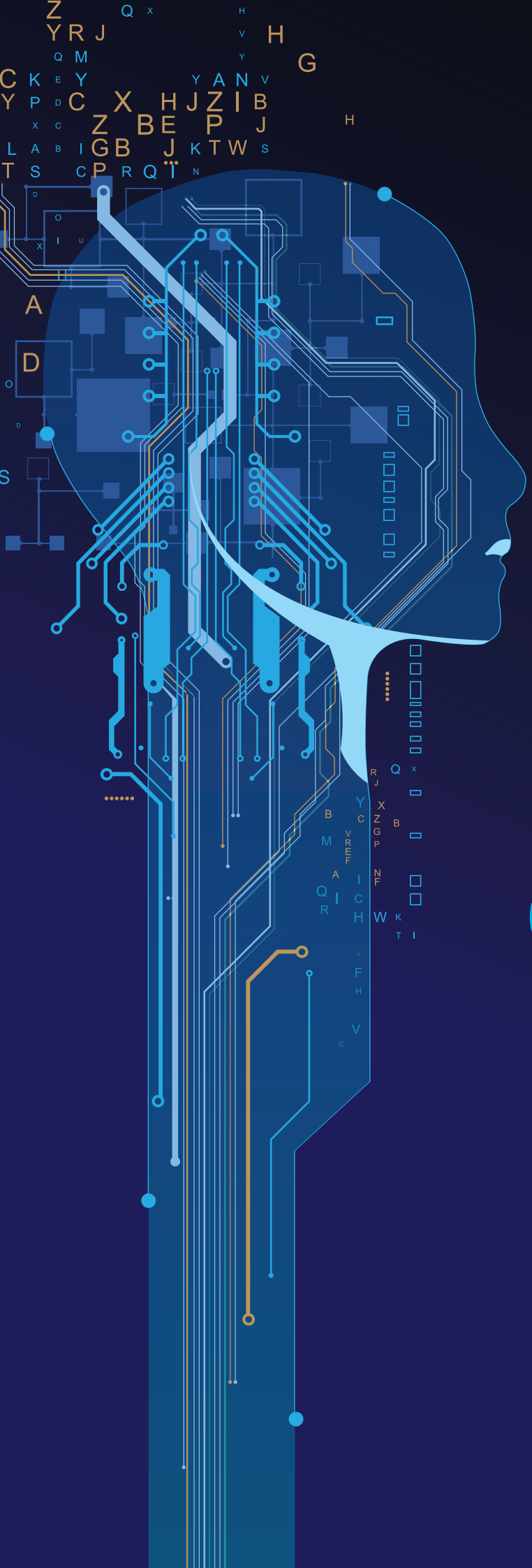


الوكالة الوطنية للأمن السيبراني
National Cyber Security Agency

مبادئ توجيهية للتطبيق والاستخدام الآمن للذكاء الاصطناعي

2024

النسخة 0.1

الوكالة الوطنية للأمن السيبراني
National Cyber Security Agency

إخلاء المسؤولية / الحقوق القانونية

صممت وأصدرت الوكالة الوطنية للأمن السيبراني (NCSA) وثيقة "مبادئ توجيهية للتطبيق والاستخدام الآمن للذكاء الاصطناعي" - الإصدار 0.1 -، لتقديم توجيهات للمؤسسات بشأن كيفية اعتماد واستخدام الذكاء الاصطناعي بشكل آمن.

تتحمل الوكالة الوطنية للأمن السيبراني مسؤولية المراجعة والحفاظ على هذه الوثيقة.

يجب أن يُذكر في أي نسخة من هذه الوثيقة، جزئيًا أو كليًا، وبغض النظر عن وسيلة النسخ، بأن الوكالة الوطنية للأمن السيبراني هي المصدر والمالك لوثيقة "مبادئ توجيهية للتطبيق والاستخدام الآمن للذكاء الاصطناعي".

يجب الحصول على تصريح خطي لأي استخدام لهذه الوثيقة لغرض التسويق التجاري من الوكالة الوطنية للأمن السيبراني. تحتفظ الوكالة الوطنية للأمن السيبراني بالحق في تقييم فعالية وقابلية تطبيق جميع النُسخ التي يتم إصدارها لأغراض تجارية.

لا يجوز تفسير الترخيص من الوكالة الوطنية للأمن السيبراني على أنه مصادقة على النسخ، ولا يجوز لجهة الإصدار بأي حال من الأحوال نشرها أو إساءة تفسيرها في أي وسيلة من وسائل الإعلام أو من خلال مناقشات شخصية أو اجتماعية.

تتحمل المؤسسات بشكل دائم مسؤولية ضمان توافر، وموثوقية، وجودة، وسلامة منتجاتها وخدماتها، بغض النظر عما إذا كانت تكنولوجيات الذكاء الاصطناعي مستخدمة أم لا.

إن اعتماد هذه المبادئ التوجيهية الطوعية لن يعفي المؤسسات من الامتثال للقوانين واللوائح المعمول بها حاليًا. وتجدر الإشارة إلى أن بعض قطاعات الصناعة (مثل قطاعات التمويل والرعاية الصحية والشؤون القانونية) قد يتم تنظيمها من خلال قوانين، أو لوائح، أو توجيهات قائمة ومحددة لقطاع بعينه.

متابعة الوثيقة

تفاصيل الوثيقة	
معرف الوثيقة	[IAG-NAT-SUAI]
النسخة	0.1
التصنيف و النوع	عامة
نبذة	لتقديم إرشادات توجيهية للمؤسسات بشأن كيفية تطبيق واستخدام الذكاء الاصطناعي بشكل آمن

مراجعة/موافقة

الوظيفة/القسم	المراجعة / الاعتماد	النسخة	التاريخ
شؤون الحوكمة و الضمان السيبراني		0.1	فبراير 2024

تاريخ الوثيقة

النسخة	الكاتب	وصف الإصدار	التاريخ
0.1	إدارة سياسات و استراتيجية الأمن السيبراني	منشور	فبراير 2024

التفويض القانوني

حدد المرسوم الأميري رقم (1) لسنة 2021 بشأن إنشاء الوكالة الوطنية للأمن السيبراني اختصاصات الوكالة الوطنية للأمن السيبراني (يشار إليها هنا فيما بعد "الوكالة الوطنية للأمن السيبراني"). تمتلك الوكالة الوطنية للأمن السيبراني سلطة الإشراف على أمن البنية التحتية الوطنية الحيوية، تنظيمها، حمايتها من خلال اقتراح وإصدار السياسات والمعايير وضمان الامتثال لها.

يتم إعداد هذا النوع من الوثائق ليأخذ في الاعتبار قوانين دولة قطر المعمول بها حاليًا. وفي حال وجود تعارض بين هذه الوثيقة وقوانين دولة قطر، تكون الأولوية لقوانين دولة قطر. وعند حذف أي حكم من أحكام هذه الوثيقة، تظل بقية الوثيقة سارية دون التأثير على أحكامها. وفي هذه الحالة، سيتطلب إجراء تعديلات لضمان الامتثال لقوانين دولة قطر ذات الصلة و المعمول بها.

المحتويات

1 المقدمة

شهدت الآونة الأخيرة تطورًا في استخدام تقنيات وتكنولوجيا الذكاء الاصطناعي التي امتدت إلى أعمال رئيسية واستخدامات عامة بسبب تطبيقاتها المتعددة واستخداماتها غير المحدودة، حيث أدركت المؤسسات القوة والتأثيرات التي من المحتمل أن يحدثها الذكاء الاصطناعي في توسيع نطاق الأعمال وتحسين تجربة المستخدم وزيادة مستوى رضا العملاء.

ومع ذلك، فإن الذكاء الاصطناعي كمفهوم كان موجودًا منذ عقود. وقد تم بحث النماذج الأساسية التي تحدد الذكاء الاصطناعي منذ أوائل الخمسينيات، ولكن عوامل مثل:

- تطور التقنيات المتقدمة في التعلم الآلي والشبكات العصبية والتعلم الذكي؛
- توافر مجموعات كبيرة من البيانات لتمكين وتعزيز التدريب؛
- التقدم في الحوسبة عالية الأداء التي تتيح التدريب والتطوير السريع؛

أدت إلى زيادة الاهتمام والاستثمار في الذكاء الاصطناعي، مما أدى إلى تطوير تطبيقات الذكاء الاصطناعي العملية التي تؤثر على المستخدم العام.

ومن الناحية التكنولوجية أصبحنا نتعامل مع الذكاء الاصطناعي دون إدراك أننا نتعامل معه في حياتنا اليومية. ومن الأمثلة على ذلك تطبيقات الرؤية الذكية، أو التعرف على الصوت، أو التطبيقات التي تسمح بإنشاء مقاطع فيديو من صورة واحدة، وتطبيقات الرعاية الصحية عن بعد، وغير ذلك من تقنيات. كما أن العديد من هذه التطبيقات كان يتطلب مهارات و معرفة بالتقنيات إلا أن ذلك تغير مع ظهور تطبيقات الذكاء الاصطناعي التوليدي مثل نظام ChatGPT الذي سمح للمستخدم العادي بطلب المساعدة من الأداة بأسلوب محادثة طبيعي وكسر الحاجز التكنولوجي وأضفى طابع النفاذ لتكون في متناول الشخص العادي.

لقد أدركت المؤسسات والدول منذ فترة طويلة إمكانات تكنولوجيا الذكاء الاصطناعي مما جعلها تعمل بكد واجتهاد لتسخير قوة الذكاء الاصطناعي، فبالنسبة للمؤسسات، يقدم الذكاء الاصطناعي إمكانيات التوسع السريع، ويوفر الرؤى للمؤسسات من البيانات التي كانت تشغل في السابق مجموعات من محركات الأقراص، فضلاً عن توفير فرص اتخاذ قرارات أفضل، وتقديم خدمات أفضل ذات جودة عالية.

وبالنسبة للدول، فيمكن للذكاء الاصطناعي أن يكون مفيدًا في تحقيق الرغبات و الأهداف الوطنية بالمجالات التالية:

- تقديم حوكمة أفضل للمواطنين والمقيمين؛
- تعزيز الاستخدام الفعال للأموال والميزانيات؛
- وضع سياسات أكثر ذكاءً؛
- إدارة الأوبئة والوقاية من الأمراض؛
- تحسين حماية البنية التحتية الحيوية؛
- تعزيز آليات الدفاع والأمن؛

بدأت دولة قطر رحلة الذكاء الاصطناعي من خلال إطلاق الاستراتيجية الوطنية للذكاء الاصطناعي في عام 2019، مع تركيز واضح على العديد من المجالات الرئيسية التي تشمل تثقيف ورفع مستوى وعي الأفراد، بالإضافة إلى بناء إمكانات وقدرات البحث والتطوير والابتكار وحوكمة التكنولوجيا الناشئة، وضمان اتباع نهج وطني مشترك يتماشى مع الأهداف المحددة في رؤية قطر 2030.

وتبع ذلك إنشاء لجنة الذكاء الاصطناعي ضمن وزارة الاتصالات وتكنولوجيا المعلومات، وفقا للأحكام الواردة في قرار مجلس الوزراء رقم (10) لسنة 2021.

تؤكد هذه المبادرات التزام دولة قطر بدمج الذكاء الاصطناعي بشكل استراتيجي في إطار الحوكمة، مما يعكس نهجًا استباقيًا لتسخير الإمكانات التحويلية لتقنيات الذكاء الاصطناعي عبر القطاعات، والبناء على الهدف طويل المدى المتمثل في جعل دولة قطر "اقتصادًا قائمًا على المعرفة".

