الهيئة الوطنية للأمن السيبراني
National Cybersecurity Authority

مشروع إرشادات الأمن السيبراني للذكاء الاصطناعي

AI Cybersecurity Guidelines
(AICG - 1: 2026)

إشارة المشاركة: شفاف

تصنيف الوثيقة: عام

تنويه: لمواكبة المتغيرات بشأن تحديثات الوثائق الصادرة عن الهيئة الوطنية للأمن السيبراني، تود الهيئة الوطنية للأمن السيبراني التنويه على أهمية الاعتماد الدائم على نسخ الوثائق المنشورة في الموقع الإلكتروني للهيئة https://nca.gov.sa

بسم الله الرحمن الرحيم

إخلاء مسؤولية: تم إعداد الإرشادات الواردة في هذه الوثيقة بناءً على أفضل الممارسات في مجال الأمن السيبراني للذكاء الاصطناعي، وهي إرشادات توعوية، تهدف إلى الحد من المخاطر المتعلّقة بالأمن السيبراني لأنظمة الذكاء الاصطناعي، والتخفيف من آثارها في الوقت المناسب. وتخلي الهيئة مسؤوليتها من أي تبعات قد تترتب بشكل مباشر أو غير مباشر على اتخاذ أي إجراءات؛ بناءً على المعلومات الواردة في هذه الوثيقة. وعند وجود تعارض بينَ ما ورد في هذه الوثيقة؛ مع أي متطلبات تشريعية أو تنظيمية؛ فإن تلك المتطلبات تحل محل ما ورد في هذه الوثيقة. وتحُث الهيئة الوطنية للأمن السيبراني، جميع الجهات؛ بإجراء تقييمات دورية لتلك المخاطر.

بروتوكول الإشارة الضوئية (TLP):

يستخدم هذا البروتوكول على نطاق واسع في العالم وهناك أربعة ألوان (إشارات ضوئية):

أحمر (شخصي وسري للمستلم فقط)

المستلم لا يحق له مشاركة المصنف بالإشارة الحمراء مع أي فرد، سواء أكان ذلك من داخل الجهة أم خارجها؛ خارج النطاق المحدد للاستلام.

برتقالي + مشدد (مشاركة في نفس الجهة)

المستلم يمكنه مشاركة المعلومات في الجهة نفسها مع الأشخاص المعنيين فحسب.

برتقالي (مشاركة محدودة)

المستلم يمكنه مشاركة المعلومات في الجهة نفسها مع الأشخاص المعنيين فحسب. ومن يتطلب الأمر منه اتخاذ إجراء يخص المعلومة.

أخضر (مشاركة في نفس المجتمع)

المستلم يمكنه مشاركة المعلومات مع آخرين في الجهة نفسها، أو جهة أخرى على علاقة معهم أو في القطاع نفسه؛ ولا يسمح بتبادلها أو نشرها من خلال القنوات العامة.

شفاف (غير محدود)

قائمة المحتويات

قائمة الأشكال والرسومات التوضيحية

قائمة الجداول

الملخص التنفيذي

تُعد الهيئة الوطنية للأمن السيبراني؛ بموجب تنظيمها الصادر بالأمر الملكي الكريم ذي الرقم (٦٨٠١) في ١١/٢/١٤٣٩هـ الجهة المُختصة بالأمن السيبراني في المملكة، والمرجع الوطني في شؤونه. وتشمل اختصاصات الهيئة ومهماتها، دون حصر؛ وضع السياسات، وآليات الحوكمة، والأطر، والمعايير، والضوابط، والإرشادات، المتعلقة بالأمن السيبراني، وذلك لدعم الأثر المهم للأمن السيبراني، الذي ازداد مع ارتفاع مخاطر الأمن السيبراني، أكثر من أي وقت مضى.

ويشهد العالم تطوراً متسارعًا في تقنيات الذكاء الاصطناعي؛ مما قد ينجم عنه ثغرات ومخاطر سيبرانية، قد تختلف في طبيعتها عن مخاطر الأمن السيبراني التقليدية. ونظرًا للسرعة الهائلة التي يشهدها الذكاء الاصطناعي، من مثل التطورات في الذكاء الاصطناعي التوليدي، والذكاء الاصطناعي التوكيلي؛ فإن الأمر يتطلب إرشادات متخصصة، تتعامل هذه المخاطر بصورة أكثر دقة، مع الأخذ في الحسبان ضمان دمج الأمن السيبراني، في جميع مراحل دورة حياة أنظمة الذكاء الاصطناعي، ابتداءً من التصميم، والتطوير إلى التشغيل والاستخدام المستمر، وحتى إنهاء الاستخدام.

وانطلاقًا من هذه التوجهات؛ قامت الهيئة الوطنية للأمن السيبراني؛ بإعداد إرشادات الأمن السيبراني للذكاء الاصطناعي (AICG-1: 2026) بعد إجراء دراسة شاملة للعديد من الإرشادات والمعايير، والأطر والضوابط الدولية، المتعلقة بالأمن السيبراني، وتحليل الوضع الحالي للمبادرات الوطنية والمتطلبات التنظيمية ذات العلاقة؛ بهدف التعامل مع المخاطر السيبرانية النوعية، المرتبطة بتقنيات الذكاء الاصطناعي، التي تواجه الجهات.

المقدمة

قامت الهيئة الوطنية للأمن السيبراني (ويُشار لها في الوثيقة بـ "الهيئة") بإصدار إرشادات الأمن السيبراني للذكاء الاصطناعي (AICG-1: 2026) بعد إجراء دراسة شاملة لعدة إرشادات، ومعايير، وأطر، وضوابط وطنية ودولية، ومراجعة لأفضل الممارسات والتجارب العالمية الرائدة، في مجال الأمن السيبراني للذكاء الاصطناعي.

تتكون إرشادات الأمن السيبراني للذكاء الاصطناعي من:

- ٤ مكونات أساسية (Main Domains).
- ١٥ مكونًا فرعيًا (Subdomains).
- ٤٢ إرشادًا (Guidelines).

الأهداف

جرى تطوير إرشادات الأمن السيبراني للذكاء الاصطناعي (AICG-1: 2026) دعمًا للأهداف الوطنية للأمن السيبراني؛ من خلال معالجة المخاطر السيبرانية النوعية، المرتبطة بأنظمة الذكاء الاصطناعي.

