CONCEPT

تقييم مخاطر الذكاء الاصطناعي والتحقق من النماذج وضمان موثوقيتها

07 - 11 Jun 2027

دبي - فندق خمس نجوم في وسط الأعمال

تقييم مخاطر الذكاء الاصطناعي والتحقق من النماذج وضمان موثوقيتها

المرجع: 1061_4785 التاريخ: 07 - 11 Jun 2027 الموقع: دبي - فندق خمس نجوم في وسط الأعمال الرسوم: SAR 19500

العمق التقني: ممارس · أسلوب التطبيق: دراسة حالة

مقدمة

تدخل أنظمة الذكاء الاصطناعي إلى قرارات الائتمان وخدمة العملاء والتوظيف وكشف الاحتيال والعمليات بوتيرة تفوق قدرة المؤسسات على تقييمها. فالنماذج تشترى أو تطور دون تسجيلها في قائمة الجرد، وتختبر من حيث الدقة وحدها، ثم تترك دون مراقبة مع انحراف البيانات، بينما يضيف الذكاء الاصطناعي التوليدي مخاطر المخرجات الملفقة وحقن الأوامر وتسرب البيانات. تمكن هذه الدورة من كور كونسبت المختصين في المخاطر والتأكيد من تقييم مخاطر الذكاء الاصطناعي واختبار النماذج وفق المعايير المعترف بها، من تقييم الأثر إلى التحقق والمراقبة المستمرة. ويعمل المشاركون على حالات من قطاعات متعددة، ويخرجون بحزمة مخاطر الذكاء الاصطناعي والتأكيد لنظام قائم في مؤسساتهم.

أهداف الدورة

- بناء قائمة جرد أنظمة الذكاء الاصطناعي وتحديثها مع تصنيف مستويات المخاطر بما يتوافق مع IEC/ISO 42001 وإطار NIST AI RMF
- إجراء تقييمات مخاطر الذكاء الاصطناعي وأثره باستخدام IEC/ISO 23894 وIEC/ISO 42005
- تقييم أدلة أداء النموذج وتحيزه ومتانته وقابليته للتفسير مقابل معايير قبول محددة
- تقييم مخاطر الذكاء الاصطناعي التوليدي والنماذج المقدمة من أطراف خارجية، ومنها حقن الأوامر وتسرب البيانات والمخرجات الملفقة
- تصميم آليات المراقبة ورصد الانحراف ومحفزات إعادة التحقق للنماذج العاملة
- إعداد حزمة مخاطر الذكاء الاصطناعي والتأكيد مع رأي تأكيدي قائم على الأدلة

الفئة المستهدفة

- مديرو المخاطر المسؤولون عن إدراج الذكاء الاصطناعي في ملف المخاطر المؤسسية
- مديرو التحقق من النماذج ومخاطر النماذج الذين يختبرون النماذج قبل التشغيل وبعده
- مديرو المراجعة الداخلية الذين يخططون لمهام التأكيد على أنظمة الذكاء الاصطناعي
- قادة علوم البيانات وتعلم الآلة المسؤولون عن توثيق النماذج وضوابطها
- مديرو الامتثال وحماية الخصوصية الذين يقيمون أثر الذكاء الاصطناعي على الأفراد
- مديرو التقنية والمشتريات الذين يقيمون منتجات الذكاء الاصطناعي وخدماته المقدمة من أطراف خارجية

محاور الدورة

اليوم الأول: بيئة مخاطر الذكاء الاصطناعي وجرد الوضع الراهن

- أنواع أنظمة الذكاء الاصطناعي ومراحل دورة حياتها وفق IEC/ISO 5338 مواصفة عمليات دورة حياة أنظمة الذكاء الاصطناعي
- مراجعة حوادث الذكاء الاصطناعي: التحيز والهلوسة وإخفاقات الأتمتة
- مبادئ OECD للذكاء الاصطناعي وخصائص الذكاء الاصطناعي الجدير بالثقة
- بناء قائمة جرد أنظمة الذكاء الاصطناعي وتصنيف حالات الاستخدام حسب المخاطر
- تقييم فجوات الوضع الراهن لحوكمة الذكاء الاصطناعي