ومع ذلك، و كما هو الحال مع أي تكنولوجيا أخرى، تأتي أنظمة الذكاء الاصطناعي مع مجموعة من المخاطر الخاصة بها. إذ أن القوة المطلقة والوصول إلى مجموعات البيانات الضخمة يمكن أن يجعلها عرضة لهجمات أكبر. علاوة على ذلك، قد تؤدي معالجة البيانات الشخصية باستخدام الذكاء الاصطناعي إلى مخاطر امتثال كبيرة تؤدي إلى انتهاك لوائح الخصوصية الوطنية والعكس صحيح، إذ يمكن أيضًا استخدام القوة لمهاجمة الآخرين، وعلى هذا النحو صنفت الأمم المتحدة الذكاء الاصطناعي على أنه تكنولوجيا مزدوجة الاستخدام وسلاح ذو حدين. حيث يشير الاستخدام المزدوج إلى التكنولوجيات التي لديها القدرة على تقديم منفعة أكبر للأفراد إذا تم استخدامها بشكل جيد، ولكنها تشكل أيضًا تهديداً كبيراً عليهم إذا لم يتم تنظيمها واستخدامها لرفاهيتهم. على سبيل المثال، العلوم النووية، حيث يمكن للطاقة النووية أن تقدم بدائل للطاقة النظيفة وبالتالي تساند مكافحة التغيرات المناخية، وفي الوقت نفسه يمكن استخدامها أيضًا لأغراض خبيثة.

تهدف هذه الوثيقة إلى توجيه الجهات المعنية لاعتماد تكنولوجيا الذكاء الاصطناعي بشكل آمن من خلال توضيح أفضل الممارسات، وتحديد المخاطر المحتملة، وتوفير استراتيجيات التخفيف لضمان نظام بيئي آمن يعتمد على الذكاء الاصطناعي.

2 الغرض والنطاق والاستخدام

1.2 الغرض

تهدف اعتبارات وتوصيات أمن المعلومات المنصوص عليها في هذه المبادئ التوجيهية إلى توجيه المؤسسات التي قررت استخدام تكنولوجيات الذكاء الاصطناعي ودمجها في أعمالها.

تركز المبادئ التوجيهية بشكل أساسي على المجالات العامة التالية:

المجال الأول: بناء ثقة الجهات المعنية بالذكاء الاصطناعي من خلال الاستخدام المسؤول للمؤسسات للذكاء الاصطناعي لإدارة المخاطر المتعلقة بأمن المعلومات والاستخدام العادل في تطبيق الذكاء الاصطناعي.

المجال الثاني: تقديم توجيهات بشأن القضايا الرئيسية التي يجب مراعاتها والتدابير التي يمكن تنفيذها للاستخدام المسؤول.

المجال الثالث: تقديم توجيهات محددة بشأن الذكاء الاصطناعي التوليدي مع التركيز على تهديداته وحلول التخفيف المحتملة.

2.2 النطاق والاعتبارات

تُعد هذه الوثيقة بمثابة توجيهات وإرشادات وطنية، ويغطي نطاقها جميع المؤسسات الخاصة والحكوميه في دولة قطر التي تستخدم أو تنوي تقديم أنظمة أو خدمات أو منتجات الذكاء الاصطناعي. يتم التركيز بشكل أساسي على مستخدمي الأعمال الذين يقومون بدمج أو اعتماد حلول الذكاء الاصطناعي ضمن أنظمة الأعمال وتكنولوجيا المعلومات الحالية.

و بشكل أساسي، فإن للأمن السيبراني والذكاء الاصطناعي ثلاثة أبعاد رئيسية:

1. **الأمن السيبراني للذكاء الاصطناعي:** يتضمن هذا البعد تقييم وإدارة مخاطر أمن المعلومات لنظام الذكاء الاصطناعي. يشمل نظام الذكاء الاصطناعي التطبيق (الواجهة الأمامية، والواجهة الخلفية، وقواعد البيانات، وغيرها)، والبنية التحتية الأساسية (الأجهزة والشبكات)، ونماذج الذكاء الاصطناعي والخوارزميات الأساسية.

    أ. **النطاق الضيق:** الحماية من الهجمات للحفاظ على سرية وسلامة وتوافر الأصول طوال دورة حياة نظام الذكاء الاصطناعي.

    ب. **النطاق الواسع:** استكمال النطاق الضيق بإضافة ميزات موثوقة مثل: جودة البيانات، الرقابة، الدقة، قابلية التفسير، الشفافية، إمكانية التتبع، وخصوصية البيانات.

2. **الذكاء الاصطناعي لمساندة الأمن السيبراني:** يُستخدم الذكاء الاصطناعي كأداة لتعزيز القدرات وإنشاء أدوات متقدمة للأمن السيبراني والتي قد تسهل، على سبيل المثال لا الحصر، الكشف المتقدم عن التهديدات، والتحليل السلوكي، والتحليل التنبؤي، وسرعة الاستجابة.
3. **الاستخدام الضار للذكاء الاصطناعي:** و هو الاستخدام العدائي للذكاء الاصطناعي للمساعدة على أو إنشاء هجمات سيبرانية معقدة من خلال جهات تهديد ضارة. على سبيل المثال لا الحصر، مقاطع الفيديو المزيفة العميقة (Deep Fake) ، والتلاعب الآلي بوسائل التواصل الاجتماعي، والهجمات الإلكترونية المدعومة بالذكاء الاصطناعي، وما إلى ذلك.

وفي سياق هذه الوثيقة، سنركز على الأمن السيبراني لأنظمة الذكاء الاصطناعي.

في حين أن أهداف المبادئ التوجيهية ليست محدودة، إلا أنها في نهاية المطاف محدودة من حيث الشكل والغرض والاعتبارات العملية. ومن المهم أيضاً ملاحظة أن المبادئ التوجيهية قد تم وضعها وفقًا للاعتبارات التالية:

1. **تجنب التفضيل لخوارزميات بعينها:** لن تركز التوجيهات على ذكاء اصطناعي بعينه أو منهجية تحليل بيانات بعينها، ولكنها تنطبق على تصميم، وتطبيق واستخدام الذكاء الاصطناعي بشكل عام. ومع ذلك، ونظرًا للتطورات الأخيرة والاهتمام المتزايد، تحتوي المبادئ التوجيهية على مجموعة من التوصيات الخاصة بشأن الذكاء الاصطناعي التوليدي.
2. **تجنب التفضيل لتكنولوجيا بعينها:** لن تركز المبادئ التوجيهية على أنظمة أو برامج أو أي تكنولوجيات بعينها، وسيتم تطبيقها بغض النظر عن لغة التطوير وطريقة تخزين البيانات.
3. **تجنب التفضيل لقطاع بعينه:** ستكون المبادئ التوجيهية بمثابة مجموعة أساسية من الاعتبارات والتدابير التي يجب على المؤسسات العاملة في أي قطاع اعتمادها. وقد تختار بعض القطاعات أو المؤسسات إدراج اعتبارات وتدابير إضافية أو تعديل هذه المجموعة الأساسية لتلبية احتياجاتها.
4. **تجنب تفضيل حجم ونموذج أعمال بعينه:** لن تركز المبادئ التوجيهية على مؤسسات ذات نطاق أو حجم معين كما ويمكن استخدامها أيضًا من قبل المؤسسات المشاركة في الأنشطة والعمليات فيما بين الشركات أو بين الشركات والمستهلكين، أو في أي نموذج أعمال آخر.

لا تحل محل السياسات هذه المبادئ التوجيهية والمعايير والإرشادات وأفضل ممارسات الأمن السيبراني وأمن المعلومات الحالية، ولكنها تكملها، وتركز على المجالات التي يجب فيها تعديل الهياكل الحالية لتتناسب مع التهديدات والمخاطر الأمنية الجديدة التي يجلبها الذكاء الاصطناعي.

3.2 الاستخدام

إن أنظمة الذكاء الاصطناعي هي مجموعة من الأنظمة الفرعية المختلفة مثل قواعد البيانات، والتعلم الآلي، ووحدات المعالجة القوية، المرتبطة ببعضها البعض من خلال نماذج وخوارزميات الذكاء الاصطناعي الأساسية. إن القواعد الأساسية للأمن السيبراني لا تتغير وتعد هذه الوثيقة بمثابة استكمال للسياسات والمعايير والإرشادات الوطنية الحالية.

إن هذه المبادئ التوجيهية من شأنها أن تساعد المؤسسات على:

- توسيع منظور المؤسسة فيما يتعلق بأمن الذكاء الاصطناعي.
- توسيع نطاق العمليات الأمنية الحالية لتشمل مبادئ الذكاء الاصطناعي المتضمنة في هذه الوثيقة لتوفير ضمان الثقة.
- فهم أهمية بناء الثقة لدى الجهات المعنية ذات الصلة أثناء استخدام أو تطبيق أنظمة الذكاء الاصطناعي في أعمالها.
- مراجعة وتعزيز تدابير الأمن السيبراني وتقييم المخاطر والحد منها.
- فهم دورة حياة منتج الذكاء الاصطناعي والقيمة التي يمكن أن تجلبها هذه التكنولوجيا الجديدة للمؤسسة.

كما يجب على المؤسسات التي تنوي تطبيق أنظمة أو حلول أو منتجات الذكاء الاصطناعي استخدام هذه المبادئ التوجيهية قبل عملية التطبيق وأثناءها. تقدم هذه المبادئ التوجيهية توصيات عملية لتعزيز الممارسات الأخلاقية وضمان التكامل السلس والآمن لتقنيات الذكاء الاصطناعي.

3 المصطلحات الرئيسية

ناشر الذكاء الاصطناعي

يعني الشركات أو الكيانات الأخرى التي تعتمد أو تدمج أو تطبق حلول الذكاء الاصطناعي في عملياتها، مثل عمليات الدعم (على سبيل المثال، معالجة طلبات القروض)، أو الخدمات الأمامية (على سبيل المثال، بوابة التجارة الإلكترونية أو تطبيق نقل الركاب)، أو بيع أو توزيع الأجهزة التي توفر ميزات مدعومة بالذكاء الاصطناعي (مثل الأجهزة المنزلية الذكية).

مقدم حلول الذكاء الاصطناعي

يعني الكيانات التي تطور حلول الذكاء الاصطناعي أو أنظمة التطبيقات التي تستفيد من تكنولوجيا الذكاء الاصطناعي، ولا تشمل فقط المنتجات التجارية الجاهزة والخدمات عبر الإنترنت وتطبيقات الهاتف الجوال وغيرها من البرامج التي يمكن للمستهلكين استخدامها مباشرة، ولكنها تشمل أيضًا تطبيقات "من شركة إلى شركة إلى المستهلك B2B2C"، على سبيل المثال، برامج الكشف عن الاحتيال المدعومة بالذكاء الاصطناعي والتي يتم بيعها للمؤسسات المالية، وتشمل أيضًا الشركات المصنعة للأجهزة والمعدات التي تدمج ميزات مدعومة بالذكاء الاصطناعي في منتجاتها، وتلك التي لا تعد حلولها منتجات مستقلة بذاتها، ولكن من المفترض أن يتم دمجها في المنتج النهائي. تقوم بعض المؤسسات بتطوير حلول للذكاء الاصطناعي خاصة بها ويمكن أن تُقدم الحلول الخاصة بها.

الذكاء الاصطناعي

يعني نظام (أجهزة أو برمجيات أو كليهما) مصمم لتنفيذ أي مهام مرتبطة بالذكاء البشري، بطريقة تحاكي العقل البشري بمستوى معين من الاستقلالية.

اتخاذ القرار الآلي (ADM)

يعني تطبيق الأنظمة الآلية في أي جزء من عملية اتخاذ القرار.

مراقب البيانات

يُعرف قانون حماية خصوصية البيانات الشخصية مراقب البيانات بأنه شخص طبيعي أو اعتباري، سواء كان يعمل منفردًا أو مشتركًا مع آخرين، يحدد كيفية معالجة البيانات الشخصية ويحدد أغراض أي من هذه المعالجات. (المادة 1 من القانون رقم (13) لسنة 2016بشأن حماية خصوصية البيانات الشخصي).

