JUPITER PRIME

Consulting • Training • Advisory

جوبيتر برايم للاستشارات والتدريب

كتيّب الدورة

دورة متقدمة في أمن وكلاء الذكاء الاصطناعي

24 يناير – 4 فبراير 2027

المنامة

info@jupiterprime.uk

jupiterprime.uk

مسجّلة في المملكة المتحدة والمملكة العربية السعودية

دورة متقدمة في أمن وكلاء الذكاء الاصطناعي

المرجع: 32_817 **التاريخ:** 24 يناير – 4 فبراير 2027 **المكان:** المنامة **الرسوم:** GBP 8000

أمن وكلاء الذكاء الاصطناعي دورة تدريبية متقدمة مدتها أسبوعين لمعماريي الأمن ومهندسي الذكاء الاصطناعي ومديري المخاطر السيبرانية، يعد فيها المشاركون نموذج تهديدات الوكيل وخطة الضوابط الأمنية لنظام مستقل تشغله مؤسستهم أو تنوي تشغيله. فالوكلاء الذين يتصفحون الويب ويستدعون الأدوات ويكتبون الشيفرة وينقلون الأموال يفتحون مسارات هجوم لم يتوقعها أمن التطبيقات التقليدي. وتقدم جوبيتر برايم للتدريب دورة أمن وكلاء الذكاء الاصطناعي لقادة الأمن التقني.

لماذا يهم أمن وكلاء الذكاء الاصطناعي الآن

تدفع هذا التخصص ثلاثة تحولات: وكلاء قائمون على النماذج اللغوية يكتسبون صلاحيات فعلية على البريد ومستودعات الشيفرة والمدفوعات وسجلات العملاء، ومهاجمون يتعلمون إخفاء التعليمات في صفحات الويب والمستندات ومخرجات الأدوات، ومجالس تسأل عمن يتحمل المساءلة حين يتصرف نظام مستقل تصرفا خاطئا. والمؤسسة التي تؤمن وكلاءها جيدا تمنحهم استقلالية مفيدة مع حصر نطاق الضرر، أما غيرها فتفتح الباب لتسريب البيانات والاحتيال وأفعال غير مصرح بها باسمها. وترسي الدورة ضوابطها على معيار ISO/IEC 42001 لنظام إدارة الذكاء الاصطناعي، وإطار إدارة مخاطر الذكاء الاصطناعي NIST AI RMF NIST AI Risk Management Framework.

المستفيدون من أمن وكلاء الذكاء الاصطناعي

- المعماريون الذين يصممون ربط الوكلاء بالأدوات ومخازن البيانات وأنظمة الهوية
- المهندسون الذين يبنون وكلاء النماذج اللغوية وينسقونهم وينشرونهم
- المحللون الذين يراقبون التنبيهات ويحققون في الحوادث ويستجيبون لإساءة استخدام الوكلاء
- المديرون الذين يقيمون المخاطر السيبرانية ويعتمدون صلاحيات الوكلاء ومستويات استقلاليتهم
- المختبرون الذين ينفذون التقييمات العدائية وتمارين الفريق الأحمر

تحقق وظيفة هندسة الأمن أكبر فائدة حين يحضر معماريوها ومطورو الوكلاء وفريق الرصد معا، لأن تصميم الصلاحيات ودفاعات مستوى الأوامر والمراقبة التشغيلية لا توقف الهجوم إلا إذا اشتركت الطبقات الثلاث في افتراضات التهديد نفسها.

قدرات أمن وكلاء الذكاء الاصطناعي

يتعلم المشاركون معاملة كل وكيل بوصفه مطلعا داخليا ذا صلاحيات قابلا للتلاعب، وهندسة دفاعات متعددة الطبقات حول تفكيره وأدواته وهويته.

- بناء نموذج تهديدات للوكيل يرسم حدود الثقة واستدعاءات الأدوات والمدخلات غير الموثوقة — مجالس معمارية الأمن
- اختبار الوكلاء ضد حقن الأوامر المباشر وغير المباشر وكسر القيود واختطاف الأهداف — قادة الفريق الأحمر والتوكيد
- تصميم صلاحيات أدوات بأقل امتياز وبيانات اعتماد محددة النطاق وبوابات موافقة بشرية — مالكو الهوية والوصول
- تحديد العزل وتصفية الحركة الخارجة والتحقق من المخرجات لمنع تسريب البيانات — مسؤولو حماية البيانات
- تطبيق المراقبة التشغيلية وتسجيل الأفعال وضوابط الإيقاف الطارئ على الوكلاء العاملين — مراكز العمليات الأمنية
- تقييم المخاطر المتبقية للوكلاء ومستويات استقلاليتهم وفق NIST AI RMFg ISO/IEC 42001 — لجان المخاطر السيبرانية