وقد حُدِّدت هذه الإرشادات؛ للتعامل مع مخاطر الأمن السيبراني، من خلال مراحل دورة حياة نظام الذكاء الاصطناعي كافة، بما يشمل التصميم، والتطوير، والتشغيل، والاستخدام المستمر، وحتى إنهاء الخدمة. كما تهدف هذه الوثيقة؛ ضمان تطبيق متطلبات الأمن السيبراني للذكاء الاصطناعي.

ويشمل نطاق الوثيقة، التقنيات الناشئة للذكاء الاصطناعي؛ مثل الذكاء الاصطناعي التوليدي (Generative AI) والذكاء الاصطناعي التوكيلي (Agentic AI) ، وبشكل عام، صممت هذه الوثيقة لوضع متطلبات الأمن السيبراني (والتي تم تحديدها في وثائق الهيئة الأخرى) في سياق أنظمة الذكاء الاصطناعي، بما يُسهم في تعزيز الأمن السيبراني لدى الجهات.

يوضح الشكل (١) أهم الأهداف التي تسعى هذه الوثيقة إلى تحقيقها، وهي التأكيد على أهمية حماية المكونات المختلفة لأنظمة الذكاء الاصطناعي، وكذلك الحماية من التهديدات الناشئة عن استخدام أنظمة الذكاء الاصطناعي.

شكل ١ :أبرز أهداف وثيقة إرشادات الأمن السيبراني للذكاء الاصطناعي

نطاق العمل وقابلية التطبيق

توصي الهيئة جميع الجهات في المملكة التي تتبنى استخدام أنظمة الذكاء الاصطناعي، أو تخطط لتبني استخدامها، باتباع هذه الإرشادات، وتطبيق الحد الأدنى من أفضل الممارسات؛ بهدف التقليل من مخاطر الأمن السيبراني، التي قد تنتج من استخدام هذه التقنيات وتبنيها.

وتؤكد الهيئة على أن هذه الإرشادات توعوية؛ وتهدف للمساعدة في تعزيز الأمن السيبراني، أثناء استخدام أنظمة الذكاء الاصطناعي وتبنيها، بهدف ضمان تطبيق أفضل الممارسات؛ للحد من مخاطر الأمن السيبراني، وزيادة القدرة على الصمود، في مواجهة هجمات الأمن السيبراني.

ونظراً للطبيعة المتغيرة لتهديدات الأمن السيبراني؛ تحث الهيئة جميع الجهات على مراجعة مخاطر الأمن السيبراني، وتقييمها بشكل دوري؛ لتحديد مدى الحاجة إلى اتخاذ أي تدابير إضافية، فيما يتعلق بالأمن السيبراني لأنظمة الذكاء الاصطناعي.

نظرة عامة عن الذكاء الاصطناعي وتهديدات الأمن السيبراني للذكاء الاصطناعي

تعريف أنظمة الذكاء الاصطناعي

بحسب تعريف الهيئة السعودية للبيانات والذكاء الاصطناعي (SDAIA)، عُرفت **أنظمة الذكاء الاصطناعي** بأنها " أنظمة برمجية تعتمد على تقنيات متقدمة، تُمكنها من التنبؤ، أو توليد المحتوى، أو تقديم التوصيات، أو اتخاذ القرارات، بمستويات مختلفة من الاستقلالية، اعتماداً على البيانات والسياق الذي تعمل فيه". [1]

تعريف الذكاء الاصطناعي التوليدي

بحسب تعريف الهيئة السعودية للبيانات والذكاء الاصطناعي (SDAIA) عُرف **الذكاء الاصطناعي التوليدي** بأنه " فرع من فروع الذكاء الاصطناعي يعتمد على تقنيات تعلم الآلة والشبكات العصبية العميقة لمحاكاة قدرة الإنسان على إنتاج بيانات ومحتوى أصيل". [2]

تعريف الذكاء الاصطناعي التوكيلي

بحسب تعريف الهيئة السعودية للبيانات والذكاء الاصطناعي (SDAIA) عُرف **الذكاء الاصطناعي التوكيلي** بأنه "نظام برمجي يعتمد على خوارزميات الذكاء الاصطناعي ويتسم بخصائص مثل: إدراك البيئة المحيطة وتفسير المعلومات وتحديد الأهداف واتخاذ القرارات ذاتياً دون الحاجة إلى إشراف بشري مستمر". [2]

مكونات أنظمة الذكاء الاصطناعي

تتكون أنظمة الذكاء الاصطناعي، من مجموعة من المكونات المترابطة، التي تستقبل البيانات، وتعالج المدخلات، وتتعلم الأنماط، وتنتج المخرجات. وتُبنى أنظمة الذكاء الاصطناعي التوليدي، على مجموعة من المكونات الأساسية، التي تدعم توليد المحتوى

[1] SDAIA, إطار تبني الذكاء الاصطناعي, Available at: https://sdaia.gov.sa/en/SDAIA/about/Files/AIAdoptionFramework.pdf.

[2] SDAIA, الذكاء الاصطناعي, Available at: https://sdaia.gov.sa/en/SDAIA/about/Pages/AboutAI.aspx

والاستدلال. وفي بعض الحالات؛ قد تتضمن هذه الأنظمة قدرات توكيلية إضافية، تُمكّن من التخطيط، واستخدام الأدوات، وتنفيذ المهمات متعددة الخطوات؛ لتحقيق أهداف محددة. وتشمل المكونات الأساسية للذكاء الاصطناعي التوليدي - على سبيل المثال لا الحصر - ما يلي:

١ .نموذج الذكاء الاصطناعي (AI Model)

وهو المكون المسؤول عن معالجة المدخلات، وإجراء الاستدلال (Inference)، وإنتاج المخرجات. وقد يشمل ذلك نموذج اللغة الكبير (LLM) أو نماذج متعددة الوسائط، أو نماذج الذكاء الاصطناعي الأخرى، التي تدعم قدرات الذكاء الاصطناعي المستهدفة.

