CONCEPT

تقييم مخاطر الذكاء الاصطناعي والتحقق من النماذج وتأكيد أنظمة الذكاء الاصطناعي



تقييم مخاطر الذكاء الاصطناعي والتحقق من النماذج وتأكيد أنظمة الذكاء الاصطناعي

نظرة عامة على الدورة

تجاوز تسارع المؤسسات في تبني تقنيات الذكاء الاصطناعي لاتخاذ قرارات الائتمان الآلية، وكشف الاحتيال، والتفاعل مع العملاء، والعمليات التشغيلية، وتيرة الممارسات المعتادة لإدارة المخاطر. وتعمل العديد من النماذج المنشورة حاليا دون توثيق منهجي في سجلات مركزية، وتقتصر اختبارات أدائها على مقاييس الدقة الأساسية فقط، مما يعرضها للتراجع السريع عند اختلاف توزيعات بيانات التشغيل الفعلي. كما تفرض بنى الذكاء الاصطناعي التوليدي مخاطر تشغيلية إضافية تشمل المخرجات غير الدقيقة واختراق الأوامر وتسرب الملكية الفكرية. يقدم هذا البرنامج أساليب تقنية متخصصة لخبراء المخاطر، وقادة التحقق، وفرق التأكيد المستقل لتقييم خطوط أنابيب تعلم الآلة بالاستناد إلى المعايير القياسية الدولية المعترف بها عبر مراحل التطوير والإنتاج. وترشد السيناريوهات الواقعية المشاركين خلال تصنيف الأثر، والتحقق الفني من الأداء، ومعالجة التحيز، وإجراءات مراقبة الانجراف، وصولا إلى إعداد ملف تقييم منهجي وقابل للدفاع عنه صممته كور كونسبت لمساندة لجان مخاطر النماذج.

أهداف الدورة

- بناء سجل تشغيلي لأنظمة الذكاء الاصطناعي وإدارته وفق مستويات مخاطر قابلة للدفاع عنها بالاستناد إلى مبادئ دورة الحياة في ISO/IEC 42001 وISO/IEC 5338.
- تنفيذ عمليات تقييم المخاطر والأثر المنهجية بتطبيق المراحل التحليلية المحددة في معايير ISO/IEC 23894 وISO/IEC 42005.
- التحقق من أداء نماذج تعلم الآلة، والعدالة الخوارزمية، والمتانة الفنية، وقابلية التفسير وفق عتبات قبول إحصائية محددة.
- تحليل ثغرات البنى التوليدية وتخفيف مخاطر حقن الأوامر وتسرب البيانات الحساسة والهلوسة المعرفية.
- بناء بروتوكولات تشغيلية لاكتشاف الانجراف في بيئة الإنتاج، وتتبع مقاييس الاستقرار، وتحديد عتبات إعادة التحقق التجريبية.
- إعداد ملف تقييم مستند إلى الأدلة لمساندة المراجعة المستقلة وتدقيق لجان المخاطر.

الفئة المستهدفة

- محللو ومديرو المخاطر المؤسسية المسؤولون عن دمج إشراف تعلم الآلة ضمن سجلات المخاطر التشغيلية
- أخصائيو التحقق من النماذج المعنيون بالفحص التقني والطعن المنهجي في الخوارزميات
- ممارسو التأكيد والمراجعة الداخلية المكلفون بالتقييم المستقل لأنظمة اتخاذ القرار الآلية
- مهندسو تعلم الآلة وقادة علوم البيانات المعنيون بإعداد بطاقات النماذج ووثائق التحقق الفني
- مستشارو الامتثال ومسؤولو الخصوصية المعنيون بتقييم أثر الخوارزميات على المستخدمين
- مسؤولو المشتريات التقنية المكلفون بمراجعة الأنظمة الخوارزمية الخارجية ونماذج الأساس التابعة لأطراف ثالثة

محاور الدورة

اليوم 1: مشهد مخاطر الذكاء الاصطناعي وحصر الأنظمة

- تصنيفات الأنظمة الخوارزمية ومراحل دورة الحياة ضمن العمليات الهندسية لمعيار ISO/IEC 5338
- تحليل الإخفاقات المنهجية بما يشمل توالد التحيز، والهلوسة، والتسلسل التلقائي للأخطاء
- مبادئ الأنظمة الجديرة بالثقة وأسس الحوكمة التشغيلية
- بناء سجل أنظمة الذكاء الاصطناعي، ومتطلبات البيانات الوصفية، وتصنيف مخاطر حالات الاستخدام
- تشخيص الحوكمة الداخلية وتحليل الفجوات الإجرائية

