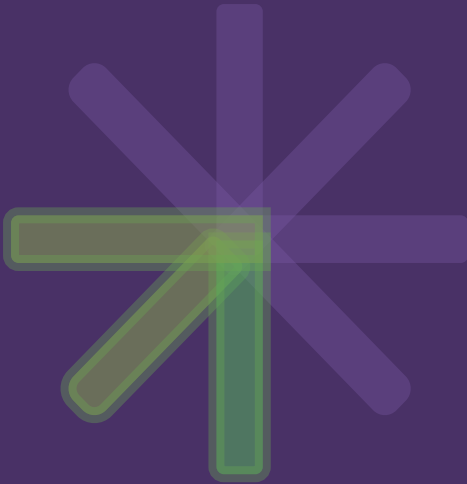


دورة حوكمة وكلاء الذكاء الاصطناعي: الضوابط الوقائية والصلاحيات والإشراف البشري





دورة حوكمة وكلاء الذكاء الاصطناعي: الضوابط الوقائية والصلاحيات والإشراف البشري

ملخص البرنامج

حوكمة وكلاء الذكاء الاصطناعي دورة تدريبية تمتد 5 أيام بمستوى متقدم، موجهة إلى القيادات العليا في التقنية والمخاطر والتدقيق، ويخرج منها كل مشارك بإطار لضبط وكلاء الذكاء الاصطناعي في جهته. فالوكلاء باتوا يخططون للمهام ويستدعون الأدوات ويعدلون السجلات بتدخل بشري محدود، فقد ينتقل أمر خاطئ واحد عبر الأنظمة قبل أن ينتبه إليه أحد. وينتقل البرنامج من مستويات الاستقلالية والتهديدات الخاصة بالوكلاء إلى تصميم الصلاحيات والضوابط الوقائية والإشراف والتأكيد. تقدم هذا البرنامج في حوكمة وكلاء الذكاء الاصطناعي المستقلين مارس للتدريب المملكة المتحدة.

المبررات المؤسسية

يعد الوكلاء بخدمة أسرع وتكلفة معالجة أقل، لكن كثيرا من المؤسسات تطلقهم بصلاحيات موظف موثوق ورقابة لا تتجاوز رقابة روبوت محادثة بسيط. ومن دون حدود واضحة للاستقلالية، ومالكين مساءلين، وسجلات تدقيق، لا يستطيع مجلس الإدارة معرفة أي وكيل تصرف، وبأي تفويض، وبأي بيانات.

- الانكشاف للخسائر التشغيلية: ينخفض حين تحدد صلاحيات الوكلاء وسقوف الإنفاق ونقاط التصعيد قبل الإطلاق.
- زمن اكتشاف الحوادث واحتوائها: يقصر حين تسجل تصرفات الوكلاء وتراقب ويمكن إيقافها.
- الجاهزية للتأكيد: تتحسن حين تربط ضوابط الوكلاء بأطر معترف بها لإدارة مخاطر الذكاء الاصطناعي.

المبادئ والأساس المعرفي

- إطار إدارة مخاطر الذكاء الاصطناعي NIST AI RMF: تنظم وظائفه الأربع في الحوكمة والتحديد والقياس والإدارة إطار الضبط الذي يبنى عبر الجلسات من 1 إلى 4.
- معيار نظام إدارة الذكاء الاصطناعي ISO/IEC 42001:2023: توجه متطلباته الأدوار والسياسات والتحسين المستمر للوكلاء في الجلسة 4.
- قائمة OWASP Top 10 for LLM Applications لمخاطر النماذج اللغوية: يغذي تناولها لحقن الأوامر والصلاحيات المفرطة تشخيص التهديدات في الجلسة 2.
- قائمة OWASP Top 10 for Agentic Applications لمخاطر التطبيقات الوكيلية: تنظم مخاطرها، ومنها إساءة استخدام الأدوات واستغلال الصلاحيات والإخفاقات المتتالية، الاختبار في الجلستين 2 و3.



المشاركون المستهدفون

- القيادات التي تعتمد إدخال وكلاء الذكاء الاصطناعي في العمليات وخدمات العملاء.
- التنفيذيون المسؤولون عن استراتيجية التقنية أو البيانات أو التحول الرقمي في المؤسسة.
- رؤساء المخاطر وأمن المعلومات الذين يضعون حدود القرارات المؤتمتة وضوابطها.
- قيادات التدقيق الداخلي المطالبة بتقديم التأكيد على العمليات المدعومة بالذكاء الاصطناعي.

يخدم البرنامج المصارف والتأمين والخدمات الحكومية والاتصالات والطاقة.

مخرجات التعلم

1. سيكون المشاركون قادرين على تصنيف حالات استخدام وكلاء الذكاء الاصطناعي حسب مستوى الاستقلالية والأثر في الأعمال.
2. سيكون المشاركون قادرين على تشخيص التهديدات الخاصة بالوكلاء، ومنها حقن الأوامر وإساءة استخدام الأدوات والصلاحيات المفرطة.
3. سيكون المشاركون قادرين على تصميم هوية الوكلاء وصلاحياتهم وفق مبدأ الحد الأدنى من الامتيازات وضوابطهم الوقائية.
4. سيكون المشاركون قادرين على تحديد نقاط الإشراف البشري وآليات الإيقاف وقواعد التصعيد.
5. سيكون المشاركون قادرين على تقييم المراقبة والتسجيل والاختبار والتأكيد على منصات الوكلاء الخارجية.
6. سيكون المشاركون قادرين على بناء إطار لضبط وكلاء الذكاء الاصطناعي يتوافق مع NIST AI RMF و ISO/IEC 42001:2023.