معالج البيانات

يُعَرِف قانون حماية خصوصية البيانات الشخصية معالج البيانات بأنه شخص طبيعي أو اعتباري يعالج البيانات الشخصية لمراقب البيانات. (المادة 1 من القانون رقم (13) لسنة 2016بشأن حماية خصوصية البيانات الشخصية).

التعلم الذكي	هو مجموعة فرعية من التعلم الآلي يستخدم التي تستخدم العصبية، التي تحاكي كيفية تفاعل الخلايا العصبية في الدماغ البشري، مما يسمح بتعلم الأنماط والعلاقات المعقدة داخل البيانات (مثل الصور والنصوص والأصوات وغيرها من البيانات)، ويتطلب قدرًا أقل من التدخل البشري، ويمكن أن يؤدي في الغالب إلى نتائج أكثر دقة من التعلم الآلي التقليدي.
الذكاء الاصطناعي التوليدي (GenAI)	هو نظام ذكي قادر على إنشاء محتوى جديد مثل الصوت والفيديو والصورة والنصوص والرموز الاصطلاحية وما إلى ذلك بناءً على أنماط وهياكل البيانات التي تم التدرب عليها.
وحدة معالجة الرسومات (GPU)	دائرة إلكترونية مصممة لمعالجة وتغيير الذاكرة بسرعة لتسريع إنشاء الصور في مخزن مؤقت للإطارات مخصص لتشغيل جهاز عرض الفيديو. تُستخدم وحدات معالجة الرسومات في الأنظمة المدمجة والهواتف الجوالة وأجهزة الكمبيوتر الشخصية ومحطات العمل ووحدات التحكم في الألعاب.
النموذج اللغوي الكبير (LLM)	هو نوع من الذكاء الاصطناعي متخصص في ابتكار نصوص تشبه النصوص البشرية.
التعلم الآلي (ML)	هو جزء من الذكاء الاصطناعي يحاكي ذكاء أنظمة الحاسوب من خلال تحسين إدراكها، أو معرفتها، أو تفكيرها، أو تصرفاتها بناءً على الخوارزميات التي يتم تدريبها على البيانات.
النموذج الأساسي للوسائط المتعددة(MfM)	هو نوع من الذكاء الاصطناعي التوليدي يمكنه معالجة وإخراج أنواع متعددة من البيانات (مثل النصوص والصور والأصوات).
البيانات الشخصية ذات الطبيعة الخاصة	يُعَرِف قانون حماية خصوصية البيانات الشخصية، البيانات الشخصية ذات الطبيعة الخاصة، بأنها البيانات الشخصية المتعلقة بالأصل العرقي والأطفال والصحة والحالة الجسدية أو النفسية والعقيدة الدينية والعلاقات الزوجية والجرائم الجنائية. يجب فهم التعريف على نطاق واسع. (المادة 16 من القانون رقم (13) لسنة 2016 بشأن حماية خصوصية البيانات الشخصية.)
وحدة معالجة الموترات (TPU)	دائرة متكاملة خاصة بتطبيق ذكاء اصطناعي تعمل على تسريع حسابات وخوارزميات الذكاء الاصطناعي، وقد قامت شركة جوجل بتطويرها خصيصًا للتعلم الآلي للشبكة العصبية لبرنامج TensorFlow.
استخدام الذكاء الاصطناعي	تطوير أو تطبيق نظام، أو منتج أو خدمة ذكاء اصطناعي.

4 المخاطر والتهديدات والتحديات

1.4 المخاطر والتهديدات

تتعرض أنظمة الذكاء الاصطناعي، مثل أي أنظمة معلومات أخرى، لمخاطر الكشف عن المعلومات غير المصرح بها، أو تعديل المعلومات غير المصرح بها أو فقدان سلامة النظام (المنصة)، أو عدم توفر المعلومات، والانتهاكات التنظيمية. إلى جانب ذلك، تشتمل أنظمة الذكاء الاصطناعي أيضًا على مخاطر كبيرة تتمثل في تحيز البيانات، وقيود الشفافية و التفسير[1]. يعرض الشكل التالي التصنيف المحتمل للتهديدات والمخاطر الرئيسية.

(شكل 1) مخاطر، وتهديدات وتحديات الذكاء الاصطناعي، على سبيل المثال لا الحصر

1.1.4 الكشف عن المعلومات غير المصرح بها

تشكل البيانات جوهر أي منصة خاصة بالذكاء الاصطناعي، كما و يزدهر النظام من خلال "التعلم" من مجموعات البيانات الضخمة، وبالتالي تخزن وتعالج منصات الذكاء الاصطناعي كميات هائلة من البيانات، بما فيها البيانات السرية أو الحساسة مثل البيانات الشخصية والبيانات المالية والسجلات الصحية وغيرها. إن الحجم الهائل للبيانات يجعلها هدفًا مثيرًا للاهتمام للجهات الفاعلة الضارة. وعلى هذا النحو، يشكل خرق البيانات تهديداً أمنياً سيبرانياً على هذه الأنظمة.

هناك عدة عوامل مثل الافتقار إلى الصحة السيبرانية كبروتوكولات الأمان الضعيفة والممارسات السيئة لأمن التطبيقات وعدم تدقيق المدخلات والتشفير غير الكافي في الأنظمة الأساسية ونقص المراقبة والتهديدات الداخلية، قد تجعل أنظمة الذكاء الاصطناعي عرضة لهذه المخاطر وقد تكون على شكل خرق/ فقدان للبيانات/سرقة للبيانات.

1 للاطلاع على تعاريف الشفافية وقابلية التفسير، يمكن الرجوع إلى مبادئ الذكاء الاصطناعي الأخلاقية والعادلة بالقسم 1.8 من الملحق المرفق بهذه الوثيقة.

في حالة حدوث خرق للبيانات، يجب على مراقب البيانات أيضًا أن يأخذ في الاعتبار الالتزامات المستمدة من اللوائح الوطنية ذات الصلة[2].

يمثل تجميع البيانات تهديداً آخر يتمثل في احتمالية كشف المعلومات الحساسة لجهات فاعلة ضارة أو غير ضارة. تعتمد أنظمة الذكاء الاصطناعي على البيانات، وأثناء عملياتها، قد تقوم بجمع مجموعات بيانات متعددة و التي قد تبدو معلوماتها في حد ذاتها غير ضارة (أقل حساسية، وبالتالي قد تكون درجات السرية منخفضة، على سبيل المثال البيانات الشخصية التي لا تعتبر ذات طبيعة خاصة، مثل العنوان الوطني)، ومع ذلك، عند دمجها مع نقاط بيانات أخرى، قد يكون لديها القدرة على إنشاء قواعد بيانات تفصيلية قد تكون حساسة فعلًا بطبيعتها[3] (درجة سرية أعلى، على سبيل المثال، إذا تم دمج العنوان الوطني مع معلومات عن الأطفال، أو الأصل العرقي، فإنها تصبح بيانات شخصية ذات طبيعة خاصة). حتى عندما تكون البيانات مسماة بأسماء مستعارة، فقد تكون أنظمة الذكاء الاصطناعي قادرة على استخدام التعرف المتقدم على الأنماط أو دمج مجموعات البيانات لإعادة تحديد هوية الأفراد دون إذن.

أيضًا وبشكل عام، تقوم حلول الذكاء الاصطناعي بتخزين البيانات لفترات طويلة (الاحتفاظ بالبيانات) حتى يتمكن النموذج من الاستمرار في الرجوع إليها وتحليلها ومقارنتها كجزء من قدراته التعليمية والتنبؤية وغيرها من القدرات. يزيد هذا التخزين للبيانات طويل المدى من خطر الكشف غير المصرح به. علاوة على ذلك، يعد هذا انتهاكًا محتملاً للالتزامات الواردة في قانون حماية خصوصية البيانات الشخصية مثل الاحتفاظ بالبيانات وتحديد الأهداف وتقليل البيانات.

و بالمثل، فإن هناك خطر الكشف غير المصرح به من خلال الهجمات على نموذج التعلم الآلي لنظام الذكاء الاصطناعي. ويشار إلى هذا باسم "هجوم استدلال العضوية"، الذي يسمح للمهاجم باكتشاف البيانات المستخدمة لتدريب نموذج تعلم آلي معين. عادةً، حيث يمكن للمهاجم شن هجمات استدلال العضوية دون الحاجة إلى الوصول إلى معلمات نموذج التعلم الآلي، وذلك من خلال مراقبة مخرجاته فقط. يمكن أن تثير هذه الهجمات مخاوف تتعلق بالأمن والخصوصية إذا تم تدريب النموذج على المعلومات الحساسة.

في أنظمة الذكاء الاصطناعي التوليدية، يتمثل الهجوم في قيام المهاجم بالتلاعب بمنبه الإدخال لإثارة سلوك غير مرغوب فيه من نموذج الذكاء الاصطناعي. يكون هذا التلاعب عن طريق الحقن الفوري الذي يشمل كسر الحماية، والتسريب السريع، وتهريب الرمز المميز الذي يمكن أن يؤدي إلى ابتكار الذكاء الاصطناعي لاستجابات غير مناسبة أو تسريب معلومات حساسة (مما يؤدي إلى انتهاك لوائح قانون حماية خصوصية البيانات الشخصية). من الممكن أن تكون هذه الهجمات فعالة تحديدًا عند استخدام أنظمة الذكاء الاصطناعي بجانب أنظمة أخرى أو في سلسلة تطبيقات برمجية.

تلك مجرد أمثلة على التهديدات المختلفةو المحتملة لإدراك خطر فقدان السرية.

2.1.4 تعديل المعلومات غير المصرح بها أو فقدان نزاهة النظام (المنصة)

نظرًا لطبيعة البيانات المتأصلة، فإن الحفاظ على سلامة البيانات من الأمور بالغة الأهمية لأنظمة الذكاء الاصطناعي، حيث إن أي هجوم على جودة البيانات من شأنه أن يؤثر على جودة مخرجات البيانات وفعاليتها في العالم الحقيقي إذ يمكن أن تحدث هذه الهجمات بطرق عديدة.

يمكن للبيانات الضارة، التي يتم إدخالها في بيانات التدريب، أن تتلاعب بسلوك نموذج الذكاء الاصطناعي، مما يؤدي إلى تنبؤات غير صحيحة أو متحيزة، واتخاذ قرارات غير دقيقة أو غير عادلة، ويشار إلى هذا أيضًا باسم **تسمم النموذج**، و قد يكون من الصعب اكتشاف هجمات تسمم النموذج، لأن تلك البيانات يمكن أن تكون غير ضارة للبشر. كما ويعد الكشف عن الهجمات أيضًا معقدًا بالنسبة لحلول الذكاء الاصطناعي التي تشمل مكونات مفتوحة المصدر أو مكونات خارجية أخرى.

يمكن للمهاجمين محاولة إجراء هندسة عكسية لنموذج الذكاء الاصطناعي لتكرار نموذج تعلم آلي معين تم التدرب عليه بناءً على استفساراته واستجاباته. سيستخدم المهاجم سلسلة من الاستعلامات المصممة خصيصًا للنموذج ويستخدم الاستجابات لإنشاء نسخة من نظام الذكاء الاصطناعي المستهدف. يُعرف هذا باسم هجوم **استخراج النماذج** ويمكن

2 على سبيل المثال القانون رقم 13 لسنة 2016 بشأن حماية خصوصية البيانات الشخصية (PDPPL) والقانون رقم 14 لسنة 2014 بإصدار قانون مكافحة الجرائم الإلكترونية.

3 أصدرت الوكالة الوطنية للأمن السيبراني سياسة تصنيف المعلومات الوطنية، النسخة الثالثة.

أن ينتهك حقوق الملكية الفكرية مما يؤدي إلى خسائر اقتصادية كبيرة.

قد تتأثر جودة البيانات أيضًا بسبب **سلسلة توريد البيانات**، والتي تشير إلى جودة البيانات (التي تم الحصول عليها أو معالجتها) المتأثرة بالتهديدات المتعلقة بأنظمة الإدخال، والعمليات المتعلقة بجمع البيانات ومعالجتها من خلال موردين خارجيين.

