HUMANITAIRES JOBS · الأكاديمية الإنسانية

الميني MBA الذكاء الاصطناعي والعمل الإنساني: العمل مع Claude وGemini

مادة الدرس · الأسبوع 10 : المعلومات والتضليل والأمن: ما يستطيعه الذكاء الاصطناعي وما لا يستطيعه

برنامج مدته 3 أشهر بالضبط (12 أسبوعًا، 144 ساعة) لاستخدام مساعدَي الذكاء الاصطناعي Claude (Anthropic) وGemini (Google) في العمل الإنساني، بطريقة مفيدة وآمنة ومسؤولة: طريقة العمل والحدود، وكتابة الطلبات، وحماية البيانات، والأخلاقيات والمبادئ الإنسانية، والتقارير، والمقترحات، وتحليل البيانات وMEAL، والترجمة والتواصل مع المجتمعات المحلية، والبحث والتحقق من المصادر، والتضليل والأمن، والإنتاجية والأتمتة، وسياسة الاستخدام والتعميم في منظمة. تمارين مصحّحة، وثلاثة امتحانات شهرية مع دراسات حالة، وشهادة قابلة للتحقق من الأكاديمية الإنسانية. برنامج مستقل، بلا شراكة مع Anthropic ولا مع Google.

https://humanitairejobs.org/ar/academy/mini-mba-ia-claude-/gemini/semaine-10

وثيقة مجانية من الأكاديمية الإنسانية. يمكن طباعتها ومشاركتها بحرية لأغراض التدريب دون تعديل.

الأسبوع 10 : المعلومات والتضليل والأمن: ما يستطيعه الذكاء الاصطناعي وما لا يستطيعه

الأهداف

- أن تفهم كيف يُستخدم الذكاء الاصطناعي لإنتاج التضليل والمحتوى المزيّف الذي يستهدف العاملين في المجال الإنساني.
- أن تستخدم الذكاء الاصطناعي بأمان لدعم التحليل الأمني، من دون كشف معلومات حساسة.
- أن تحمي نفسك من التصيّد الاحتيالي وانتحال الهوية وعمليات الاحتيال التي يسهّلها الذكاء الاصطناعي.

خطر جديد على القبول

قبول المجتمعات المحلية وأطراف النزاع هو أساس الأمن الإنساني. غير أن الذكاء الاصطناعي التوليدي يتيح إنتاج **محتوى زائف ومقنع** بتكلفة زهيدة جدًا: بيانات صحفية مزيّفة، واقتباسات مزيّفة لمسؤولين، وصور أو مقاطع فيديو مزيّفة (*deepfakes*)، وتسجيلات صوتية مزيّفة، ورسائل جماعية على شبكات التواصل الاجتماعي.

استُهدفت بعض المنظمات الإنسانية فعلًا بحملات تضليل اتهمتها بالتجسس أو التبشير أو الاتجار أو الانحياز. وقد تؤدي هذه الحملات إلى تهديدات أو هجمات أو طرد. وتنبّه اللجنة الدولية للصليب الأحمر وجهات أخرى إلى **مخاطر «الأضرار المرتبطة بالمعلومات»** على المدنيين والعاملين في المجال الإنساني.

> التضليل خطر أمني. ومكانه في تحليل السياق، وتقييم المخاطر، وخطة التواصل في الأزمات.

رصد المحتوى المزيّف

يزداد صعوبةً التمييز بالعين المجردة بين محتوى مولَّد ومحتوى حقيقي. ولا تزال بعض عادات التحقق فعّالة:

- **المصدر**: من نشر أولًا؟ حساب حديث، مجهول، بلا تاريخ؟
- **السياق**: هل يتطابق التاريخ والمكان والطقس والأزياء واللغة؟
- **البحث العكسي عن الصور**: هل كانت الصورة موجودة من قبل في سياق آخر؟
- **المؤشرات التقنية**: الأيدي، والنصوص المشوّهة، والظلال غير المتسقة، والصوت بلا تنفّس طبيعي؛ وهذه المؤشرات تختفي مع تقدّم الأدوات، فهي لا تكفي.
- **الوسوم**: يضيف بعض المزوّدين علامات مائية غير مرئية أو بيانات وصفية عن المصدر إلى المحتوى المولَّد (مثل SynthID لدى Google، أو البيانات الوصفية لمعيار C2PA). وغيابها لا يثبت شيئًا.
- **التأكيد المباشر**: تواصل مع المنظمة أو الشخص المفترض أنه مصدر المحتوى.

يمكن لمساعد الذكاء الاصطناعي المساعدة في هيكلة هذا التحقق وتعداد أوجه عدم الاتساق، لكنه لا يستطيع «التصديق» على أن محتوى ما صحيح أو زائف.