هيكل البرنامج

اليوم 1: حوكمة وكلاء الذكاء الاصطناعي ومستويات الاستقلالية

ما الذي يجعل حوكمة الوكيل مختلفة عن حوكمة أنظمة الذكاء الاصطناعي الأخرى؟

- بنية الوكيل: التخطيط والذاكرة والأدوات والتصرفات
- مستويات الاستقلالية ورسم التفويض
- حصر حالات الاستخدام وتصنيف الأثر
- تطبيق وظيفة الحوكمة في NIST AI RMF على الوكلاء
- المالكون المساءلون وسجلات الوكلاء



اليوم 2: تشخيص تهديدات الوكلاء وأنماط إخفاقهم

كيف يمكن تضليل الوكيل أو إساءة استخدامه أو تسببه في ضرر متسلسل؟

- حقن الأوامر وهجمات التعليمات غير المباشرة
- الصلاحيات المفرطة وسيناريوهات إساءة استخدام الأدوات
- تسميم الذاكرة ومسارات تسرب البيانات
- الأخطاء المتتالية في سير العمل متعدد الوكلاء
- نموذج تهديدات الوكلاء باستخدام OWASP Top 10 for Agentic Applications

اليوم 3: تصميم الصلاحيات والضوابط الوقائية والإشراف

ما الضوابط التي تبقى الوكيل ضمن حدود التفويض المقصود؟

- هوية الوكيل وبيانات اعتماده والوصول بالحد الأدنى من الامتيازات
- ضوابط المدخلات والمخرجات والتصرفات
- نقاط الإشراف البشري وعتبات الموافقة
- مفاتيح الإيقاف الطارئ وحدود المعدل وتصميم التراجع
- اختبار الفريق الأحمر وخطط الاختبار قبل الإطلاق

اليوم 4: المراقبة والتأكيد ومنصات الأطراف الخارجية

كيف يوثق سلوك الوكيل ويحقق ويدار مع الوقت؟

- تسجيل التصرفات وسجلات التدقيق وإمكانية التتبع
- مؤشرات المراقبة أثناء التشغيل وتنبيهات الشذوذ
- الاستجابة لحوادث الوكلاء والمراجعة اللاحقة
- العناية الواجبة تجاه منصات الوكلاء الخارجية وشروط التعاقد
- الدمج في نظام الإدارة وفق ISO/IEC 42001:2023

اليوم 5: تطبيق حوكمة الوكلاء على المؤسسة

أي الوكلاء يمضي إلى التشغيل، وبأي ضوابط، وبأي أدلة؟



- قرارات الاستقلالية لحالات الاستخدام من الجلسة 1
- اختيار الضوابط مقابل نموذج التهديدات من الجلسة 2
- خطة التأكيد باستخدام مؤشرات المراقبة من الجلسة 4
- التقارير إلى مجلس الإدارة ومستوى تقبل مخاطر وكلاء الذكاء الاصطناعي
- إنجاز إطار ضبط وكلاء الذكاء الاصطناعي

المنهجية

يعتمد البرنامج على تمارين منظمة حول وكلاء نموذجيين في خدمة العملاء والعمليات المالية والدعم التقني. وتبدأ كل جلسة بعرض موجز، ثم تحليل جماعي، ثم اختبار الضوابط المقترحة أمام سيناريوهات الهجوم والإخفاق.

- نشاط مقترح: رسم صلاحيات وكيل مشتريات وإزالة الوصول غير الضروري.
- نشاط مقترح: تتبع حقن أوامر غير مباشر عبر وكيل لمعالجة المستندات.
- نشاط مقترح: تصميم عتبات الموافقة لوكيل يصدر المبالغ المستردة.

المخرج الختامي

يعد كل مشارك إطارا لضبط وكلاء الذكاء الاصطناعي في جهته، يتضمن سجلا للوكلاء، وتصنيفات الاستقلالية، ونموذجا للتهديدات، ومعايير الصلاحيات والضوابط الوقائية، ونقاط الإشراف، وخطة للتأكيد. ويمكن اعتماده سياسة مؤسسية أو استخدامه للبت في إطلاق أي وكيل.

الأسئلة الشائعة

هل يناسب تدريب حوكمة وكلاء الذكاء الاصطناعي القيادات التي لا تملك خلفية تقنية؟

نعم، فهو بمستوى متقدم للتنفيذيين الذين يعتمدون وكلاء الذكاء الاصطناعي أو يقدمون التأكيد عليهم، ويفترض معرفة بحوكمة المخاطر أو التقنية.

ما الأطر التي يقوم عليها تدريب حوكمة وكلاء الذكاء الاصطناعي؟

يقوم على NIST AI RMF، والمعيار ISO/IEC 42001:2023، وقائمة OWASP Top 10 for LLM Applications، وقائمة OWASP Top 10 for Agentic.



كيف يستخدم إطار ضبط وكلاء الذكاء الاصطناعي في العمل؟

تستخدمه أساسا لسياسة الوكلاء، وللموافقة على إطلاق مقترح أو رفضه، ولتخطيط أعمال التأكيد.

بم تختلف حوكمة وكلاء الذكاء الاصطناعي عن برامج أخلاقيات الذكاء الاصطناعي؟

تركز على الضبط التشغيلي للوكلاء الذين يتصرفون في الأنظمة والبيانات، لا على المبادئ الأخلاقية للذكاء الاصطناعي عموما.