٢. مكونات مستودعات البيانات والمعرفة(Data & Knowledge Repositories components)

وتشمل مصادر تخزين البيانات، ومصادر المعرفة التي تستخدمها أنظمة الذكاء الاصطناعي؛ بما في ذلك وثائق الجهة، وقواعد البيانات، وقواعد المعرفة، ومصادر البيانات الداخلية أو الخارجية الأخرى.

٣ .مكونات الاسترجاع وإدارة السياق(Retrieval & Context Management components)

يُعنى هذا المكون باسترجاع المعلومات، وتوفير البيانات السياقية لنموذج الذكاء الاصطناعي؛ لتحسين جودة الاستجابة، ودقتها، وملاءمتها. ويمكن تطبيق هذه الإمكانيات من خلال آليات الاسترجاع؛ بما في ذلك تقنية التوليد المعزز بالاسترجاع (Retrieval-Augmented Generation (RAG)).

٤ .مكونات طبقة التطبيق والتفاعل(Application and interaction layer components)

وهي الواجهات التي يتفاعل من خلالها المستخدمون، والأنظمة الخارجية، مع نظام الذكاء الاصطناعي؛ بما في ذلك واجهات المحادثات، وأوامر المستخدم، والمساعد الافتراضي، وواجهات برمجة التطبيقات (APIs).

كما تتضمن أنظمة الذكاء الاصطناعي، التي تدمج القدرات التوكيلية، مكونات إضافية تُمكّنها من تنفيذ المهمات بشكل مستقل أو شبه مستقل، بما يتجاوز مجرد توليد المحتوى. ومن هذه المكونات - على سبيل المثال لا الحصر - ما يلي:

٥ .مكونات التخطيط(Planning Components)

يُعنى هذا المكون بتقسيم الهدف الرئيسي إلى مهمات أصغر، وتحديد تسلسل التنفيذ، وتنسيق إنجاز المهمات.

٦ .مكونات الذاكرة (Memory Components)

يحتفظ هذا المكون بالمعلومات السياقية، والتفاعل التاريخي، والنتائج الوسيطة، والمعلومات المتعلقة بالمهام والتي تدعم الاستمرارية، واتخاذ القرارات، خلال التنفيذ.

٧ .مكونات تكامل الأدوات (Tool Integration Components)

يتيح هذا المكون التفاعل مع الأدوات، والتطبيقات، والخدمات، وقواعد البيانات، وواجهات برمجة التطبيقات، وأنظمة الجهة، وبروتوكولات التكامل القياسية الخارجية؛ لدعم تنفيذ المهمات واسترجاع المعلومات. ويشمل ذلك مكونات مثل خادم بروتوكول سياق النموذج (MCP) وبوابات واجهة برمجة التطبيقات، وأدوات الربط، والمكونات الإضافية، ومهيئات الخدمة التي تسمح لنماذج الذكاء الاصطناعي، أو الوكلاء بالوصول الآمن، إلى الأدوات ومصادر البيانات المعتمدة.

يوضح الشكل(٢)، بعض مكونات أنظمة الذكاء الاصطناعي التوليدي والتوكيلي؛ والحاجة إلى تفاعلها خارج أنظمة الذكاء الاصطناعي.

شكل ٢: بعض مكونات أنظمة الذكاء الاصطناعي التوليدي والتوكيلي

نظرة عامة عن أبرز تهديدات الأمن السيبراني للذكاء الاصطناعي

مع التوسع المتسارع في تبني تقنيات الذكاء الاصطناعي، والاستفادة من قدراتها المتقدمة في إنجاز الأعمال، برزت تهديدات سيبرانية جديدة، يجب التعامل معها، والتخفيف من آثارها. وبسبب اعتماد أنظمة الذكاء الاصطناعي على كميات هائلة من البيانات، والخوارزميات المعقدة، وبيئات حوسبة مترابطة؛ فإنها بذلك تُنشئ أسطح هجوم جديدة، التي يُمكن للمهاجمين استغلالها. وتؤدي مواطن الضعف الجديدة إلى تسميم البيانات، وعكس النماذج، وهجمات التهرب، وهجمات اختراقات سلاسل الإمداد لمجموعات بيانات التدريب، أو مستودعات النماذج. ونظرًا إلى أن مخرجات أنظمة الذكاء الاصطناعي وقراراته، قد تؤثر بشكل مباشر في العمليات، والخدمات الحيوية؛ فإن أي تلاعب أو تسريب للنماذج الأساسية، قد يؤدي إلى معلومات مضللة، أو إجراءات غير مصرح بها، أو فقدان لسرية المعلومات الحساسة، أو سلامتها، أو توافرها. وبسبب هذه الخطورة؛ فإنه يتطلب الحماية من مثل هذه التهديدات، واتباع نهجًا شاملًا يتضمن معالجة البيانات بشكل آمن، وحوكمة قوية للنماذج، ومراقبة مستمرة للسلوك غير الاعتيادي، وإجراء اختبارات دورية لتقييم قدرة الأنظمة على مقاومة هجمات الأمن السيبراني. كما ينبغي أن تتكامل هذه التدابير مع ضوابط الأمن السيبراني، الصادرة عن الهيئة، مثل الضوابط الأساسية للأمن السيبراني. ويبين الشكل (٣) عددًا من أبرز التهديدات والهجمات التي قد تستهدف أنظمة الذكاء الاصطناعي.

شكل ٣: تهديدات الأمن السيبراني للذكاء الاصطناعي

مكونات وهيكلة إرشادات الأمن السيبراني للذكاء الاصطناعي

المكونات الأساسية والفرعية

يوضح الجدول (١) الآتي المكونات الأساسية والفرعية لإرشادات الأمن السيبراني للذكاء الاصطناعي