3.1.4 عدم توفر المعلومات

في أي نظام مركزي للبيانات، يعد توفر المعلومات من المتطلبات الأساسية. كما أن البنية التحتية للذكاء الاصطناعي حالها حال أي نظام تكنولوجيا معلومات آخر معرضة لخطر عدم التوفر بسبب مجموعة عدة عوامل، قد يكون منها عطل الأجهزة في أحد مكونات نظام الذكاء الاصطناعي، سوء تصميم النظام، نقص الموارد، عدم المرونة، أو وجود عمليات تشغيلية رديئة.

يمكن أن تكون هجمات رفض الخدمة (DoS) أحد هذه التهديدات[4]. فمن خلال إغراق البنية التحتية لنموذج الذكاء الاصطناعي بحركة المرور، يمكن للمهاجم أن يجعل خدمة الذكاء الاصطناعي غير قابلة للاستخدام.

4.1.4 انتهاكات الخصوصية

الخصوصية هي حق أساسي، علاوة على ذلك، يحدد قانون حماية خصوصية البيانات الشخصية التزامات الكيانات التي تعالج البيانات الشخصية لحماية خصوصية الأشخاص.

يمكن لأنظمة الذكاء الاصطناعي جمع ومعالجة كميات كبيرة من البيانات، الأمر الذي قد يثير مخاوف تتعلق بالخصوصية وتحديات الامتثال الإلزامي لقانون حماية خصوصية البيانات الشخصية واللوائح الأخرى المعمول بها. يمكن أن يؤدي جمع البيانات الشخصية ومعالجتها وتحليلها على نطاق واسع من خلال أنظمة الذكاء الاصطناعي إلى:

المراقبة والتوصيف: القدرة على تتبع الأفراد و مراقبة سلوكهم، إذ يمكن لتكنولوجيا الذكاء الاصطناعي مثلاً التعرف على الوجوه و بالتالي مراقبة وسائل التواصل الاجتماعي الخاصة بالأفراد، مما يعرض خصوصيتهم و استقلاليتهم للخطر و يحرمهم من حرية التعبير [5].

عدم الامتثال التنظيمي: يمكن أن يكون سياق وتعقيد حلول الذكاء الاصطناعي عائقاً للتأكد من أن معالجة البيانات متوافقة مع قانون حماية خصوصية البيانات الشخصية ومعايير الخصوصية الأُخرى. على سبيل المثال: يتم حذف البيانات الشخصية عندما لا تكون هناك حاجة إليها أو عندما يمارس الأفراد حقوقهم في طلب الحذف؛ حق الأفراد في الوصول إلى البيانات الشخصية؛ وحق الأفراد في الحصول على معلومات عن تقنيات معالجة البيانات الشخصية.

علاوة على ذلك، فإنه من المحتمل أيضًا الكشف عن البيانات الشخصية، من مجموعات كبيرة من بيانات مجهولة المصدر، بسبب تجميع البيانات.

مشاركة البيانات: يمكن أن تشمل منصات الذكاء الاصطناعي التعاون بين أطراف متعددة أو استخدام أدوات وخدمات مقدم حلول الذكاء الاصطناعي. ففي العديد من الحالات، ينتهي الأمر بمعالجي ومراقبي البيانات إلى مشاركة البيانات من خلال أنظمة وحلول ذكاء اصطناعي متعددة، و يمكن أن يتسبب هذا في عمليات نقل البيانات عبر العديد من الجهات المختصة وإلى كيانات أخرى مما يؤدي إلى تحديات الامتثال للخصوصية وانتهاكات تنظيمية.

5.1.4 أخلاقيات وتحيز البيانات

تزدهر أنظمة وخدمات الذكاء الاصطناعي على البيانات. فكلما زادت البيانات، كلما كان من الممكن تدريب النموذج بشكل أفضل على هذه المجموعات من البيانات وتحسينه لتقديم أداء أفضل. تعد جودة البيانات من الأمور الجوهرية في هذا الشأن، ومن الممكن أن تؤدي الجودة الرديئة للبيانات إلى إبعاد النموذج تمامًا عن أهميته وأثره. فإذا كانت

4 أصدرت الوكالة الوطنية للأمن السيبراني المبادئ التوجيهية ذات الصلة في يناير 2023 بعنوان: "مبادئ توجيهية للتأمين ضد هجمات حجب الخدمة الموزعة (DDoS)".

5 هذا النوع من المراقبة والتوصيف يمكن أن يشكل انتهاكًا مباشرًا للحقوق، لأنه قبل معالجة البيانات الشخصية، يجب إبلاغ الفرد وفقًا لأحكام قانون حماية خصوصية البيانات الشخصية.

البيانات المستخدمة لتدريب نماذج الذكاء الاصطناعي غير كاملة، قد يؤدي ذلك إلى نتائج غير دقيقة أو غير متوقعة، وبالتالي يؤدي إلى التحيز.

ويمكن أن يكون لهذه التحيزات أثر كبير، وخاصة في القطاع الحكومي، عند استخدامها في الأنظمة المخصصة لتقديم خدمات عامة. كما أن تنبؤات النظام يمكن أن يكون لها تأثير سلبي على صنع السياسات، حيث يمكن أن تؤدي أيضًا إلى اضطرابات/استياء بين الجمهور الذي قد يتأثر بهذه القرارات.

و بشكل أساسي، فإن هناك نوعين من التحيزات، التحيز المعرفي وتحيز البيانات.

فالتحيز المعرفي يرتبط بحالات خطأ اللاوعي في التفكير التي تؤثر على حكم الفرد وقراراته، وتشمل التمييز أو التحيز ضد عرق معين، أو مجموعة، أو جنس، أو خصائص سكانية معينة، وعادة ما يكون ذلك مجهولًا للشخص المتسم بالتحيز، وأحيانًا يتسرب هذا من خلال البيانات التاريخية المستخدمة لتدريب نظام الذكاء الاصطناعي.

أما **بالنسبة لتحيز البيانات** فيتعلق بالأخطاء الناتجة بشكل أساسي بسبب جودة البيانات، وتشمل عدم استكمال مجموعة البيانات، أو وجود مجموعة بيانات شاذة لا تمثل جميع الجهات المعنية، أو مجموعة بيانات محدودة.

من الأمثلة على هذه التحيزات التي يمكن أن تؤثر سلبًا على المجتمع، التفاوتات في التعرف بين الخلفيات العرقية من خلال أدوات التعرف على الوجه التي تستخدمها السلطات الأمنية، وأدوات التوظيف التي تربط أسماء النساء بأدوار النساء التقليدية، وانتشار المعلومات المضللة ذات الدوافع السياسية والاستخدام الدائم لوجهات النظر العالمية المتحيزة، والتحيز العنصري في المخططات المالية من بين أمور أُخرى.

2.4 التحديات

1.2.4 الحاجة إلى تدابير لبناء الإمكانات والقدرات

لتلبية تطلعات الذكاء الاصطناعي العالمية والوطنية والتنظيمية، فإن هناك حاجة بشكل متزايد إلى المزج الصحيح للإمكانات لترجمة احتياجات العمل إلى متطلبات حلول، بناء تطبيق أنظمة الذكاء الاصطناعي، دمج الذكاء الاصطناعي في العمليات، وتفسير النتائج. تحتاج الإدارة إلى فهم كيفية عمل الذكاء الاصطناعي لدمج وتطبيق واستخدام وصيانة أنظمة الذكاء الاصطناعي بشكل فعال.

كما أن هناك حاجة إلى تنفيذ خطة متعددة الجوانب للتعامل بفاعلية مع هذا التحدي. وعلى المدى القصير والمتوسط، تحتاج جهات اعتماد الذكاء الاصطناعي إلى تدريب القوى العاملة الحالية لتعزيز الخبرات وتضييق فجوة المهارات. و تقوم العديد من الشركات في جميع أنحاء العالم بتدريب المطورين على إنشاء حلول الذكاء الاصطناعي، وعلى فريق تكنولوجيا المعلومات تطبيق هذه الحلول، وعلى الموظفين استخدام الذكاء الاصطناعي في وظائفهم اليومية. ومع ذلك، نحتاج على المدى الطويل إلى تجديد نظامنا التعليمي لضمان تزويد الأطفال الصغار بالتوجيهات المناسبة والمسار الوظيفي المناسب لممارسة مهنة في الذكاء الاصطناعي أو تعتمد عليه.[6]

2.2.4 المشهد التنظيمي الديناميكي

عالميًا، فإن البيئة القانونية والتنظيمية للذكاء الاصطناعي في حالة تغير مستمرة، إذ تتصارع الحكومات والمجتمعات المدنية والأفراد والصناعات مع التأثير المحتمل للذكاء الاصطناعي على البشر والمجتمعات البشرية، وتطرح أسئلة مثل "هل سيسيطر الذكاء الاصطناعي على الوظائف البشرية؟" إلى "هل سيستعبد الذكاء الاصطناعي البشر يومًا ما؟".

تواجه الحكومات في جميع أنحاء العالم حالياً تحدياً لتحقيق التوازن بين الاستفادة من الفوائد المحتملة للذكاء الاصطناعي للبشرية، والسيطرة المحتملة على إساءة استخدامه، لتجنب تطور التكنولوجيا إلى أنظمة مستقلة قد لا يمكن السيطرة عليها.

تقوم المؤسسات الدولية كالأمم المتحدة، الاتحاد الأوروبي، الجمعيات الصناعية، المنظمات غير الربحية والبلدان المعنية بتقييم تكنولوجيا الذكاء الاصطناعي واقتراح لوائح ومعايير وتوجيهات مختلفة لتنظيم ومراقبة هذه التكنولوجيا.

6 استراتيجية قطر الوطنية للذكاء الاصطناعي، وزارة الاتصالات وتكنولوجيا المعلومات 2019.

ونظرًا لأن هذه المتطلبات المتعلقة بالذكاء الاصطناعي قيد التطوير عالميًا، كونها تتطور بسرعة، فمن المتوقع حدوث تغييرات وتحديثات متكررة، و تحتاج المؤسسات إلى متابعة التطورات القانونية وأن تكون جاهزة لتغيير سياساتها وعملياتها بشكل ديناميكي لضمان الامتثال الكامل.

3.2.4 قابلية التدقيق

في المخطط التقليدي للأشياء، فإن برامج الامتثال والتدقيق هي الركيزة التي تحافظ على الضوابط والتوازن وتتأكد من أن الأمور بالحالة التي من المفترض أن تكون عليها.

ومع ذلك، وبالنسبة لأنظمة الذكاء الاصطناعي، ينبثق التحدي من جوانب القلق بشأن إمكانية تفسير الأنظمة، نظرًا لأن هذه الأنظمة بطبيعتها معقدة ويصعب فهمها. وبشكل أكثر تحديدًا، غالبًا ما تكون النماذج الأساسية معقدة وتشمل طبقات عديدة إذ يتم تدريب هذه النماذج على كميات هائلة من البيانات، مما يجعل من الصعب على البشر فهم الأسباب الدقيقة وراء قراراتهم، ويُعرف هذا بتأثير "الصندوق الأسود".

وتجدر الإشارة إلى أن الأدوات التقليدية المتوفرة للتدقيق والامتثال مجهزة لتقييم وتدقيق هذه المشكلة.