اليوم الثاني: معايير مخاطر الذكاء الاصطناعي وأنظمة إدارته

- متطلبات نظام إدارة الذكاء الاصطناعي في IEC/ISO 42001 وضوابط الملحق A
- إرشادات إدارة مخاطر الذكاء الاصطناعي في IEC/ISO 23894 وربطها بالمواصفة ISO 31000
- الوظائف الأساسية في إطار RMF AI NIST لإدارة مخاطر الذكاء الاصطناعي: Govern وMap وMeasure وManage
- فئات المخاطر في ملف الذكاء الاصطناعي التوليدي AI NIST 600-1
- نموذج جودة أنظمة الذكاء الاصطناعي في IEC/ISO 25059

اليوم الثالث: تقييم المخاطر والتحقق من النماذج

- تقييم أثر نظام الذكاء الاصطناعي باستخدام IEC/ISO 42005
- مراجعة بطاقة النموذج Card Model وصحيفة البيانات Datasheet للتحقق من اكتمال التوثيق
- التحقق من الأداء: اختبار البيانات المحجوزة ومصفوفة الالتباس واختيار عتبة القرار
- قياس التحيز باستخدام مقاييس العدالة في IEC/ISO 24027 TR
- مراجعة أدلة قابلية التفسير من مخرجات SHAP وLIME

اليوم الرابع: المخاطر المتقدمة والضوابط والمراقبة

- مراجعة تهديدات الذكاء الاصطناعي التوليدي بقائمة OWASP Top 10 لتطبيقات النماذج اللغوية الكبيرة
- رسم خريطة الهجمات العدائية باستخدام ATLAS MITRE
- استبيان العناية الواجبة لنماذج الأطراف الخارجية والنماذج التأسيسية
- مراقبة انحراف البيانات وانحراف المفهوم باستخدام مؤشر استقرار المجتمع PSI
- تصميم الإشراف البشري ومحفزات إعادة التحقق من النموذج

اليوم الخامس: دراسات الحالة وحزمة مخاطر الذكاء الاصطناعي والتأكيد

- دراسة حالة نموذج التصنيف الائتماني: نتائج التحقق ومساءلتها
- دراسة حالة روبوت محادثة لخدمة العملاء: مراجعة نتائج اختبار الفريق الأحمر للذكاء الاصطناعي التوليدي
- سجل مخاطر الذكاء الاصطناعي وربط الضوابط لنظام من مؤسسة المشارك
- إعداد مسودة حزمة مخاطر الذكاء الاصطناعي والتأكيد
- الدفاع عن الرأي التأكيدي أمام لجنة مخاطر نماذج من الزملاء

المهارات المكتسبة

- تصنيف مخاطر حالات استخدام الذكاء الاصطناعي
- التحقق من النماذج
- تقييم التحيز الخوارزمي
- مراجعة قابلية التفسير
- تحليل تهديدات الذكاء الاصطناعي التوليدي
- ربط الضوابط بأنظمة الذكاء الاصطناعي
- العناية الواجبة لموردي الذكاء الاصطناعي
- إعداد تقارير التأكيد على الذكاء الاصطناعي

لماذا تحضر هذه الدورة

- تعود إلى عملك بحزمة مخاطر الذكاء الاصطناعي والتأكيد لنظام فعلي، وقد دافعت عنها أمام لجنة من الزملاء
- تطرح على فرق علوم البيانات والموردين الأسئلة التي تكشف ضعف التحقق أو غياب الضوابط
- تقدم للإدارة العليا ومجلس الإدارة صورة واضحة عن أنظمة الذكاء الاصطناعي الأعلى خطورة وأسباب ذلك
- تقارن ممارسات التأكيد على الذكاء الاصطناعي مع مختصين في المخاطر والتقنية من قطاعات ودول مختلفة

الخاتمة

لا يمكن الوثوق بالذكاء الاصطناعي إلا حين تقيم مخاطره قبل التشغيل ويختبر سلوكه ويراقب بعده. تنتقل هذه الدورة من بيئة مخاطر الذكاء الاصطناعي والمعايير الحاكمة، مرورا بتقييم الأثر والتحقق من النماذج، وصولا إلى تهديدات الذكاء الاصطناعي التوليدي ونماذج الأطراف الخارجية والانحراف. ويحول اليوم الأخير هذه المعارف إلى حزمة مخاطر الذكاء الاصطناعي والتأكيد يعود بها المشارك إلى مؤسسته، فيمتلك منهجية قابلة للتكرار ومستندة إلى المعايير لكل نظام ذكاء اصطناعي يكلف بتقييمه.