اليوم 2: معايير الذكاء الاصطناعي وأطر حوكمة المخاطر

- متطلبات نظام الإدارة وضوابط الملحق A التشغيلية في معيار ISO/IEC 42001
- إرشادات إدارة المخاطر ومسارات التقييم المعتمدة في ISO/IEC 23894
- وظائف الحوكمة الجوهرية التي تشمل التحديد والقياس والمعالجة التشغيلية
- ملفات مخاطر البنى التوليدية وتصنيفات الثغرات المتخصصة
- خصائص جودة الأنظمة ومعايير التحقق المحددة في معيار ISO/IEC 25059

اليوم 3: منهجيات تقييم الأثر والتحقق من النماذج

- تقييمات أثر الأنظمة بالاعتماد على المنهجيات المحددة في معيار ISO/IEC 42005
- تدقيق وثائق المنظومة عبر بطاقات مواصفات النماذج وسجلات منشأ البيانات
- تقنيات التحقق التجريبي باستخدام عينات الاختبار المستقلة ومصفوفات الالتباس وضبط عتبات التصنيف
- تحليل الأثر المتباين واختبار التكافؤ الديموغرافي استنادا إلى إرشادات ISO/IEC TR 24027
- التحقق من أهمية الخصائص وقابلية التفسير المحلية باستخدام أطر SHAP وLIME

اليوم 4: الثغرات المتقدمة وضوابط التهديدات والمراقبة المستمرة

- تخفيف الثغرات استنادا إلى تصنيف OWASP Top 10 لتطبيقات النماذج اللغوية الكبيرة
- تصنيف الهجمات العدائية واختبارات المراوغة بالاستناد إلى مصفوفة MITRE ATLAS
- استبيانات تأكيد النماذج الخارجية والتحقق من سلاسل إمداد الموردين
- تتبع الانجراف الإحصائي باستخدام مؤشر استقرار المجتمع ومقاييس التباعد الإحصائي
- مسارات الإشراف البشري وتصميم عتبات محفزات إعادة التحقق التجريبية

اليوم 5: التقييم العملي وبناء ملف التأكيد

- تقييم نموذج تسجيل الائتمان للأفراد وفحص مقاييس التحقق واختبار العتبات
- اختبار المساعد الحواري لخدمة العملاء وتقييم أمان الأوامر وحدود المخرجات
- هيكلة سجل المخاطر وتعيين الضوابط الداخلية لنموذج مؤسسي قيد التشغيل
- تجميع ملف التقييم الكامل ووثائق التحقق من النماذج وتأكيدها
- محاكاة جلسة دفاع أمام لجنة مخاطر النماذج لعرض نتائج التأكيد المستقل

المهارات المكتسبة

- تصنيف المخاطر الخوارزمية
- اختبار التحقق من النماذج
- تدقيق العدالة والتحيز الخوارزمي
- تحليل قابلية تفسير النماذج
- تخفيف تهديدات الأنظمة التوليدية
- ربط ضوابط مخاطر الذكاء الاصطناعي
- العناية الواجبة بموردي الخوارزميات
- إعداد تقارير تأكيد النماذج

لماذا تحضر هذه الدورة

- اكتساب بروتوكولات اختبار منهجية للتحقق من كفاءة نماذج تعلم الآلة تحت مختلف ظروف التشغيل.
- صياغة أسئلة فنية دقيقة لكشف الافتراضات غير المختبرة في خطوط أنابيب علوم البيانات ونماذج الموردين.
- تزويد الإدارة التنفيذية وجهات الرقابة بتقييمات هيكلية محكمة للأنظمة الخوارزمية ذات الحساسية التشغيلية العالية.
- الخروج بملف تقييم متكامل للنماذج تمت مراجعته وتدقيقه عبر محاكاة عملية لفرق المراجعة المستقلة.

الخاتمة

يتطلب تحقيق قيمة مستدامة من أنظمة تعلم الآلة رقابة منضبطة تسبق إطلاق النماذج وتستمر طوال تشغيلها الفعلي. يتناول هذا البرنامج المسار الكامل للمخاطر الخوارزمية، انطلاقا من حصر الأنظمة وتطبيق المعايير الحاكمة، مرورا بمقاييس التحقق التجريبي وضوابط ثغرات الأنظمة التوليدية، وصولا إلى كشف الانجراف. ومن خلال بناء الأدلة عبر دراسات حالات تطبيقية، يكتسب الممارسون إطارا متوافقا مع المعايير لتقييم ونقد وتأكيد حلول الذكاء الاصطناعي باحترافية واستقلالية عبر مؤسساتهم.