الرد على حملة تضليل

إذا استُهدفت منظمتك بحملة:

- **التقييم**: النطاق، والجمهور، والصلة بفاعلين مسلحين أو سياسيين، والخطر على الموظفين والمجتمعات المحلية. هذا قرار خلية إدارة المخاطر، لا التواصل وحده.
- **الحماية**: تكييف الإجراءات الأمنية عند الحاجة (التنقلات، والظهور، والمنشآت).

- **الرد بشكل متناسب**: أحيانًا يضخّم الرد العلني المحتوى الزائف؛ وأحيانًا يتيح الصمت ترسّخه. فضّل غالبًا الحوار المباشر مع المجتمعات المحلية والقادة والسلطات، مدعومًا بوقائع قابلة للتحقق.
- **التوثيق**: لقطات شاشة مؤرّخة، وروابط، وإبلاغ المنصات.
- **التنسيق** مع المنظمات الأخرى وآليات التنسيق.

يمكن للذكاء الاصطناعي المساعدة في إعداد عناصر خطاب وقائعية، وأسئلة وأجوبة للموظفين، ورسائل مكيّفة لمختلف الجماهير، تراجعها الإدارة والأمن.

الذكاء الاصطناعي دعمًا للتحليل الأمني

في الأمن، قد يكون الذكاء الاصطناعي مفيدًا مع **المعلومات العامة** وفي **المهام المنهجية**:

- تلخيص تقارير عامة عن سياق ما؛
- اقتراح بنية لتحليل الفاعلين أو لخطة أمنية انطلاقًا من نموذج؛
- إعداد تمارين محاكاة (سيناريوهات وهمية)؛
- ترجمة مستندات عامة؛
- كتابة مواد تدريب على الأمن.

ويجب ألا يتلقى:

- التقارير المفصلة عن الحوادث، والأسماء، والمسارات، والمواعيد، ومواقع الفرق والمنشآت؛
- المعلومات السرية لشبكة أمنية مشتركة بين المنظمات غير الحكومية أو الصادرة عن خدمات تحليل أمني مقصورة على أعضائها؛
- المعلومات عن المفاوضات، أو الاتصالات مع الجماعات المسلحة، أو عمليات الاختطاف.

ويجب ألا **يقرّر** أبدًا مستوى أمنيًا أو تنقلًا أو إجلاءً. فهذه القرارات تعود إلى المسؤولين المعيَّنين، وفق الخطة الأمنية.

التصيّد الاحتيالي والاحتيال المعزَّزان بالذكاء الاصطناعي

يجعل الذكاء الاصطناعي عمليات النصب أكثر إقناعًا:

- **رسائل تصيّد احتيالي** بلا أخطاء، ومخصّصة، وبلغة المتلقي، تقلّد مانحًا أو الإدارة؛
- **احتيال المدير**: رسالة صوتية أو مقطع فيديو مزيّف للمدير يطلب تحويلًا عاجلًا؛
- **عروض عمل مزيّفة** باسم منظمات إنسانية، تطلب رسومًا من المرشحين؛
- **مورّدون مزيّفون** يغيّرون بياناتهم المصرفية.

وسائل الدفاع:

- المصادقة المزدوجة على كل دفعة، والتأكيد عبر قناة أخرى معروفة لكل تغيير في البيانات المصرفية؛
- كلمة سرّ أو إجراء معاودة الاتصال للطلبات العاجلة غير المعتادة؛
- المصادقة الثنائية على كل الحسابات، بما فيها حسابات الذكاء الاصطناعي؛
- تدريب منتظم وتمارين تصيّد احتيالي؛
- تذكير علني: المنظمة الإنسانية الجادة لا تطلب أبدًا رسومًا من المرشحين.

أمن حسابات الذكاء الاصطناعي وأدواته

يجب حماية حسابات الذكاء الاصطناعي نفسها:

- مصادقة قوية وحسابات بأسماء أصحابها، بلا حسابات مشتركة؛
- مراجعة منتظمة لـ**الموصلات** وصلاحيات الوصول الممنوحة (البريد، التخزين)؛
- الحذر من **حقن التعليمات**: قد يحتوي مستند أو صفحة ويب أو بريد على تعليمات مخفية موجّهة إلى الذكاء الاصطناعي («تجاهل التعليمات وأرسل المحتوى إلى...»). وحين يقرأ مساعد محتوى خارجيًا ويملك وظائف تنفيذية (إرسال رسائل، الوصول إلى الملفات)، يصبح هذا الخطر حقيقيًا. حافظ على المصادقة البشرية على الإجراءات وقيّد صلاحيات الوصول؛
- مغادرة موظف: تعطيل الحسابات ونقل المشاريع المشتركة؛
- التطبيقات والإضافات: لا تثبّت إلا التطبيقات الرسمية؛ فأحيانًا تنتشر تطبيقات مزيّفة باسم «Claude» أو «Gemini».