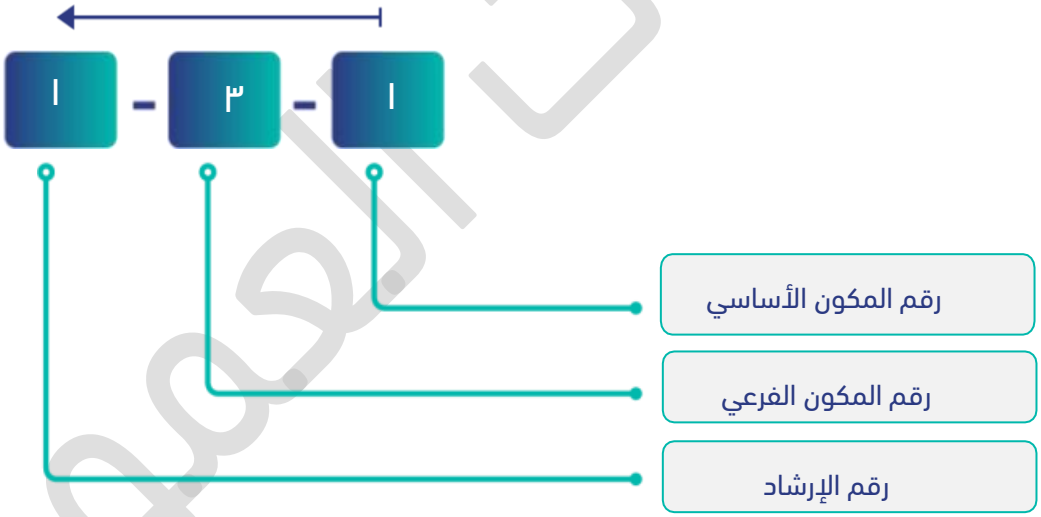
١- حوكمة الأمن السيبراني Cybersecurity Governance	١-١	إدارة مخاطر الأمن السيبراني Cybersecurity Risk Management	١-٢	الأمن السيبراني ضمن إدارة المشاريع المعلوماتية والتقنية Cybersecurity in Information and Technology Project Management
	١-٣	الأمن السيبراني المتعلق بالموارد البشرية Cybersecurity in Human Resources	١-٤	برنامج التوعية والتدريب بالأمن السيبراني Cybersecurity Awareness and Training Program
٢- تعزيز الأمن السيبراني Cybersecurity Defense	٢-١	إدارة الأصول Asset Management	٢-٢	إدارة هويات الدخول والصلاحيات Identity and Access Management
	٢-٣	إدارة أمن الشبكات Network Security Management	٢-٤	حماية البيانات والمعلومات Data and Information Protection
	٢-٥	إدارة النسخ الاحتياطية Backup and Recovery Management	٢-٦	إدارة الثغرات Vulnerability Management
	٢-٧	اختبار الاختراق Penetration Testing	٢-٨	إدارة سجلات الأحداث ومراقبة الأمن السيبراني Cybersecurity Event Logs and Monitoring Management
	٢-٩	حمايات تطبيقات الويب Web Application Security		
٣- صمود الأمن السيبراني Cybersecurity Resilience	٣-١	جوانب صمود الأمن السيبراني في إدارة استمرارية الأعمال Cybersecurity Resilience Aspects of Business Continuity Management (BCM)		
٤- الأمن السيبراني المتعلق بالأطراف الخارجية Third-Party Cybersecurity	٤-١	الأمن السيبراني المتعلقة بالأطراف الخارجية Third-Party Cybersecurity		

جدول ١: المكونات الأساسية والفرعية لإرشادات الأمن السيبراني للذكاء الاصطناعي

هيكلية وثيقة إرشادات الأمن السيبراني للذكاء الاصطناعي

يوضح الشكلان (٤) و(٥) أدناه معنى رموز إرشادات وهيكلية الإرشادات؛ على التالي.

شكل ٤: معنى رموز إرشادات الأمن السيبراني للذكاء الاصطناعي



شكل ٥: هيكلية إرشادات الأمن السيبراني للذكاء الاصطناعي

طريقة هيكلية إرشادات الأمن السيبراني للذكاء الاصطناعي

يوضح الجدول ٢ أدناه طريقة هيكلية إرشادات الأمن السيبراني للذكاء الاصطناعي.

I	اسم المكون الأساسي
رقم مرجعي للمكون الفرعي	اسم المكون الفرعي
الهدف	
الإرشادات	
رقم مرجعي للإرشاد	بنود الإرشاد

جدول ٢: هيكلية الإرشادات

إرشادات الأمن السيبراني للذكاء الاصطناعي

١ حوكمة الأمن السيبراني (Cybersecurity Governance)

١-١	إدارة مخاطر الأمن السيبراني (Cybersecurity Risk Management)
الهدف	ضمان إدارة مخاطر الأمن السيبراني للذكاء الاصطناعي، وفق منهجية واضحة؛ لحماية الأصول المعلوماتية والتقنية للجهة، وذلك وفقًا للسياسات والإجراءات التنظيمية، والأنظمة والتشريعات ذات العلاقة.
الإرشادات	
١-١-١	تحديد ممارسات إدارة مخاطر الأمن السيبراني، لجميع مراحل دورة حياة أنظمة الذكاء الاصطناعي، وتوثيقها واعتمادها، ومن ثم تنفيذها.
١-١-٢	إجراء تقييم لمخاطر الأمن السيبراني؛ بهدف تصنيف وكلاء أنظمة الذكاء الاصطناعي، وفقًا للتأثير المحتمل، واحتمالية الحدوث، وإمكانية التراجع، وتطبيق تدابير الأمن السيبراني المناسبة.
١-٢	الأمن السيبراني ضمن إدارة المشاريع المعلوماتية والتقنية (Cybersecurity in Information and Technology Project Management)
الهدف	ضمان تضمين متطلبات الأمن السيبراني للذكاء الاصطناعي، في ممارسات إدارة المشاريع المعلوماتية والتقنية من خلال كافة مراحل دورة حياة أنظمة الذكاء الاصطناعي.
الإرشادات	
١-٢-١	تطبيق ممارسات التصميم الآمن (security by design)، وتحديد إجراءات وتنفيذها؛ لتأمين تطوير الذكاء الاصطناعي وتشغيله بدءًا من تحديد متطلبات الأمن السيبراني، والحصول على البيانات من مصادر موثوقة وآمنه. والعمل على تدريب النماذج وضبطها، وإعداد إدارة الأوامر (prompt configurations)، وتقييم النماذج، ونشرها، وتشغيلها، وصيانتها، وانتهاءً بإيقافها والتخلص الآمن منها.
١-٢-٢	التعامل مع البيانات المسترجعة من الأدوات الخارجية، أو قواعد الذاكرة، أو واجهات برمجة التطبيقات (APIs) التابعة لأطراف خارجية باعتبارها غير موثوقة، ومن ثم تطبيق عمليات تحقق وتطهير على أي بيانات يتم إدخالها في نافذة السياق (Context window) لمنع تنفيذ الأوامر الخبيثة المتخفية في صورة بيانات.
١-٢-٣	تجنب إعادة تدريب النماذج، أو التعلّم العشوائي، في بيئة التشغيل الفعلية؛ من غير وضع ضوابط مباشرةً؛ ووجود المبررات اللازمة، على أن يتم إجراء تقييم لمخاطر الأمن السيبراني.
١-٢-٤	تقييم قدرات وصمود الأمن السيبراني لوكلاء الذكاء الاصطناعي، خلال تطويره وتشغيله.
١-٢-٥	إجراء مراجعات للأمن السيبراني، واختبارات الأمن السيبراني، والاعتماد الرسمي بصورة منهجية، لجميع الأكواد البرمجية المُولَّدة بالذكاء الاصطناعي؛ لمعالجة مخاطر الأمن السيبراني، وذلك قبل استخدامها في بيئات التطوير(Development)، أو التكامل (Integration) أو الإنتاج (Production) .

