает команду из документа и меняет данные. Сторонний инструмент с «отравленным» описанием меняет поведение	4	1	1	2	-



		агента. Атаки на инструменты агента с избыточными полномочиями; подтверждение человеком не спасает, если не видно, что меняется. Результат: реестр инструментов и полномочий учебной системы. Поиск инструментов с избыточными полномочиями, составление правил допуска сторонних инструментов. Задание: разобрать один публичный инцидент с ИИ-системой и предложить лимиты ресурсов для учебной системы.					
4.	Память агента и красная команда	Память агента как канал атаки: вредная инструкция, сохранённая в долговременной памяти, срабатывает в чужом сеансе; что можно и нельзя хранить в памяти агента. Демонстрация: Вредная инструкция, сохранённая в долговременной памяти агента, срабатывает в чужом сеансе. Затем красная команда: группа разбирает первые найденные уязвимости. Красная команда: составление плана атак (15–20 сценариев по классам угроз) и проведение атак в учебной системе с фиксацией доказательств. Результат: план и журнал атак. Составление плана из 15–20 атак, их проведение в учебной системе с фиксацией доказательств. Задание: составить реестр инструментов, полномочий и данных одной ИИ-системы.	4	1	2	1	-
5.	Отчёт о проверке и защитные меры	Многослойная защита: разделение данных и инструкций, минимальные	4	-	2	2	-



		полномочия, изоляция инструментов, фильтры входа и выхода, ограничение исходящих данных, подтверждение и лимиты как последние рубежи; журналы, признаки атаки, оповещения и аварийная остановка. Результат: план защитных мер и правила мониторинга. Выбрать меры защиты под найденные уязвимости, настроить оповещение и повторным прогоном доказать результат. Задание: оформить черновик отчёта о проверке со статистикой воспроизводимости атак.					
6.	Изменения системы и итоговый отчёт	Регламент изменений ИИ-системы: смена модели, инструкции или инструмента как изменение системы; регрессионный прогон атак после обновления модели. Результат: регламент изменений. Составление регламента изменений ИИ-системы и итогового отчёта о проверке.	2	-	1	1	-
7.	Итоговая аттестация	Зачет	2	-	-	-	2
	ИТОГО		24	6	8	8	2