4.2.4 طبيعة الاستخدام المزدوج للذكاء الاصطناعي: البرامج الضارة التي تم إنشاؤها وتشغيلها من خلال الذكاء الاصطناعي

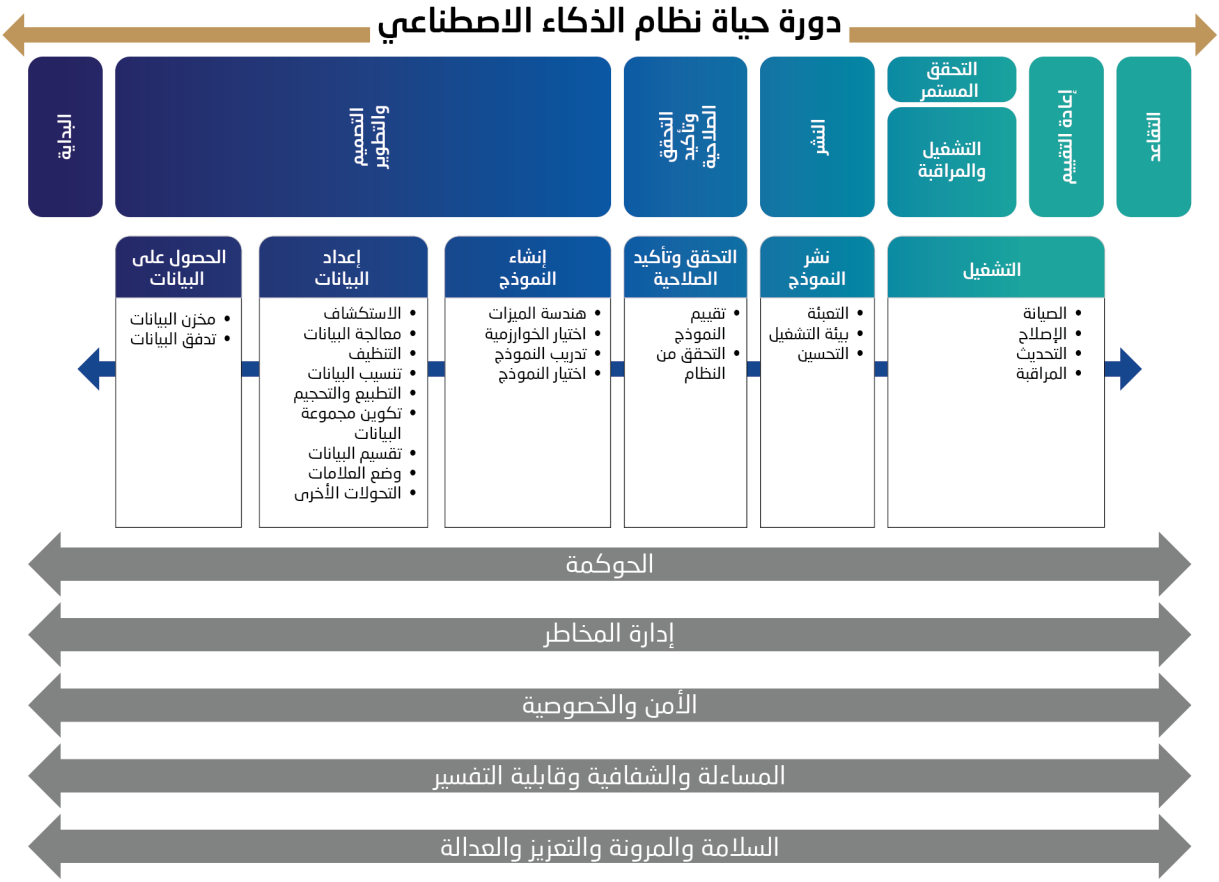
صنفت الأمم المتحدة أنظمة الذكاء الاصطناعي على أنها تكنولوجيا ذات استخدام مزدوج مثل التكنولوجيا النووية. وحتى على نطاق أقل، فإن تكنولوجيا الذكاء الاصطناعي التجارية، بما في ذلك تكنولوجيا الذكاء الاصطناعي الابتكاري، لديها القدرة على إساءة استخدامها لأغراض ضارة. إذ يمكن للجهات الفاعلة الضارة استخدام تقنيات ذكية لتجاوز التدابير الأمنية والضوابط المعمول بها لإنشاء برامج ضارة بمستوى متطور كما لو تم إنشاؤها بواسطة جهة تهديد ترعاها الدولة إذ يمكن أن تعيد تعريف قدرات جهة تهديد تقليدية قليلة أو معدومة الخبرة.

يتم تدريب البرامج الضارة القائمة على الذكاء الاصطناعي من خلال التعلم الآلي لتكون أسرع وأكثر فعالية من البرامج الضارة التقليدية. على عكس البرامج الضارة التي تستهدف العديد من الأشخاص بهدف مهاجمة نسبة صغيرة منهم بنجاح، يتم تدريب البرامج الضارة القائمة على الذكاء الاصطناعي على التفكير بنفسها، وتحديث إجراءاتها بناءً على سيناريو معين، وعلى وجه التحديد تستهدف ضحاياها وأنظمتهم.

5 المبادئ التوجيهية

إن الهدف من هذه المبادئ التوجيهية هو تحفيز المؤسسات في رحلتها لاعتماد واستخدام الذكاء الاصطناعي بطريقة آمنة. لذا، ولوضع ذلك في منظوره الصحيح، يجب أن يكون لدى ناشري الذكاء الاصطناعي متطلبات استراتيجية لاعتماد الذكاء الاصطناعي والتي تزيد من تحفيز صُناع القرار كأسباب حاجتهم إلى الذكاء الاصطناعي، وكيف يتجهون للحصول على حلول الذكاء الاصطناعي المناسبة، وما يحتاجون إليه لتأمين حلول الذكاء الاصطناعي، وما هي القيمة التي يحصلون عليها من حلول الذكاء الاصطناعي، وتأكيد صلاحيته والتحقق من أن حلول الذكاء الاصطناعي آمنة بشكل مقبول.

توصي الأقسام التالية بقائمة من أفضل الممارسات الأمنية لنظام ذكاء اصطناعي آمن وموثوق يمكن للمؤسسات اعتماده وتنفيذه ضمن أنظمة الذكاء الاصطناعي القائمة حاليًا أو أثناء إنشاء أنظمة ذكاء اصطناعي جديدة. تم تصميم المبادئ التوجيهية حول الركائز الثلاثة الراسخة، الأفراد والعمليات والتكنولوجيا، ويتبع دورة حياة منتجات الذكاء الاصطناعي، و يمكن للمؤسسات وفقًا لتقديرها الخاص اختيار الممارسات والضوابط التي تناسب بيئتها.



(شكل 2) مهام دورة حياة الذكاء الاصطناعي من حيث الضوابط[7]

كما ذُكر سابقًا، فإن هذه المبادئ التوجيهية لا تحل محل السياسات والمعايير والارشادات وأفضل الممارسات الموجودة حاليًا و المتعلقة بالأمن السيبراني، ولكنها تكملها، وتركز على المجالات التي يجب فيها تعديل الهياكل الحالية لتتناسب مع التهديدات والمخاطر الأمنية الجديدة التي يجلبها الذكاء الاصطناعي.

يجب على كل مؤسسة تعتزم استخدام الذكاء الاصطناعي استيعاب أن أنظمة المعلومات الأكثر تعقيدًا تتكون في النهاية من عدد كبير من عناصر معالجة أبسط يتم تجميعها معًا من خلال طبقات إضافية من الأجهزة والبرمجيات. بالإضافة إلى ذلك، يمكن استهداف أنظمة الذكاء الاصطناعي من خلال الخدمات السحابية التي قد تكون قيد الاستخدام، أو المكونات المادية مثل وحدات معالجة الرسومات (GPUs) ووحدات معالجة الموترات (TPUs)[8].

إن صحة أمن المعلومات الأساسية من الأمور الجوهرية، إذ يجب أن تمتلك الجهات القائمة على تطبيق الذكاء الاصطناعي جميع المعلومات الأساسية اللازمة وضوابط الأمن السيبراني وهياكل الحوكمة والسياسات المعمول بها.

يجب على المؤسسات في دولة قطر الاستمرار في الالتزام بلوائح الأمن السيبراني التي تنشرها الوكالة الوطنية للأمن السيبراني، مثل السياسة الوطنية لتصنيف البيانات، ومعيار تأمين المعلومات الوطنية، إلى جانب سياسات ومعايير وأُطر وتوجيهات الأمن والخصوصية الأخرى المنشورة من وقت لآخر كما و يمكن للمؤسسات أيضًا اختيار اتباع المعايير والأُطر الدولية مثل مجموعة معايير ISO 27000.

7 مقتبس من معيار دورة حياة نظام الذكاء الاصطناعي (ISO 22898-2023).

8 وحدات معالجة الرسومات (GPUs) ووحدات معالجة الموترات (TPUs) هي معالجات متخصصة مصممة لتسريع أعباء أعمال الذكاء الاصطناعي، ويمكنها تقديم نواقل هجمات جديدة. ويمكن لعيوب التصميم في المعالجات والأجهزة الأخرى أن تؤثر على مجموعة من المنتجات.

1.5 الأفراد

يمكن أن يؤدي استخدام الذكاء الاصطناعي إلى جلب مخاطر كبيرة والتزامات إضافية على المؤسسة. ولضمان الاستخدام الفعال والمقبول لأنظمة الذكاء الاصطناعي في المؤسسة، فإن إنشاء حوكمة الذكاء الاصطناعي من الأمور الضرورية.

يهدف هذا القسم إلى توجيه المؤسسات التي تستخدم الذكاء الاصطناعي في تطوير هياكل الحوكمة الداخلية المناسبة التي تسمح لها بفرض الرقابة على تكنولوجيات الذكاء الاصطناعي وفهم كيفية تأثير الذكاء الاصطناعي على الموارد البشرية.

1.1.5 الاستخدام الأخلاقي للذكاء الاصطناعي

يتم تمكين حوكمة مؤسسة تستخدم الذكاء الاصطناعي من خلال تطبيق مبادئها، و بالتالي، يجب على المؤسسات تحديد مجموعة من مبادئ الاستخدام الأخلاقي والعادل عند تطوير، أو نشر منتجات، أو خدمات، أو أنظمة الذكاء الاصطناعي.

و عند وضع مبادئ أخلاقية للذكاء الاصطناعي، يجب على المؤسسة مراجعة قيمها المؤسسية القائمة في ضوء "مبادئ الذكاء الاصطناعي الأخلاقية والعادلة" المنصوص عليها في الملحق "8-1" من هذه الوثيقة. تتوافق مبادئ الذكاء الاصطناعي الأخلاقية والعادلة المنصوص عليها في هذه المبادئ التوجيهية مع أفضل الممارسات العالمية كما تتوافق مع الاستراتيجية الوطنية للذكاء الاصطناعي بدولة قطر[9].

2.1.5 هياكل الحوكمة الداخلية

تساعد هياكل وتدابير الحوكمة الداخلية على ضمان الرقابة القوية عند استخدام المؤسسة للذكاء الاصطناعي. يمكن تهيئة هياكل الحوكمة الداخلية القائمة بمؤسسة تطبيق الذكاء الاصطناعي لتناسب التحديات الجديدة التي يجلبها الذكاء الاصطناعي. على سبيل المثال، يمكن إدارة المخاطر المرتبطة باستخدام الذكاء الاصطناعي ضمن هيكل إدارة المخاطر في المؤسسة، في حين يمكن تقديم الاعتبارات الأخلاقية كقيم مؤسسية وإدارتها من خلال لجان مراجعة الأخلاقيات أو هياكل مماثلة.

يمكن أن يؤثر نظام الذكاء الاصطناعي على ثقافة المؤسسة من خلال تغيير وإدخال مسؤوليات، وأدوار ومهام جديدة، و يجب تخصيص أدوار المسؤولية والرقابة على المراحل والأنشطة المختلفة المستخدمة في تطبيق الذكاء الاصطناعي للأفراد و/أو الإدارات المناسبة.

يمتلك هؤلاء الأفراد:

- سلطة معالجة مخاطر الذكاء الاصطناعي.
- مسؤولية إنشاء ومراقبة عمليات معالجة مخاطر الذكاء الاصطناعي.
- سلطة اتخاذ قرارات بشأن المستوى المناسب للمشاركة البشرية في صنع القرارات المعززة بالذكاء الاصطناعي.
- مسؤولية صيانة، ومراقبة، وتوثيق ومراجعة نماذج الذكاء الاصطناعي.