١-٢-٦	تحديد إجراءات الموافقة على مشاريع أنظمة الذكاء الاصطناعي ذات التأثير العالي؛ وتنفيذها؛ قبل نشرها على بيئة الإنتاج، مع ضمان تقييم مخاطر الأمن السيبراني، ومعالجتها.
١-٢-٧	تطبيق استقلالية وكيل الذكاء الاصطناعي بشكل متدرج؛ مع ضمان الإشراف البشري المستمر لمراقبة القرارات والإجراءات المتعلقة بالأمن السيبراني وفهمها.
١-٣	**الأمن السيبراني المتعلق بالموارد البشرية (Cybersecurity in Human Resources)**
الهدف	ضمان إدارة مخاطر الأمن السيبراني ومتطلباته للذكاء الاصطناعي، المتعلقة بالعاملين (موظفين ومتعاقدين) في الجهة بفعالية قبل عملهم، وأثنائه، وعند انتهائه، وذلك وفقًا للسياسات والإجراءات التنظيمية للجهة، والمتطلبات التشريعية والتنظيمية ذات العلاقة.
الإرشادات	
١-٣-١	إجراء المسح الأمني للعاملين، الذين لديهم إمكانية الوصول لمكونات أنظمة الذكاء الاصطناعي الداخلية أو العاملين الذين لديهم صلاحيات الوصول الهامة والحساسة، لبيئات تشغيل الذكاء الاصطناعي، أو غيرها من مهام الذكاء الاصطناعي، ذات المخاطر المرتفعة.
٢-٣-١	ضمان الالتزام بالسرية، والاستخدام المقبول، والتعامل الآمن، لجميع العاملين الذين قد تكشف أدوارهم عن بيانات حساسة، أو عن الأوامر، أو المخرجات، أو الإعدادات لأنظمة الذكاء الاصطناعي.
١-٤	**برنامج التوعية والتدريب بالأمن السيبراني (Cybersecurity Awareness and Training Program)**
الهدف	ضمان تلقي المختصين والمستخدمين، التوعية والتدريب الخاص بالأمن السيبراني للذكاء الاصطناعي، بما يتناسب مع أدوارهم، ومخاطر الأمن السيبراني المرتبطة بأنظمة الذكاء الاصطناعي.
الإرشادات	
١-٤-١	تضمين متطلبات الأمن السيبراني للذكاء الاصطناعي، في برامج تدريب العاملين، المشاركين في تطوير أنظمة الذكاء الاصطناعي، أو إعداده أو تشغيله، بما يشمل ذلك: • التصميم الآمن للأوامر (Prompt design) وتصميم الأدوات، والتعامل مع المخرجات، والإبلاغ عن حوادث الأمن السيبراني. • التعامل الآمن مع البيانات الحساسة في الأوامر (Prompts) والملفات المرفوعة، والتعرف على تهديدات الأمن السيبراني، المرتبطة بالذكاء الاصطناعي (مثل حقن الأوامر "Prompt Injection" أو تسريب البيانات)، والتحقق من المخرجات، بحثًا عن مخاطر الأمن السيبراني المحتملة، بالإضافة إلى التوعية بتدابير الأمن السيبراني ذات العلاقة، للحماية من تلك المخاطر.
٢-٤-١	تضمين التوعية بالأمن السيبراني لأنظمة الذكاء الاصطناعي، ضمن برنامج التوعية الشامل للأمن السيبراني في الجهة، ويشمل ذلك:

	• الاستخدام الآمن للأوامر (Prompts) المدخلة من المستخدم، عند التفاعل مع نظام الذكاء الاصطناعي؛ بما في ذلك تجنب إدخال معلومات حساسة في البيانات المدخلة، والذي قد ينتج عنه مخاطر أمن سيبراني، من مثل تسريب البيانات. • مخاطر الأمن السيبراني المرتبطة بالبرمجيات المولدة بواسطة الذكاء الاصطناعي، وضوابط المراجعة والاختبارات اللازمة قبل الاستخدام.

٢ تعزيز الأمن السيبراني (Cybersecurity Defense)