مصطلحات أمن وكلاء الذكاء الاصطناعي الأساسية

- وكيل الذكاء الاصطناعي — نظام يستخدم نموذجا لغويا للتخطيط وتنفيذ أفعال عبر الأدوات أو الواجهات البرمجية دون توجيه بشري خطوة بخطوة.
- حقن الأوامر — هجوم يدرج تعليمات خبيثة في مدخلات النموذج فيتجاهل الوكيل مهمته أو قواعده المقصودة.
- حقن الأوامر غير المباشر — تعليمات مخفية في صفحات الويب أو الملفات أو مخرجات الأدوات يقرؤها الوكيل لاحقا وينفذها.
- تحديد نطاق صلاحيات الأدوات — حصر الوظائف والبيانات والأفعال التي يجوز للوكيل استدعاؤها وشروط ذلك.
- بيئة عزل الوكيل — بيئة تنفيذ معزولة تقيد وصول الوكيل إلى الملفات والشبكة والعمليات لاحتواء الضرر.
- الاستقلالية المفرطة — الخطر الناشئ حين يملك الوكيل استقلالية أو صلاحيات أو وظائف تفوق ما تتطلبه مهمته.

أمن وكلاء الذكاء الاصطناعي يوما بيوم

اليوم 1: مشهد تهديدات الوكلاء

- معمارية الوكلاء من مخطط وذاكرة وحلقات أدوات — جرد سطح الهجوم
- تصنيف مستويات الاستقلالية — مقدار الحرية الملائم لكل حالة استخدام
- أنماط اختراق الوكلاء الحديثة — دروس من الحوادث المعلنة
- دوافع المهاجمين ضد الوكلاء — الاحتيال والتجسس والتخريب
- سجل أصول الوكلاء — الأنظمة والبيانات وبيانات الاعتماد التي يلمسها الوكلاء

اليوم 2: نمذجة التهديدات للأنظمة المستقلة

- مخططات حدود الثقة لسير عمل الوكلاء — مواضع دخول المدخلات غير الموثوقة
- رسم تدفق البيانات عبر الأدوات والذاكرة — تعرض السجلات الحساسة
- كتابة حالات إساءة الاستخدام للوكلاء — سيناريوهات سوء الاستخدام قبل البناء
- أشجار الهجوم لاختطاف الأهداف — ترتيب أولويات الاستثمار الدفاعي
- إيقاع مراجعة نموذج التهديدات — التحديث مع تغير الأدوات والأوامر

اليوم 3: دفاعات حقن الأوامر والتلاعب

- تقنيات الحقن المباشر وكسر القيود — التعرف على أنماط التجاوز
- الحقن غير المباشر عبر المستندات والويب والبريد — قواعد الثقة في المحتوى
- تسلسل التعليمات وتحصين أوامر النظام — قرارات الأسبقية
- مصنفات المدخلات ونماذج الحراسة — تصفية المحتوى العدائي
- الإبراز والفصل بين البيانات والتعليمات — الحد من هجمات الالتباس

اليوم 4: إساءة استخدام الأدوات والإضافات والصلاحيات

- تصميم الأدوات بأقل امتياز — وظائف ضيقة بدلا من قدرات عامة
- التحقق من المعاملات وقوائم السماح — منع الوسائط الخطرة
- مخاطر الوكيل المرتبك في سلاسل الأدوات — تصعيد الامتيازات بين الأدوات
- بوابات الموافقة البشرية للأفعال عالية الأثر — المدفوعات والحذف والرسائل
- فحص الإضافات والموصلات من أطراف ثالثة — التعرض عبر سلسلة الإمداد

اليوم 5: هوية الوكلاء والتحكم في الوصول

- دورة حياة الهويات غير البشرية للوكلاء — الإصدار والتدوير والإلغاء
- بيانات اعتماد محددة النطاق قصيرة العمر وتبادل الرموز — الحد من الوصول المسروق
- السلطة المفوضة والتصرف نيابة عن المستخدم — التصرف عن المستخدمين بأمان
- الثقة بين الوكلاء المتعددين والمصادقة البينية — منع انتحال الهوية
- إنفاذ سياسة انعدام الثقة على الوكلاء — قرارات تفويض مستمرة

اليوم 6: تسريب البيانات وحماية الذاكرة

- قنوات التسريب عبر الروابط والصور واستدعاءات الأدوات والسجلات — مسارات التسرب
- تصفية الحركة الخارجة والتحقق من المخرجات — منع الإفصاح الخفي عن البيانات
- ضوابط الوصول لمخازن الاسترجاع وقواعد البيانات المتجهية — صلاحيات على مستوى المستند
- تسميم الذاكرة وهجمات الاستمرار — سلامة الوكيل على المدى الطويل
- حجب البيانات الحساسة في نوافذ السياق — الخصوصية بالتصميم

اليوم 7: العزل وفصل بيئات التنفيذ

- بيئات تشغيل الوكلاء في الحاويات والأجهزة الافتراضية — اختيار قوة العزل
- بيئات عزل تنفيذ الشيفرة للوكلاء — حدود الملفات والشبكة والعمليات
- عزل وكلاء المتصفح — التعامل مع المواقع العدائية
- حصص الموارد وحدود المعدل — منع الوكلاء الجامحين
- الفصل بين بيئتي الاختبار والإنتاج — مسارات ترقية آمنة