3.1.5 بناء الإمكانات والقدرات داخل المؤسسة

يعد الذكاء الاصطناعي مجالًا جديدًا نسبيًا يتطلب مهارات جديدة متعددة يتم تطويرها وتعزيزها داخل المؤسسة، تشتمل على التعلم الآلي، علوم البيانات، الشبكات العصبية، الفهم القانوني، الأخلاقيات، و غيرها من المهارات.

ولذلك، يجب على الإدارة العليا اتخاذ خطوات لتحسين المهارات المتعلقة بالذكاء الاصطناعي بين الموظفين. كما ويجب أن تنظر إلى ما هو أبعد من رفع المهارات الفنية كإعطاء حافز للمهارات المتعلقة بالأخلاق والقانون والخصوصية.

9 في حالة وجود تعارض، فإن المبادئ المنصوص عليها في الاستراتيجية الوطنية للذكاء الاصطناعي المحدثة لدولة قطر أو أي منشور مستقبلي من قبل وزارة الاتصالات وتكنولوجيا المعلومات و/أو الجهة الوطنية المعنية بحوكمة الذكاء الاصطناعي، تكون لها الأولوية على المبادئ المنصوص عليها في هذه المبادئ التوجيهية.

4.1.5 سياسات وقيود المستخدم المقبولة

يجب على المؤسسات القائمة على تطبيق الذكاء الاصطناعي تحديد سياسات المستخدم المقبولة ("AUPs") لضمان فهم المستخدمين لكيفية استخدام نظام الذكاء الاصطناعي بشكل آمن. بشكل عام، تستهدف سياسات المستخدم المقبولة هذه المستخدمين النهائيين لمساعدتهم على فهم ما يجب فعله وما لا يجب فعله.

يمكن إنشاء سياسة الاستخدام المقبول باستخدام:

1. قائمة بحالات الاستخدام المحتملة، بناءً على المخاطر المتوقعة والاحتمالية ودرجة الخطورة.
2. تعريف المخاطر المنخفضة وحالات الاستخدام المقبولة.

قد تجد المؤسسة أنه من الجيد صياغة قائمة "ما يجب فعله وما لا يجب فعله" فيما يتعلق بالاستخدام. على سبيل المثال:

- عدم إدخال أي معلومات تعريفية شخصية.
- عدم إدخال أي معلومات حساسة.
- عدم إدخال أي عنوان بروتوكول إنترنت IP خاص بالشركة دون الرجوع إلى فريق أمن تكنولوجيا المعلومات.
- إيقاف تشغيل تتبع السجلات القديمة.
- مراقبة المخرجات عن كثب، بحثًا عن الأخطاء الواقعية والبيانات المتحيزة أو غير المناسبة.
- عدم إدخال بيانات ضارة تتلاعب بشكل غير مقبول بأداء و/أو نتائج نموذج الحل.

ينبغي تعريف سياسات المستخدم المقبولة بلغة بسيطة لا لبس فيها.

5.1.5 الرقابة البشرية في صنع قرارات الذكاء الاصطناعي

وهي عوامل اتخاذ القرارات الاستراتيجية في البيانات من مصادر مختلفة، على سبيل المثال، الوعي بالموقف لاتخاذ قرار مستنير. إذ أن العديد من الأدوات المتاحة لاتخاذ القرار الآلي (ADM)، وبالتالي فإن الذكاء الاصطناعي ليس حلاً واحدًا يناسب الجميع، إلا أن الاهتمام باتخاذ القرارات المدعومة بالذكاء الاصطناعي يتزايد عالميًا.

الخطوة الأولى في تحديد الرقابة البشرية هي وجود هدف واضح يتمثل في استخدام الذكاء الاصطناعي. تستطيع المؤسسات القائمة على استخدام الذكاء الاصطناعي اتخاذ قرار بشأن أهدافهم التجارية، ويمكن مقارنتها بمخاطر استخدام الذكاء الاصطناعي في عملية صنع القرار. علاوة على ذلك، ينبغي خاصة بالنسبة للمؤسسات العاملة في بلدان متعددة، أن تنظر في كيفية تفاعل أنظمة الذكاء الاصطناعي مع المعايير والقيم والتوقعات المجتمعية الموجودة مسبقًا كما و يجب أن تشكل الرقابة البشرية جزءًا لا يتجزأ من تصميم النظام وأدائه. وقد يشمل ذلك تنفيذ الإجراءات والضوابط الإلزامية وقواعد التصعيد.

علاوة على ذلك، يجب اعتبار أن بعض اللوائح مثل اللائحة الأوروبية العامة لحماية البيانات (GDPR)، وكذلك بعض لوائح الذكاء الاصطناعي الناشئة الأخرى تقيد مدى اتخاذ الذكاء الاصطناعي للقرارات المستقلة في حالات استخدام محددة.

يمكن تحديد ثلاث نماذج واسعة النطاق للدرجات المختلفة للرقابة البشرية في عملية صنع القرار، تشمل ما يلي:

1. **الرقابة البشرية في الحلقة:** في هذا النموذج، تكون الرقابة البشرية نشطة ومشاركة وتحتفظ بالسيطرة الكاملة على نظام الذكاء الاصطناعي. يقدم نظام اتخاذ القرار الآلي توصيات أو مدخلات فقط، ولكن القرار النهائي يتخذه البشر.
2. **الرقابة البشرية خارج الحلقة:** في هذا النموذج لا توجد رقابة بشرية على نظام الذكاء الاصطناعي وتنفيذ القرارات. يكون لنظام اتخاذ القرار الآلي السيطرة الكاملة على عملية اتخاذ القرار، ولا يحق لأي من البشر تجاوزها.

3. **الرقابة البشرية فوق الحلقة / الرقابة البشرية أعلى الحلقة:** في هذا النموذج، تكون الرقابة البشرية بمثابة دور رقابي أو إشراف، مع القدرة على السيطرة عندما يواجه نظام اتخاذ القرار الآلي أحداثًا غير متوقعة أو غير مرغوب فيها.

يتأثر النموذج المحدد الذي يتم اختياره بعوامل مثل اللوائح وإدارة المخاطر المرغوبة.

2.5 العملية

1.2.5 إدارة المخاطر

المبادئ العامة لإدارة المخاطر هي نهج متكامل ومنظم وشامل. يجب أن تأخذ إدارة المخاطر في الاعتبار النظام بأكمله، بكل تقنياته ووظائفه، وتأثيره على البيئة والجهات المعنية.

ومع ذلك، فإن أنظمة الذكاء الاصطناعي هي أنظمة معقدة ويمكن أن تقدم مخاطر جديدة أو ناشئة للمؤسسة، مع عواقب إيجابية أو سلبية على الأهداف، أو تحدث تغييرات في احتمالية المخاطر القائمة حاليًا. وعلى هذا النحو، فإن هناك حاجة إلى وجود إطار عمل قابل للتكيف لإدارة المخاطر في مجال الذكاء الاصطناعي و الذي يجب أن يكون جزءًا متكاملاً من إطار إدارة المخاطر الخاص بالمؤسسة التي ستطبق أنظمة الذكاء الاصطناعي، ويجب أن يتمتع بالصفات التالية:

الشمولية: السعي إلى الحوار مع مجموعات داخلية وخارجية متنوعة، لتوصيل الأضرار والمنافع، دمج الملاحظات والوعي في عملية إدارة المخاطر.

الديناميكية: إن إطار إدارة المخاطر ديناميكي للأسباب التالية:

- إن طبيعة أنظمة الذكاء الاصطناعي في حد ذاتها ديناميكية، وذلك بسبب التعلم المستمر والتحسين والتقييم والتحقق من صحتها. تتمتع بعض أنظمة الذكاء الاصطناعي بالقدرة على التكيف والتحسين، مما يؤدي إلى إحداث تغييرات ديناميكية من تلقاء نفسها.
- توقعات العملاء بشأن أنظمة الذكاء الاصطناعي عالية ومن المتوقع أن تتغير بسرعة كما تتغير الأنظمة.
- تتغير المتطلبات القانونية والتنظيمية المتعلقة بالذكاء الاصطناعي بشكل متكرر.

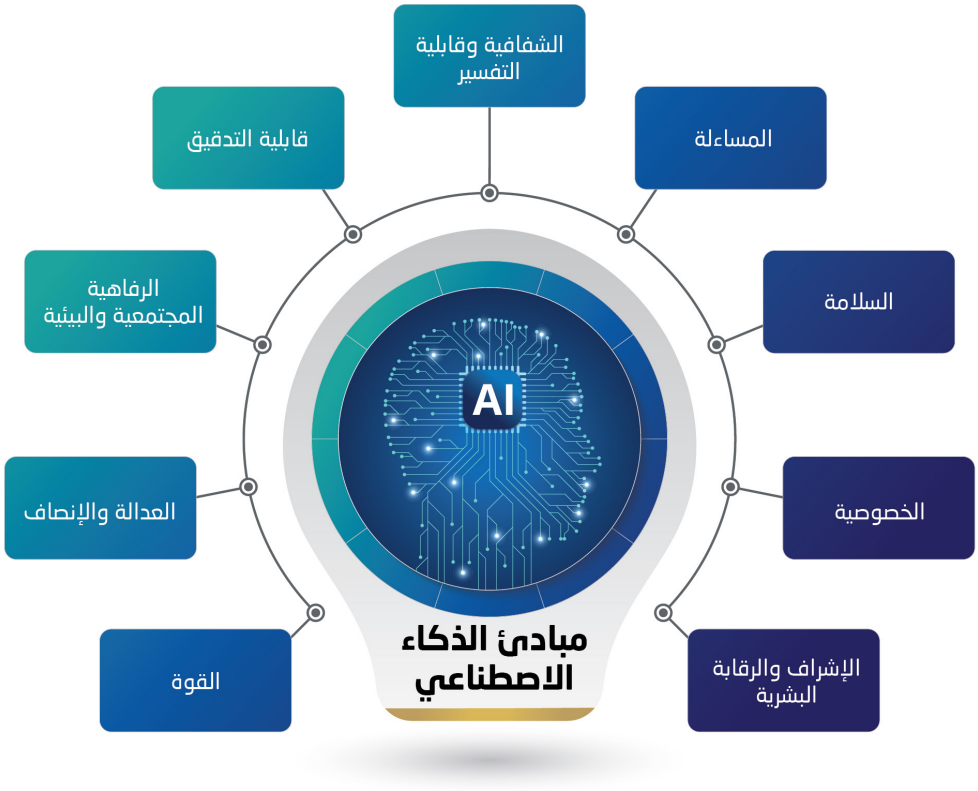
الحساسية للعوامل البشرية والثقافية: مراقبة المشهد، وكيفية تفاعل أنظمة أو مكونات الذكاء الاصطناعي مع الأنماط المجتمعية الموجودة مسبقًا والتي يمكن أن تؤدي إلى تأثيرات على النتائج العادلة، الخصوصية، حرية التعبير، العدالة، السلامة، الأمن، التوظيف، البيئة، وحقوق الإنسان على نطاق واسع.

مساندة التحسين المستمر: يجب أخذ تحديد المخاطر غير المعروفة سابقًا والمتعلقة باستخدام أنظمة الذكاء الاصطناعي بعين الاعتبار في عملية التحسين المستمر. إذ يجب على المؤسسات مراقبة النظام البيئي للذكاء الاصطناعي لمعرفة نجاحات الأداء وأوجه القصور والدروس المستفادة، والحفاظ على الوعي بنتائج وتقنيات أبحاث الذكاء الاصطناعي الجديدة.