٢-١	إدارة الأصول (Asset Management)
الهدف	ضمان امتلاك الجهة، سجلًا دقيقًا ومفصلًا للأصول المعلوماتية والتقنية، المرتبطة بالذكاء الاصطناعي؛ بما يدعم متطلبات الأمن السيبراني، والمتطلبات التشغيلية، ويحافظ على سرية تلك الأصول، وسلامتها، وتوافرها.
الإرشادات	
٢-١-١	إنشاء قائمة جرد للأصول الخاصة بجميع مكونات نظام الذكاء الاصطناعي، وتحديثها بشكل مستمر.
٢-١-٢	تحديد ملاك الأصول وحال دورة حياة الأصول، لجميع مكونات نظام الذكاء الاصطناعي؛ بما في ذلك الأصول التجريبية، وأصول ما قبل النشر والاطلاق، وأصول بيئة الانتاج، والأصول المقيدة، والأصول التي تم الانتهاء من استخدامها، وكذلك الخدمات، والإعدادات.
٢-١-٣	تحديد وتنفيذ الإجراءات المتعلقة بالأصول والمكونات بعد الانتهاء من استخدام أنظمة الذكاء الاصطناعي ولمرحلة ما بعد التشغيل، مثل إجراءات التخلص الآمن من تلك الأصول.
٢-٢	إدارة هويات الدخول والصلاحيات (Identity and Access Management)
الهدف	ضمان الوصول المنطقي (Logical access) الآمن، والمقيد إلى الأصول المعلوماتية والتقنية، المرتبطة بالذكاء الاصطناعي، وذلك لمنع الوصول غير المشروع، وقصر صلاحيات الوصول على المستخدمين، والخدمات، ووكلاء الذكاء الاصطناعي (AI agents) المصرّح لهم فحسب، وعند الحاجة لإنجاز المهمات المسندة إليهم.
الإرشادات	
٢-٢-١	تحديد الصلاحيات، وسياسات التصاريح، وفقًا لمهمات المستخدم، وذلك لحسابات المستخدمين البشريين، وحسابات الخدمة، وحسابات وكلاء الذكاء الاصطناعي، وضمان تطبيق مبادئ التحكم بالدخول والصلاحيات (مبدأ فصل المهام، ومبدأ الحد الأدنى من الصلاحيات والامتيازات، ومبدأ الحاجة إلى المعرفة والاستخدام).
٢-٢-٢	إجراء المراجعة الدورية للهويات، وصلاحيات الوصول، والرموز، وحسابات الخدمة المستخدمة؛ من قبل أنظمة الذكاء الاصطناعي.
٢-٢-٣	وجود القدرة على التحكم في الصلاحيات بما يخدم التخفيف من المخاطر المتعلقة بالذكاء الاصطناعي بما في ذلك تحديد المعدل (rate limiting) وتحديد حدود الاستخدام (usage threshold) وآليات أخرى؛ للتخفيف من مخاطر استخراج النماذج، وتسريب البيانات.
٢-٣	إدارة أمن الشبكات (Network Security Management)
الهدف	ضمان الأمن السيبراني لمسارات الشبكة، والواجهات، والاتصالات المستخدمة في أنظمة الذكاء الاصطناعي؛ بما في ذلك حركة البيانات، المتعلقة بالتدريب، والضبط والاسترجاع والاستدلال، وكذلك الاتصالات مع الأدوات، والخدمات الخارجية.

الإرشادات	
٢-٣-١	عزل شبكة المكونات الداخلية لأنظمة الذكاء الاصطناعي، عن المكونات الخارجية، وتقييد البروتوكولات، والمنافذ، والوجهات، ومسارات الخروج (egress path) بالمتطلبات المعتمدة فحسب؛ وذلك لفرض حدود الثقة، واكتشاف الحالات غير العادية.
٢-٣-٢	عزل شبكة المكونات الخارجية لأنظمة الذكاء الاصطناعي، وتقييد الوصول والاتصالات الأخرى إليها، وفق الحاجة التشغيلية فحسب، مع تطبيق الضوابط المناسبة لحماية الأنظمة المواجهة للإنترنت، وتقليل سطح الهجوم و اكتشاف الحالات غير العادية.
٢-٤	**حماية البيانات والمعلومات (Data and Information Protection)**
الهدف	ضمان حماية البيانات، والمعلومات التي تستخدمها أنظمة الذكاء الاصطناعي، أو تُنتجها، أو ترتبط بها بأي صورة كانت، وذلك على امتداد دورة حياتها.
الإرشادات	
٢-٤-١	حماية سرية بيانات مكونات أنظمة الذكاء الاصطناعي، وسلامتها، ويشمل ذلك بيانات التدريب، وبيانات التقييم، وأوزان النموذج، والأوامر (Prompts) والتضمينات، والسجلات، والسياقات، والمخرجات في حالة التخزين، وأثناء النقل، وأثناء المعالجة، وتقييد الوصول لها والكشف عنها؛ بحيث يقتصر الوصول إليها على الموظفين، المصرح لهم فحسب.
٢-٤-٢	حماية البيانات المصنفة ضمن أنظمة الذكاء الاصطناعي، في حالة التخزين، وأثناء النقل، وأثناء المعالجة؛ وفقًا لتصنيفها حسب المتطلبات التشريعية والتنظيمية ذات العلاقة.
٢-٤-٣	حماية مكونات ذاكرة الذكاء الاصطناعي، من التعديل غير المشروع، والتحقق من صحة المحتوى المخزن؛ لمنع تسميم الذاكرة والتلاعب بها.
٢-٤-٤	تحديد ضوابط (Guardrails) للكشف عن المدخلات الخبيثة وتنفيذها، وكذلك تحديد الإجراءات غير المصرح بها، والكشف عن المعلومات الحساسة، وأي تهديدات أخرى للأمن السيبراني، وضمان إجراء الاختبارات، قبل النشر، وبعد التغييرات الجوهرية.
٢-٥	**إدارة النسخ الاحتياطية (Backup and Recovery Management)**
الهدف	ضمان وجود ترتيبات آمنة للنسخ الاحتياطية، والاستعادة، والرجوع إلى الإصدارات السابقة (Rollback) لأنظمة الذكاء الاصطناعي.
الإرشادات	
٢-٥-١	تطبيق آليات آمنة للنسخ الاحتياطي والاستعادة؛ لضمان توافر أنظمة الذكاء الاصطناعي، وموثوقيتها، مع ضمان استعادتها ضمن الوقت المحدد.
٢-٥-٢	اختبار إجراءات النسخ الاحتياطي، والاستعادة، والرجوع إلى الإصدارات السابقة (Rollback) لأنظمة الذكاء الاصطناعي بشكل دوري.

