JUPITER PRIME

Consulting • Training • Advisory

جوبيتر برايم للاستشارات والتدريب

كتيّب الدورة

دورة الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي

info@jupiterprime.uk

jupiterprime.uk

مسجّلة في المملكة المتحدة والمملكة العربية السعودية

دورة الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي

الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي دورة تدريبية مدتها أسبوع لمحللي الاحتيال ومسؤولي الأمن ومديري المخاطر الذين يحافظون على موثوقية نماذج الكشف حين يختبرها المجرمون، ويخرج المشاركون منها بسجل تهديدات نموذج الاحتيال وخطة تحصينه. فالمحتالون يختبرون أنظمة التقييم عمدا، فيدخلون معاملات مصطنعة وتصنيفات مسممة وبيانات اعتماد مسروقة للإفلات من ضوابط تعلم الآلة. وتقدم جوبيتر برايم للتدريب الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي بمنظور يضع الأمن أولا.

لماذا يهم الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي الآن

تشكل ثلاث فئات هذا الموضوع. فمجموعات الاحتيال المنظمة تتبادل أساليب التهرب وتتكيف خلال أيام حين يوقفها النموذج. وفرق الأمن التي اعتادت حماية الشبكات والأجهزة صارت مسؤولة عن النماذج ومسارات التدريب بوصفها أسطح هجوم جديدة. ولجان المخاطر والمدققون يطلبون دليلا على أن الكشف ما زال يعمل بعد الهجمات وانحراف البيانات. والنموذج المحصن يحد من الخسائر والرفض الخاطئ، أما النموذج المكشوف فيمرر الاحتيال بصمت بينما يتحمل العملاء الإزعاج. وتعتمد الدورة على معيار ISO/IEC 42001 لنظام إدارة الذكاء الاصطناعي وإطار AI RMF Management Framework لتنظيم تقييم التهديدات وضوابط نماذج الكشف. NIST AI Risk

المستفيدون من الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي

- محللو الاحتيال الذين يحققون في الحالات ويعيدون التصنيفات إلى النماذج
- مسؤولو الأمن الذين يحمون مسارات البيانات وبيانات الاعتماد ونقاط اتصال النماذج
- مديرو المخاطر الذين يحددون تقبل الخسارة ويراجعون فاعلية الضوابط
- علماء البيانات الذين يعيدون تدريب نماذج الشذوذ والتقييم ويتحققون منها
- قادة العمليات الذين يديرون قوائم التحقيق ومستويات الخدمة

تستفيد مجموعة مشتركة من فريقي الاحتيال والأمن أكثر حين تحضر معا، إذ تجتمع سيناريوهات الهجوم وضوابط النموذج في سجل واحد.

قدرات الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي

يتدرب المشاركون على النظر إلى نموذج الاحتيال بوصفه أصلا معرضا للهجوم ووضع ضوابط متناسبة حوله.

- تقييم مسارات الهجوم على نماذج الكشف بنمذجة تهديدات منظمة — مهندسو بنية الأمن
- اختبار متانة النموذج بمحاكاة التهرب والتسميم — قادة علوم البيانات
- تحديد ضوابط حماية بيانات التدريب ومخازن الخصائص — مسؤولو الخصوصية
- تصميم قواعد فرز الحالات التي توجه أنماط التجاوز المريبة إلى المحققين — عمليات الاحتيال
- بناء مراقبة للانحراف وإشارات الهجوم والتلاعب بالدرجات — لجان المخاطر
- عرض التهديدات المتبقية وأولويات التحصين — الإدارة العليا

مصطلحات الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي الأساسية

- هجوم التهرب — صياغة مدخلات، مثل معاملات مجزأة أو مموهة، تجعل النموذج يقيم الاحتيال على أنه سليم.
- تسميم البيانات — إدخال سجلات زائفة أو خاطئة التصنيف في بيانات التدريب لإضعاف الكشف لاحقا.
- استخراج النموذج — استجواب النموذج مرارا لنسخ سلوكه أو معرفة عتباته.
- كشف الشذوذ — أساليب ترصد السلوك البعيد عن الأنماط المعتادة دون حاجة إلى حالات احتيال مصنفة.
- انحراف المفهوم — تغير في السلوك الواقعي يجعل النموذج المدرب أقل دقة مع الوقت.
- الفريق الأحمر — محاولات مخططة ومأذون بها لتجاوز الضوابط بهدف اكتشاف الثغرات قبل المجرمين.

الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي يوما بيوم

اليوم 1: نموذج الاحتيال سطحا للهجوم

- رسم مسار الكشف — تحديد نقاط البيانات والنموذج والقرار
- نمذجة التهديدات بأسلوب STRIDE — حصر الهجمات على كل مكون
- مخاطر المطلعين وبيانات الاعتماد — حماية الوصول إلى القواعد والدرجات
- ISO/IEC 42001 وإطار NIST AI RMF — تنظيم ضوابط مخاطر النماذج
- سجل تهديدات نموذج الاحتيال — توثيق الأصول والتهديدات والمالكين

اليوم 2: أساليب الكشف ونقاط ضعفها

- التقييم الموجه مقابل كشف الشذوذ غير الموجه — اختيار نطاق التغطية
- غابات العزل والمرمزات التلقائية — رصد الأنماط غير المرئية سابقا
- انكشاف هندسة الخصائص — تقدير الإشارات التي يستطيع المهاجم تزييفها
- ضبط العتبات ونطاقات الدرجات — موازنة الخسائر والرفض الخاطئ
- جودة التصنيفات وحلقات التغذية الراجعة — منع إفساد الحقيقة المرجعية

اليوم 3: الهجمات العدائية والدفاعات

- محاكاة التهرب — اختبار المعاملات المجزأة والمموهة والمضبوطة السرعة
- سيناريوهات تسميم البيانات — اكتشاف سجلات التدريب المعبث بها
- استخراج النموذج واستجوابه — تحديد المعدل وضبط الاستجابات
- التدريب العدائي والنماذج المجمعة — تحصين النماذج ضد التلاعب
- تمارين الفريق الأحمر — تخطيط هجمات مأذون بها على الضوابط

اليوم 4: حماية البيانات والفرز والمراقبة

- تشفير بيانات التدريب وضبط الوصول — تأمين مخازن الخصائص
- الترميز المستعار لبيانات العملاء — الحد من الانكشاف في التحليل
- فرز أنماط التجاوز — توجيه حالات التهرب المشتبه بها إلى المحققين
- لوحات الانحراف وإشارات الهجوم — مراقبة توزيعات الدرجات
- الاستجابة لحوادث اختراق النموذج — التراجع وإعادة التدريب

اليوم 5: ممارسة سجل تهديدات نموذج الاحتيال وخطة تحصينه

- نمذجة تهديدات مسار الحالة — تطبيق STRIDE على تدفق الكشف
- محاكاة هجوم الحالة — تطبيق اختبارات التهرب والتسميم
- اختيار ضوابط الحالة — تطبيق حماية البيانات والمراقبة
- إحاطة لجنة المخاطر — عرض التهديدات المتبقية والأولويات
- سجل تهديدات نموذج الاحتيال وخطة تحصينه — استكمال الضوابط والمالكين

تطبيق الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي بعد الدورة

- إدارة جلسة نمذجة تهديدات بأسلوب STRIDE على مسار كشف الاحتيال القائم مع زملاء الأمن وعلوم البيانات.
- جدولة اختبار تهرب صغير للفريق الأحمر على نموذج تقييم واحد وتسجيل النتائج في سجل التهديدات.

أسئلة حول الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي

ما الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي ولماذا يهم الآن؟

هو ممارسة حماية نماذج كشف الاحتيال من التهرب والتسميم والسرقة. ويهم الآن لأن مجموعات الاحتيال المنظمة تختبر النماذج وتتكيف معها بسرعة، فتتحول ضوابط تعلم الآلة إلى أهداف.

من أكثر المستفيدين من الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي؟

يستفيد أكثر محللو الاحتيال ومسؤولو الأمن ومديرو المخاطر، ومعهم علماء البيانات الذين يعيدون تدريب النماذج وقادة العمليات الذين يديرون قوائم التحقيق.

كيف يختلف الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي عن دورة تحليلات الاحتيال البنكي؟

يتعامل مع النموذج نفسه بوصفه الأصل المعرض للخطر. فدورة تحليلات الاحتيال تبني الكشف وتضبط التنبيهات، أما هذه الدورة فتغطي نمذجة التهديدات والاختبار العدائي وحماية البيانات ومراقبة الاختراق.

ماذا تعود به من الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي؟

تعود بسجل تهديدات وخطة تحصين مبنيين على حالة عملية، يضمان سيناريوهات الهجوم ونتائج الاختبار والضوابط المختارة والمالكين المحددين.

الأمن السيبراني لنماذج كشف الاحتيال بالذكاء الاصطناعي: الطريق إلى الأمام

يوجه سجل تهديدات نموذج الاحتيال وخطة تحصينه أولا قرار لجنة المخاطر بشأن الضوابط التي تمولها. ثم ينفذها موظفو الأمن والاحتيال وعلوم البيانات ويختبرونها. وبعد أول تمرين للفريق الأحمر تستحق سيناريوهات الهجوم وتنبيهات الانحراف مراجعة جديدة، لأن الأساليب في قطاعات مثل المدفوعات الإلكترونية ومطالبات التأمين تتغير بسرعة.