الذكاء الاصطناعي هو نموذج جديد تمامًا. ومع قدرته على التأثير و/أو اتخاذ القرارات التي قد تؤثر على حياة البشر، فمن المهم أن تلتزم هذه الأنظمة بمعايير عالية من بالثقة. تحدد وتوضح المبادئ التالية للذكاء الاصطناعي[10] مدى موثوقية نظام الذكاء الاصطناعي. تتوافق المبادئ الأخلاقية والعادلة للذكاء الاصطناعي المنصوص عليها في هذا الدليل الإرشادي مع أفضل الممارسات العالمية والاستراتيجية الوطنية للذكاء الاصطناعي في دولة قطر. تحدد وتوضح المبادئ في (شكل3)[11] مدى موثوقية نظام الذكاء الاصطناعي.

10 في حالة وجود تعارض، فإن المبادئ المنصوص عليها في الاستراتيجية الوطنية للذكاء الاصطناعي المحدثة لدولة قطر أو أي منشور مستقبلي من قبل وزارة الاتصالات وتكنولوجيا المعلومات و/أو الجهة الوطنية المعنية بحوكمة الذكاء الاصطناعي، تكون لها الأولوية على المبادئ المنصوص عليها في هذه المبادئ التوجيهية.

11 شرح تفصيلي في الملحق 8-1 قسم المبادئ الأخلاقية والعادلة للذكاء الاصطناعي من هذه الوثيقة



(شكل 3) المبادئ الأخلاقية والعادلة للذكاء الاصطناعي

يتطلب بناء نظام ذكاء اصطناعي جدير بالثقة تحقيق التوازن بين كل من هذه المبادئ في سياق المؤسسات والجهات المعنية ونظام الذكاء الاصطناعي نفسه. وهذا يستلزم دمج إدارة المخاطر في دورة حياة نظام الذكاء الاصطناعي، بدءًا من مرحلة التخطيط والتصميم وحتى مرحلة التشغيل والمراقبة.

إن الإطار التفصيلي لإدارة المخاطر يقع خارج نطاق هذه الوثيقة. ومع ذلك، فإن الهدف هو توفير الخطوط العريضة لإطار عمل إدارة مخاطر أنظمة الذكاء الاصطناعي.

إن من الحقائق الأساسية التي يجب على المؤسسات وضعها في عين الاعتبار عند إدارة مخاطر الذكاء الاصطناعي هي:

1. أنظمة الذكاء الاصطناعي معقدة البيانات. تؤثر مجموعات البيانات في النهاية على كيفية عمل النموذج و/أو تقديم الاستدلالات، حيث يجب على هذه المؤسسات بذل العناية في تحديد التهديدات ذات الصلة، المخاطر المتعلقة بالبيانات، دورة حياة البيانات، العمليات المتعلقة بالحصول على البيانات، معالجتها، تخزينها، إخراجها، الجهات المعنية بها، وما إلى ذلك. كما ويجب على المؤسسات أيضًا أن تأخذ بعين الاعتبار جميع متطلبات قانون حماية خصوصية البيانات الشخصية وتضمن الامتثال لها.
2. يمكن أن تكون أنظمة الذكاء الاصطناعي معقدة للغاية، وبالتالي تحتاج هذه المؤسسات إلى النظر في المخاطر المتعلقة بقابلية تفسير النظام، والمخاطر التنظيمية خاصة فيما يتعلق بالخصوصية والعدالة، وقابلية التدقيق، والتأثير البشري، والمجتمعي.
3. تحتاج نماذج الذكاء الاصطناعي إلى اختبار دقيق عبر مجموعات كبيرة من البيانات، وإعادة التحقق من صحتها باستمرار للتأكد من استمرار توافقها مع الأهداف والمبادئ الأساسية للمؤسسات التي تنشرها.
4. على الرغم من بذل العناية اللازمة والتكاليف والتعقيدات وإدارة المخاطر، فإن أنظمة الذكاء الاصطناعي تمتلك درجة من عدم القدرة على التنبؤ.
5. بناءً على أهمية النظام والوظيفة التي يخدمها نظام الذكاء الاصطناعي، فمن الضروري تحديد دور البشر وتدخلهم بوضوح خلال مرحلة التخطيط والتصميم نفسها، وتنفيذه ومراقبته بشكل فعال خلال دورة نظام الذكاء الاصطناعي.

6. فيما يلي بعض التوجيهات الواجب اتباعها عند تحديد معايير المخاطر:

أ. يجب على المؤسسات اتخاذ خطوات معقولة لفهم الثغرات في جميع أجزاء نظام الذكاء الاصطناعي، بما في ذلك البيانات المستخدمة والبرمجيات ونماذج الرياضيات والامتداد المادي وجوانب الرقابة البشرية للنظام.

ب. الوعي بأن الذكاء الاصطناعي هو نطاق تكنولوجي سريع الحركة، وينبغي تقييم طرق القياس باستمرار

ت. وضع نهج ثابت وديناميكي لتحديد مستوى المخاطر.

7. التوجيهات الواجب اتباعها عند تحديد الرغبة في المخاطرة:

أ. يجب على الجهة المنظمة تقييم الاستخدام المقصود للذكاء الاصطناعي كجزء من رغبتها في المخاطرة.

ب. يجب أخذ قدرة المؤسسة في مجال الذكاء الاصطناعي ومستوى المعرفة والقدرة على التخفيف من مخاطر الذكاء الاصطناعي المحققة في الاعتبار عند تحديد مدى تقبلها لمخاطر الذكاء الاصطناعي.

8. يجب تحديد مخاطر الذكاء الاصطناعي وقياسها كميًا أو وصفها نوعيًا وترتيبها حسب الأولوية وفقًا لمعايير وأهداف المخاطر ذات الصلة بالمؤسسة ودمجها في دورة الذكاء الاصطناعي.

9. تتطلب أنظمة الذكاء الاصطناعي أن يكون لدى المؤسسة القدرة على الاستجابة للمخاطر والتخفيف منها بطريقة ديناميكية واستباقية، لأن المخاطر في مجال الذكاء الاصطناعي، يمكن أن تتغير بسرعة. إن توفر المشهد العام و النهج الاستباقي للمؤسسة ضروري للاستجابة للمخاطر. ولذلك يجب على المؤسسة أن تثبت استعدادها لتعديل أو إلغاء المشاريع، إذا رأت ذلك ضروريًا.

2.2.5 إدارة عمليات الذكاء الاصطناعي

1.2.2.5 التفاعل والتواصل مع الجهات المعنية

يجب تثقيف الإدارة بشأن المفهوم العام للذكاء الاصطناعي كتكنولوجيا، قدراتها، وتأثيرها المحتمل على المؤسسة والمجتمع، وكذلك تحديات تنفيذها. يعد بناء أنظمة الذكاء الاصطناعي عملية معقدة وضخمة وربما باهظة الثمن. كما ويجب أن تتوافق الإدارة مع دورات التطوير وكذلك مع إدارة فشل أنظمة الذكاء الاصطناعي.

على الإدارة العليا أن توفر معلومات عامة حول ما إذا كان سيتم استخدام الذكاء الاصطناعي وكيفية استخدامه في منتجات و/أو خدمات المؤسسة. يجب أن يكون التواصل مفتوحًا وسهلاً، ويمكن أن يأخذ شكل وصف عام للمنتج أو وضع علامات/علامة المائية على المنتجات الرقمية.

من المستحسن الكشف عن الطريقة التي قد يؤثر بها قرار الذكاء الاصطناعي على الفرد (الموظف أو العميل)، وما إذا كان القرار قابلاً للتراجع. يجب على الإدارة إبلاغ المستخدمين إذا كان لديهم خيار إلغاء الاشتراك في استخدام الذكاء الاصطناعي إذ ينبغي تحديد قنوات واضحة لمراجعة القرار وردود الأفعال.

2.2.2.5 حفظ السجلات (التوثيق)

يعد التوثيق أحد المكونات الرئيسية لضمان قابلية التفسير وقابلية التدقيق في أنظمة الذكاء الاصطناعي، بما في ذلك على سبيل المثال لا الحصر التصميم، البناء، النموذج، مجموعات البيانات، الاختبارات والتحقق من صحتها ويجب على المؤسسات التأكد من عدم التلاعب بهذه السجلات وتأمينها وفقًا لذلك.

3.2.2.5 المراقبة الأمنية (إدارة السجلات)

عندما تقوم المؤسسات بتقييم تكامل نظام ما أو حل للذكاء الاصطناعي في عملياتها، يجب عليها البحث عن أنظمة الذكاء الاصطناعي التي تم تصميمها وتطويرها بقدرات تتيح التسجيل التلقائي للأحداث ("السجلات") أثناء تشغيل المنتج. ويجب أن تتوافق قدرات التسجيل هذه مع المعايير المعترف بها أو المواصفات الشائعة. كما و يجب أن تكون

الأنظمة قادرة على تسجيل السجلات المتعلقة باتخاذ القرار إلى أقصى حد ممكن، إلى جانب سجلات نظام التشغيل والتطبيقات والأمن.

4.2.2.5 ملاحظة البيانات

بالإضافة إلى مراقبة مكونات النظام مثل الأجهزة والأنظمة والتطبيقات، يجب أيضًا مراقبة البيانات. يُعرف علم مراقبة البيانات باسم "ملاحظة البيانات". فهو يوفر للمؤسسات القائمة على تطبيق الذكاء الاصطناعي رؤية واسعة النطاق لمشهد البيانات الخاص بهم وتبعيات البيانات متعددة الطبقات، مثل مسارات البيانات والبنية التحتية للبيانات وتطبيقات البيانات. ومن خلال المراقبة المستمرة والتتبع والتنبيه والتحليل واستكشاف الأخطاء وإصلاحها، تهدف ممارسات ملاحظة البيانات إلى تقليل ومنع أخطاء البيانات أو انقطاعها ضمن اتفاقيات مستوى الخدمة المقبولة.

تشمل ملاحظة البيانات بشكل عام إما ملاحظة أو مراقبة ما يلي:

أ. محتوى البيانات؛

ب. تدفق وسير البيانات؛

ت. البنية التحتية والحصر؛

ث. المستخدم والاستخدام والاستعانة بالبيانات؛

ج. المخصصات المالية.

5.2.2.5 عمليات الكشف

يجب على المؤسسات القائمة على تطبيق الذكاء الاصطناعي اتخاذ خطوات استباقية لإظهار الشفافية والمساءلة فيما يتعلق بتطوير واستخدام الذكاء الاصطناعي.

يجب على الإدارة توصيل السياسات والبيانات المتعلقة باستخدام الذكاء الاصطناعي بشكل فعال بالإضافة إلى أهداف المؤسسة المنشودة، وإدارة المخاطر المرتبطة بها من بين أمور أخرى إلى الجهات المعنية لتعزيز ثقتهم في استخدام الذكاء الاصطناعي.

6.2.2.5 دمج حلول الذكاء الاصطناعي

سيركز الدمج الناجح لنظام الذكاء الاصطناعي على إمكانات المؤسسة وليس على الجدوى التقنية والبراعة. يعتمد المتطلبان الأكثر أهمية للدمج الفعال لنظام الذكاء الاصطناعي على البيانات والاختبار.

ستضمن البيانات عالية الجودة فعالية النموذج الأساسي إذ تحتاج المؤسسات القائمة على تطبيق الذكاء الاصطناعي إلى النظر بجدية في التأكد من أن البيانات التي تم الحصول عليها لبناء النموذج صالحة وكافية من حيث الكمية وتتوافق مع المتطلبات القانونية. ويجب عليهم أيضًا النظر في تأمين البيانات باستخدام ضوابط كافية، بالإضافة إلى التأكد من تأمين البيانات أثناء انتقالها طوال سلسلة توريد البيانات.