٦-٢	إدارة الثغرات (Vulnerability Management)
الهدف	ضمان تحديد الثغرات الأمنية، التي تؤثر على نماذج الذكاء الاصطناعي، والتطبيقات، والبنية التحتية، والاعتماديات (Dependencies) وعمليات التكامل (Integrations) وتقييمها، ومعالجتها، والإبلاغ عنها في الوقت المناسب.
الإرشادات	
١-٦-٢	تحديد الثغرات الأمنية في أنظمة الذكاء الاصطناعي، وتتبعها، ومعالجتها.
٢-٦-٢	التواصل والاشتراك، مع خدمات المعلومات الاستباقية، للثغرات، والتهديدات على أنظمة الذكاء الاصطناعي، مع مراقبة التوصيات الأمنية الصادرة من المطورين، والتحديثات على حالة الخدمات، وتنبيهات الأمن السيبراني.
٣-٦-٢	تعطيل أو تقييد خصائص الذكاء الاصطناعي؛ عند اكتشاف ثغرة أمن سيبراني، إلى أن يتم ضمان المعالجة والتحقق منها.
٧-٢	اختبار الاختراق (Penetration Testing)
الهدف	تقييم مدى فعالية إرشادات الأمن السيبراني، التي تحمي أنظمة الذكاء الاصطناعي؛ وذلك من خلال محاكاة أساليب الهجوم، والتهديد ذات الصلة؛ لتحديد الثغرات الأمنية القابلة للاستغلال، والأخطاء في الإعدادات (Misconfigurations) ونقاط الضعف قبل مرحلة الإطلاق (Deployment) وبعد ذلك تباعًا، بما يتناسب مع مستوى المخاطر السيبرانية لهذه الأنظمة، وتعقيدها، ومدى تعرضها للتهديدات.
الإرشادات	
١-٧-٢	إجراء اختبارات اختراق لمكونات نظام الذكاء الاصطناعي، واختبارات اختراق مخصصة للذكاء الاصطناعي؛ قبل النشر وبشكل دوري، وبما يضمن اختبار المخرجات، وأداء النظام.
٢-٧-٢	إجراء اختبار لأنظمة الذكاء الاصطناعي من خلال طرق مختلفة من مثل؛ اختبارات الفريق الأحمر (Red-teaming) وإجراء تقييم مقاومة الهجمات لأنظمة الذكاء الاصطناعي (Adversarial evaluation)، بالإضافة إلى التحقق من صحة إجراءات التجاوز البشري (Human override).
٣-٧-٢	معالجة الملاحظات والثغرات المكتشفة من خلال اختبارات وتقييمات الأمن السيبراني، وتحديد آليات التحصين المناسبة.
٨-٢	إدارة سجلات الأحداث ومراقبة الأمن السيبراني (Cybersecurity Event Logs and Monitoring Management)
الهدف	ضمان تغطية قدرات التسجيل، والمراقبة، والكشف عن الأحداث، والإجراءات، والتغييرات، والسلوكيات غير الطبيعية (Anomalies) الخاصة بالذكاء الاصطناعي؛ اللازمة لدعم الالتزام، والاستجابة لحوادث الأمن السيبراني، وضمان التحقق المستمر.

الإرشادات	
٢-٨-١	تسجيل أحداث الأمن السيبراني، على مكونات أنظمة الذكاء الاصطناعي ومراقبتها؛ بما في ذلك أحداث هويات الدخول والصلاحيات، والأحداث الهامة والحساسة على الأصول، والوصول إلى مجموعات البيانات، ومصادر المعرفة، والتغيير على إعدادات، وتفعيل المكونات والأدوات.
٢-٨-٢	تسجيل أحداث الأمن السيبراني المرتبطة بالإجراءات التي يتخذها أنظمة الذكاء الاصطناعي ومراقبة تلك الأحداث؛ بما في ذلك عمليات تحميل البيانات وتنزيلها، وإنشاء البيانات أو تعديلها أو حذفها، وأي إجراءات قد تؤثر على السلامة أو السرية أو التوافر.
٢-٨-٣	حماية سجلات الذكاء الاصطناعي، ذات الصلة بالأمن السيبراني، والاحتفاظ بها، وتحليلها؛ لدعم متابعة الالتزام، والاستجابة لحوادث الأمن السيبراني، ومعالجة ثغرات الأمن السيبراني، والتحقيق فيها.
٢-٩	**حماية تطبيقات الويب (Web Application Security)**
الهدف	ضمان تصميم التطبيقات المدعومة بالذكاء الاصطناعي، وواجهات برمجة التطبيقات (APIs) وواجهات الويب، والوكلاء (Agents) والمكونات البرمجية (Plugins) وعمليات التكامل المرتبطة بها، وتطويرها، وإعدادها، وتشغيلها بصورة آمنة.
الإرشادات	
٢-٩-١	حماية تطبيقات وخدمات الويب المرتبطة بأنظمة الذكاء الاصطناعي، وذلك باستخدام بروتوكولات اتصال آمنة، وتحديد معدل الطلبات، والحماية من هجمات حجب الخدمة، ونقاط خروج مُتحكم بها، ومرشحات لحظر القنوات غير الآمنة.

3 صمود الأمن السيبراني (Cybersecurity Resilience)

١-٣	جوانب صمود الأمن السيبراني في إدارة استمرارية الأعمال Cybersecurity Resilience Aspects of Business Continuity Management (BCM)
الهدف	ضمان توافر متطلبات صمود الأمن السيبراني في إدارة استمرارية أعمال الجهة، وضمان معالجة الآثار المترتبة على أنظمة الذكاء الاصطناعي، نتيجة الكوارث الناتجة عن المخاطر السيبرانية، وذلك بما يتناسب مع درجة أهميتها، ومدى الاعتماد عليها.
الإرشادات	
١-١-٣	تحديد إجراءات الرجوع إلى الإصدارات السابقة (Rollback) والإيقاف الآمن للتشغيل (Safe shutdown) وتحديد البدائل لأنظمة الذكاء الاصطناعي، عند وقوع حادثة أمن سيبراني، ما أمكن ذلك.

4 الأمن السيبراني المتعلق بالأطراف الخارجية (Third-Party Cybersecurity)

١-٤	الأمن السيبراني المتعلق بالأطراف الخارجية Third-Party Cybersecurity
الهدف	ضمان تحديد مخاطر الأمن السيبراني الناشئة عن جهات، أو مكونات، أو خدمات الأطراف الخارجية، المستخدمة في دورة حياة الذكاء الاصطناعي، وتقييمها، وإدارتها، ومراقبتها باستمرار.
الإرشادات	
١-١-٤	تقييم ومعالجة مخاطر الأمن السيبراني، التي قد تنشأ من استخدام أطراف وجهات خارجية لأنظمة الذكاء الاصطناعي.
٢-١-٤	تقييم الأمن السيبراني لمكونات الذكاء الاصطناعي التابعة لجهات خارجية في مكونات أنظمة الذكاء الاصطناعي؛ قبل الاستخدام، وبشكل مستمر.
٣-١-٤	ضمان التزام مطوري الذكاء الاصطناعي، بمتطلبات الأمن السيبراني للذكاء الاصطناعي الخاصة بالجهة، بموجب العقود، ومراقبة هذا الالتزام وتقييمه بشكل دوري.
٤-١-٤	تحديد متطلبات الخروج والنقل والتخلص أو الاتلاف الآمن لمكونات أنظمة الذكاء الاصطناعي، عند تغيير المطور أو إنهاء العقد.