اليوم 8: الفريق الأحمر والتقييم العدائي

- تخطيط الفريق الأحمر للوكلاء وقواعد الاشتباك — قرارات النطاق
- التوليد الآلي للأوامر العدائية — توسيع تغطية الهجمات
- مجموعات القياس لحقن الأوامر وإساءة استخدام الأدوات — قياس المتانة
- تصنيف خطورة ثغرات الوكلاء — أولوية المعالجة
- اختبار الانحدار بعد تحديث النموذج أو الأوامر — صون الدفاعات

اليوم 9: المراقبة التشغيلية والاستجابة للحوادث

- تسجيل أفعال الوكلاء والتقاط المسارات — الأدلة الجنائية الرقمية
- كشف الشذوذ السلوكي في استدعاءات الأدوات — مؤشرات الاختراق المبكرة
- مفاتيح الإيقاف الطارئ وقواطع الدارة — إيقاف الوكلاء المنحرفين
- أدلة الاستجابة لحوادث الوكلاء — الاحتواء والإزالة والتعافي
- ربط الضوابط بمعيار ISO/IEC 42001 وإطار NIST AI RMF — أدلة المساءلة

اليوم 10: ممارسة نموذج تهديدات الوكيل وخطة الضوابط الأمنية

- تطبيق نمذجة التهديدات على سير عمل وكيل المشارك — أولويات المخاطر
- تطبيق تحديد نطاق الصلاحيات على مجموعة أدواته — قرارات أقل امتياز
- تطبيق نتائج الفريق الأحمر على اختيار الدفاعات — فجوات الضوابط
- تطبيق تصميم المراقبة على النشر الإنتاجي — تغطية الرصد
- تجميع نموذج تهديدات الوكيل وخطة الضوابط الأمنية — المالكون والضوابط ونقاط المراجعة

تطبيق أمن وكلاء الذكاء الاصطناعي بعد الدورة

- إعداد مخطط حدود الثقة وقائمة حالات إساءة الاستخدام لوكيل قيد التطوير، وعرض أعلى خمسة مخاطر على مالك المنتج.
- تنفيذ اختبار حقن أوامر غير مباشر بتعليمات مزروعة في مستندات يقرؤها الوكيل، وتسجيل الدفاعات التي صمدت والتي أخفقت.

أسئلة حول أمن وكلاء الذكاء الاصطناعي

ما أمن وكلاء الذكاء الاصطناعي، ولماذا يهم الآن؟

هو تخصص حماية وكلاء النماذج اللغوية المستقلين من التلاعب وإساءة الاستخدام وتسريب البيانات. ويهم الآن لأن الوكلاء يملكون صلاحيات فعلية متزايدة على الشيفرة والمدفوعات والسجلات، والمهاجمون يخفون التعليمات بالفعل في المحتوى الذي يقرؤه هؤلاء الوكلاء.

من يستفيد أكثر من تدريب أمن وكلاء الذكاء الاصطناعي؟

يستفيد ممارسو الأمن والذكاء الاصطناعي ذوو الخبرة أكثر من غيرهم: معماريو الأمن ومطورو الوكلاء ومهندسو الرصد ومديرو المخاطر السيبرانية. وتستفيد إذا كنت تصمم وكلاء أو تبنيهم أو تعتمدهم أو تراقبهم وهم يتصرفون باسم مؤسستك.

كيف تختلف الدورة المتقدمة في أمن وكلاء الذكاء الاصطناعي عن دورة عامة في الأمن السيبراني؟

تركز على التهديدات الخاصة بالذكاء الاصطناعي المستقل. فالدورة العامة تغطي الشبكات والأجهزة الطرفية والتطبيقات، أما هذه الدورة فتعالج حقن الأوامر وإساءة استخدام الأدوات وهوية الوكلاء والعزل واختبار الفريق الأحمر لأنظمة تفكر وتتصرف.

ماذا تعود به من تدريب أمن وكلاء الذكاء الاصطناعي؟

تعود بنموذج تهديدات الوكيل وخطة الضوابط الأمنية. ويرسم حدود الثقة ومسارات الهجوم، ويحدد ضوابط الصلاحيات والعزل والمراقبة، ويسجل نتائج الفريق الأحمر، ويسمي المالكين ونقاط المراجعة لكل دفاع.

أمن وكلاء الذكاء الاصطناعي: الطريق إلى الأمام

يوجه نموذج تهديدات الوكيل وخطة الضوابط الأمنية أولا قرار انتقال الوكيل من الاختبار إلى الإنتاج والصلاحيات الممنوحة له. ويستخدمه معماريو الأمن لاعتماد التصاميم، والمهندسون لتطبيق الضوابط، ومديرو المخاطر لقبول التعرض المتبقي. وبعد أول حادثة إنتاجية أو تحديث كبير للنموذج يستحق مستوى الاستقلالية المحدد المراجعة، لأن السلوك الفعلي يختلف غالبا عن التصميم.