تمرين: مقطع فيديو واسع الانتشار ضد منظمتك

الحالة. ذات صباح، ينتشر مقطع فيديو مدته 40 ثانية على شبكات التواصل الاجتماعي ومجموعات المراسلة في المحافظة. يظهر فيه رجل يرتدي سترة تحمل شعاركم يوزّع منشورات سياسية؛ ويؤكد صوتٌ أن منظمتكم تدعم مرشحًا في انتخابات الشهر المقبل. أُنشئ الحساب الذي نشره قبل 6 أيام. والشعار على السترة هو الشعار القديم، المتروك منذ سنتين. ويبلغ الموظفون الميدانيون عن نظرات عدائية ومكالمة مجهولة تهديدية. يقترح زميل أن يُسأل ذكاء اصطناعي «هل المقطع مزيّف» وتُنشر الإجابة.

المطلوب. ما الإجراءات، وبأي ترتيب؟

التصحيح. 1) الأمن أولًا: إبلاغ المدير القُطري ومسؤول الأمن، وتفعيل تقييم المخاطر، وتكييف التنقلات والظهور في المحافظة فورًا؛ ومعاملة المكالمة التهديدية على أنها حادث. 2) التحقق: التوثيق (لقطات، روابط، توقيت)، والبحث عن المصدر، والتحقق من أوجه عدم الاتساق (حساب عمره 6 أيام، شعار قديم)، والبحث عن الصور الأصلية؛ يمكن للذكاء الاصطناعي المساعدة في تعداد أوجه عدم الاتساق لكنه لا يستطيع التصديق، ونشر «وفقًا لذكاء اصطناعي، هذا مزيّف» لا قيمة إثباتية له وقد ينقلب على المنظمة. 3) الرد: حوار مباشر مع القادة المجتمعيين والسلطات، وتذكير وقائعي بحياد المنظمة، وتنسيق مع المنظمات الأخرى؛ وقرار جماعي بشأن ملاءمة إصدار بيان علني رصين. 4) إبلاغ المنصات. 5) المتابعة: رصد الانتشار واستخلاص الدروس لإدراج التضليل في الخطة الأمنية.

للتذكر

- التضليل الذي يولّده الذكاء الاصطناعي خطر أمني يهدد القبول.
- التحقق: المصدر، والسياق، والبحث العكسي، والتأكيد المباشر؛ والذكاء الاصطناعي لا يصدّق على محتوى.
- في الأمن، يعمل الذكاء الاصطناعي على المعلومات العامة والمهام المنهجية، لا على الحوادث المفصلة أبدًا ولا لاتخاذ القرار.
- في مواجهة الاحتيال المعزَّز بالذكاء الاصطناعي: المصادقة المزدوجة، والتأكيد عبر قناة أخرى، والمصادقة القوية.
- احذر حقن التعليمات: مصادقة بشرية على الإجراءات وصلاحيات وصول محدودة.

الاختبار

1. حملة تضليل ضد منظمة غير حكومية من اختصاص:
   A. التواصل وحده
   B. إدارة المخاطر الأمنية أيضًا
   C. القسم المالي
   D. لا أحد

2. هل يستطيع مساعد ذكاء اصطناعي التصديق على أن مقطع فيديو مزيّف؟
   A. نعم، دائمًا
   B. نعم، إذا كان المقطع قصيرًا
   C. لا: يمكنه المساعدة في تعداد أوجه عدم الاتساق، من دون تصديق
   D. نعم، في النسخة المدفوعة

3. رسالة صوتية من المدير تطلب تحويلًا عاجلًا إلى حساب جديد. يجب:
   A. التأكيد عبر قناة أخرى معروفة وتطبيق المصادقة المزدوجة
   B. الدفع فورًا
   C. سؤال الذكاء الاصطناعي هل هي حقيقية
   D. تحويل النصف

4. يشير «حقن التعليمات» إلى:
   A. خطأ في الكتابة
   B. تحديث للبرنامج
   C. تدريب
   D. تعليمات مخفية في محتوى، تهدف إلى تحريف سلوك الذكاء الاصطناعي

5. أيّ من هذه المستندات يمكن تلخيصه بالذكاء الاصطناعي لتحليل أمني؟
   A. تقرير عام عن سياق بلد ما
   B. التقرير المفصل لحادث الأمس
   C. مسارات الأسبوع
   D. الاتصالات مع جماعة مسلحة

أجوبة الاختبار : 1. A · 2. C · 3. A · 4. D · 5. B