الملاحق

الملحق (أ) المصطلحات والتعاريف

يبين الجدول رقم ٣ الآتي، بعض المصطلحات التي ورد ذكرها في هذه الوثيقة، وتعريفاتها.

المصطلح	التعريف
وكيل الذكاء الاصطناعي AI Agent	كيان برمجي، قادر على استقبال المدخلات، واستدعاء الأدوات، أو الوظائف الخارجية، والعمل بتحكم ذاتي؛ لتحقيق أهداف محددة.
نظام الذكاء الاصطناعي AI System	بحسب تعريف الهيئة السعودية للبيانات والذكاء الاصطناعي (SDAIA) عُرفت أنظمة الذكاء الاصطناعي بأنها "أنظمة برمجية تعتمد على تقنيات متقدمة، تُمكنها من التنبؤ، أو توليد المحتوى، أو تقديم التوصيات، أو اتخاذ القرارات، بمستويات مختلفة من الاستقلالية، اعتماداً على البيانات والسياق الذي تعمل فيه". كما يشمل هذا المصطلح الخدمات التي تدعمها تقنية الذكاء الاصطناعي، وجميع مكونات نظام الذكاء الاصطناعي (أنظمة الذكاء الاصطناعي التوليدية، والتوكيلية) المحددة في هذه الوثيقة مثل مكونات نماذج الذكاء الاصطناعي، مكونات مركزية البيانات والمعرفة، مكونات الاسترجاع وإدارة السياق، ومكونات طبقة التطبيق، والتفاعل، ومكونات التخطيط، ومكونات الذاكرة، ومكونات تكامل الأدوات.
الذكاء الاصطناعي التوكيلي Agentic AI	بحسب تعريف الهيئة السعودية للبيانات والذكاء الاصطناعي (SDAIA) عُرف **الذكاء الاصطناعي التوكيلي**، بأنه "نظام برمجي يعتمد على خوارزميات الذكاء الاصطناعي ويتسم بخصائص مثل: إدراك البيئة المحيطة وتفسير المعلومات وتحديد الأهداف واتخاذ القرارات ذاتياً دون الحاجة إلى إشراف بشري مستمر".
الذكاء الاصطناعي التوليدي Generative AI	بحسب تعريف الهيئة السعودية للبيانات والذكاء الاصطناعي (SDAIA) عُرف **الذكاء الاصطناعي التوليدي** بأنه " فرع من فروع الذكاء الاصطناعي يعتمد على تقنيات تعلم الآلة والشبكات العصبية العميقة لمحاكاة قدرة الإنسان على إنتاج بيانات ومحتوى أصيل".
التضمينات Embeddings	تمثيلات عددية للبيانات، تستخدمها أنظمة الذكاء الاصطناعي؛ لأغراض استرجاع المعلومات، أو مطابقة التشابه، أو العمليات الأخرى القائمة على النماذج.
ضبط دقيق Fine-tuning	عملية تدريب إضافية أو مواءمة لنموذج ذكاء اصطناعي مُدرب سابقًا، باستخدام بيانات خاصة بمهمة معينة أو بجهة محددة.
استدلال Inference	هي عملية استخدام نموذج مُدرب؛ للتنبؤ، أو اتخاذ القرارات، بناءً على بيانات جديدة.

المصطلح	التعريف
أوزان النموذج **Model Weights**	المعلمات (Parameters) التي اكتسبها نموذج الذكاء الاصطناعي، خلال عملية التعلم، والتي تحدد سلوكه وتؤثر في مخرجاته.
أمر **Prompt**	النص أو التعليمات أو السياق، الذي يجري تزويد نموذج الذكاء الاصطناعي به؛ للتأثير في استجابته، أو الإجراء الذي سيتخذه.
حقن الأوامر **Prompt Injection**	أسلوب هجومي، يستهدف التلاعب بالأوامر، أو السياق، أو المحتوى المسترجع، بهدف تغيير سلوك النموذج، أو تجاوز ضوابط الحماية، أو تحريضه على تنفيذ إجراءات، غير مصرح بها.
بيانات التدريب **Training Data**	البيانات المستخدمة لتدريب نموذج الذكاء الاصطناعي على حال معينة، أو تدريبه مسبقًا، أو ضبط دقته (Fine-tune) أو مواءمته بأي شكل آخر.
الأداة / المكونات البرمجية **Tool / Plugin**	وظيفة خارجية، أو خدمة، أو واجهة، يمكن لنظام الذكاء الاصطناعي استدعاؤها، أو استخدامها أثناء التشغيل.

جدول ٣: المصطلحات والتعريفات

الملحق (ب) قائمة الاختصارات

يوضح الجدول ٤ أدناه الاختصارات المستخدمة في هذه الوثيقة.

الاختصار	معناه
AI	الذكاء الاصطناعي Artificial Intelligence
AICG	إرشادات الأمن السيبراني للذكاء الاصطناعي AI Cybersecurity Guidelines
API	واجهة برمجة التطبيقات Application Programming Interface
BCM	إدارة استمرارية الأعمال Business Continuity Management
ECC	الضوابط الأساسية للأمن السيبراني Essential Cybersecurity Controls
LLM	نموذج اللغة الكبير

الاختصار	معناه
	Large Language Model
MCP	بروتوكول سياق النموذج Model Context Protocol
NCA	الهيئة الوطنية للأمن السيبراني National Cybersecurity Authority
RAG	تقنية التوليد المعزز بالاسترجاع Retrieval-Augmented Generation
TLP	برتوكول الإشارة الضوئية Traffic Light Protocol

جدول ٤: قائمة الاختصارات


الهيئة الوطنية
للأمن السيبراني
National Cybersecurity Authority