يجب على المؤسسات التأكد من وجود ضوابط كافية ضمن النظام، والعمليات للحماية من التحيز إذ يجب إنشاء أدلة التشغيل لاختبار هذه التهديدات بشكل منتظم، مع التركيز بشكل خاص على تقنيات الهجوم المعروفة مثل الهجمات العدائية[12] على سبيل المثال: المراوغة، أو استخراج النماذج، أو تسميم النماذج[13]. تستهدف الهجمات العدائية نماذج أو أنظمة الذكاء الاصطناعي في بيئة إنتاجها، لكن هجمات تسميم النماذج تستهدف نماذج الذكاء الاصطناعي في بيئة التطوير أو الاختبار.

12 تتلاعب الهجمات العدائية بالبيانات المدخلة لإحداث أخطاء أو سوء تصنيف، وتجاوز التدابير الأمنية والتحكم في عملية صنع القرار بشأن أنظمة الذكاء الاصطناعي.

13 في عملية تسميم النماذج، يقوم المهاجمون بإدخال بيانات ضارة في بيانات التدريب للتأثير على المخرجات، مما يؤدي في بعض الأحيان إلى انحراف كبير في السلوك عن نموذج الذكاء الاصطناعي.

يعد اختبار النموذج جزءًا مهمًا من دورة تطوير الذكاء الاصطناعي. إذ تحتاج المؤسسة القائمة على تطبيق الذكاء الاصطناعي إلى إنشاء عمليات واضحة لضمان اختبار النموذج وفقًا لأهداف المؤسسة في كل مرحلة رئيسية في دورة تطوير الذكاء الاصطناعي. وهذا وحده سيضمن أن النموذج يتم تنفيذه كما هو متوقع وأنه يحمي المؤسسة من التحيزات الراسخة والمحتملة التي تتسلل إلى النموذج.

يجب أن تسعى المؤسسات جاهدة لبناء الثقة في أنظمة الذكاء الاصطناعي الخاصة بها من خلال جعلها قابلة للتفسير، قابلة للتكرار، قوية، مناسبة، وآمنة. كما ويجب أن يكون نموذج الذكاء الاصطناعي مناسبًا للغرض ويتناسب مع رغبة المؤسسة في المخاطرة، وأن يكون مضمونًا من خلال التحقق والتأكد الأمني و أن يكون قابلاً للتدقيق عند تصميمه.

7.2.2.5 مراقبة موثوقية البيانات

تتطور موثوقية نظام الذكاء الاصطناعي طوال دورة حياته حيث يجب على المؤسسات القائمة على استخدام الذكاء الاصطناعي تطوير عمليات لضمان مراقبة موثوقية البيانات طوال دورة الذكاء الاصطناعي وذلك يتطلب وضع ضوابط لتحديد الإجراءات الواجب اتخاذها عندما لا يعمل النموذج كما هو متوقع.

يجب على المؤسسات دراسة إمكانية استخدام الأدوات التي تسمح لنموذج الذكاء الاصطناعي بالإبلاغ عن درجة عدم اليقين إلى جانب التنبؤ أو المخرجات. يمكن أن تكون هذه الأفكار ذات قيمة للمشغل البشري وتعزز الثقة في بناء نظام ذكاء اصطناعي قوي.

يجب أن تأخذ المؤسسات القائمة على تطبيق الذكاء الاصطناعي أيضًا في الاعتبار العمليات والأدوات الممكنة لمراقبة وقياس انحرافات النماذج[14].

8.2.2.5 قابلية تدقيق أنظمة الذكاء الاصطناعي

تحتاج المؤسسات إلى التأكد من أن عمليات التدقيق الداخلي تغطي أنظمة الذكاء الاصطناعي. بالمثل، يجب عليها التأكد من أن الآليات التي تسهل إمكانية تدقيق نظام الذكاء الاصطناعي (مثل إمكانية تتبع عملية التطوير، مصادر بيانات التدريب، وتسجيل العمليات، المخرجات والتأثير الإيجابي والسلبي لنظام الذكاء الاصطناعي) مدمجة في نظام الذكاء الاصطناعي أثناء مرحلة التصميم.

ستساهم القدرة على تدقيق أنظمة الذكاء الاصطناعي في وجود ذكاء اصطناعي جدير بالثقة، في المجالات التي تُستخدم فيها أنظمة الذكاء الاصطناعي في الصناعات والعمليات الحيوية نظراً لتأثير ذلك على الحقوق الأساسية للمستخدمين، ينبغي النظر في عمليات التدقيق المستقلة لإثبات شفافية أنظمة الذكاء الاصطناعي. كما ويجب على قسم التدقيق الداخلي بناء القدرات اللازمة وشراء الأدوات التي ستساعدهم على تسهيل التدقيق الناجح لأنظمة الذكاء الاصطناعي.

يمكن استخدام تقنيات مختلفة مثل مراجعة المستندات، أو الإرشادات التفصيلية للعملية، أو تحليل البيانات، أو التقييمات الأمنية لإجراء التدقيق. يجب أن تكون فرق التدقيق على دراية بالتحديات مثل تعقيد أنظمة الذكاء الاصطناعي والنماذج الأساسية والمشهد التنظيمي الديناميكي ونقص القدرات الفنية ذات الصلة من بين أمور أخرى. كما و يجب أن تكون عملية تدقيق البيانات قادرة على تحديد المشكلات المحتملة في النظام إن وجدت، مثل تحيز البيانات والأخلاقيات والامتثال التنظيمي وما إلى ذلك.

تعد تقييمات الامتثال والشهادات وفقًا للمعايير المعمول بها وذات الصلة طريقة جيدة لإثبات قابلية التدقيق وبناء الثقة في أنظمة الذكاء الاصطناعي.

9.2.2.5 الإبلاغ عن الحوادث

لزيادة الثقة في الذكاء الاصطناعي، يجب على المؤسسات إنشاء قنوات إبلاغ فعالة للجهات المعنية للإبلاغ عن الأداء

14 يشير انحراف النموذج إلى مشكلة اضمحلال قدرة النموذج على التنبؤ بناءً على التغيرات التي تحدث في البيئة.

غير المناسب، أو عند اكتشاف خرق للبيانات[15] أو جوانب قلق أخرى بشأن سلوك نظام الذكاء الاصطناعي.

يجب أن تكون القناة سهلة الاستخدام ومتاحة للمستخدمين الداخليين والخارجيين. كما و يجب أن يتم عمل مراقبة بشرية لتقارير المستخدمين والرد عليها بالشكل المناسب.

3.5 التكنولوجيا

يمكن أن تكون أنظمة الذكاء الاصطناعي، حالها حال أية أنظمة معلومات أُخرى تقوم بمعالجة البيانات وتحليلها عرضة للخطر بسبب الأجهزة الأساسية والبرمجيات (البرمجيات الثابتة ونظام التشغيل والبرمجيات الافتراضية والتطبيقات وما إلى ذلك)، وبروتوكولات الشبكة والأمن، والتشفير غير الكافي، وعدم وجود مراقبة كافية وضعف ضوابط الوصول والتهديدات الداخلية.

و كما تم ذكره سابقًا، لا يمكن التغاضي عن الصحة الأساسية للأمن السيبراني إذ يجب على المؤسسات القائمة على تطبيق الذكاء الاصطناعي التأكد من امتثال أية أنظمة معلومات (بما في ذلك أنظمة الذكاء الاصطناعي) لمعايير أمن المعلومات. تتضمن الفقرات الفرعية التالية الضوابط الفنية و فوائدها من منظور نظام الذكاء الاصطناعي.

1.3.5 تصميم وبنية النظام

إن بناء أنظمة قوية وآمنة ليس بمحض الصدفة، ولكن يتطلب ذلك فهمًا ذكيًا للمتطلبات (بما في ذلك المتطلبات التجارية والوظيفية والفنية والتنظيمية والمالية وما إلى ذلك)، ونمذجة التهديدات لفهم المخاطر، ونمذجة النظام من بين أمور أخرى إذ يتطلب بناء نظام ذكاء اصطناعي جدير بالثقة دمج مبادئ الأخلاق والخصوصية والأمان في النظام عند تصميمه.

يمكن لنظام الذكاء الاصطناعي المصمم بشكل آمن أن يحمي المؤسسات من التهديدات مثل التطبيق غير الآمن على الجمهور[16].

2.3.5 التحكم في الوصول

يعد التحكم في الوصول أمرًا بالغ الأهمية في نظام الذكاء الاصطناعي ويرجع ذلك بشكل أساسي إلى تعقيده وضخامته. يمتلك نظام الذكاء الاصطناعي بطبيعته إمكانية الوصول إلى مجموعات كبيرة من البيانات، وقد يكون لديه عدة أطراف ثالثة كجزء من سلسلة التوريد الخاصة به والتي قد تساهم في عمليات مختلفة داخل النظام، وبالتالي، فإنه من المهم للغاية تنظيم الوصول داخل نظام الذكاء الاصطناعي. يوصى بشدة أن تستخدم المؤسسات القائمة على تطبيق الذكاء الاصطناعي مفهوم "الثقة المعدومة" أثناء تصميم وبناء نظام الذكاء الاصطناعي.

إن النظام القوي للتحكم في الوصول الذي يعتمد على مفهوم "الثقة المعدومة" يمكن أن يساعد المؤسسات على الحماية من التهديدات كضعف التحكم في الوصول[17] والوصول المفرط للوكلاء[18].

3.3.5 أمن الشبكات

يتطلب نظام الذكاء الاصطناعي ويتكون من وحدات / أنظمة متعددة متصلة ببعضها البعض، قد تكون عبر الحدود المادية (والتنظيمية). ويتم توفير هذا الاتصال الأساسي عن طريق "الشبكات". وبالتالي، فإنه من الضروري التأكد من أن الشبكة آمنة بشكل افتراضي وتصميمي. يمكن أن تساعد الشبكة الآمنة المؤسسات القائمة على تطبيق الذكاء

15 يجب أن تتماشى إدارة حوادث الذكاء الاصطناعي مع إدارة خرق البيانات لضمان اكتشاف جميع الحوادث والتعامل معها بشكل صحيح.

16 يمكن أن يكون هذا على سبيل المثال نموذجًا تم تطبيقه مباشرة على خادم استدلال غير آمن أو يمكن تنزيله مباشرة. بالإضافة إلى ذلك، تكون واجهة برمجة تطبيقات الاستدلال أو خدمة الويب ضعيفة وغير مصححة وغير محدثة، ولديها أذونات زائدة لحسابات الخدمة على خوادم الاستدلال.

17 يحدث هذا عندما لا تتمتع مجموعة التكنولوجيا الأساسية بالتحكم الكافي في الوصول ويكون المهاجم قادرًا على تنزيل النموذج أو يتم تصميم واجهات برمجة التطبيقات مع عدم وضع التحكم في الوصول في الاعتبار.

18 يحدث هذا عندما يتمتع وكيل من الجمهور بإمكانية الوصول إلى واجهات برمجة التطبيقات الداخلية الخاصة/المقيدة أو يكون لدى وكيل من الجمهور حق الوصول إلى النماذج الخاصة/المقيدة أو يكون لدى الوكيل حق الوصول إلى الأنظمة المالية.

الاصطناعي على حماية نفسها من التهديدات مثل هجمات حجب الخدمة[19].

4.3.5 أمن سلسلة التوريد

في سياق أنظمة الذكاء الاصطناعي، نحتاج إلى النظر في الجوانب التالية لسلسلة التوريد:

1. الموردون الذين يقدمون خدمات مباشرة كجزء من النظام البيئي: هؤلاء هم الموردون الذين يمكن أن يقدموا خدمات تتعلق بإدارة البيانات وإدارة الشبكات وغيرها.
2. الموردون الذين يقدمون خدمات غير مباشرة مثل الموردين الذين يقدمون الأجهزة/البرمجيات الأساسية وغيرها.

وبغض النظر عن ذلك، يجب على المؤسسات أن تكفل التزام الموردين بالأمن والتقيد بأفضل الممارسات والامتثال للمتطلبات التنظيمية، ويستلزم ذلك وجود عمليات شراء تدعم الأمن من خلال التأكد من أن النضج الأمني للمورد والتزامه بالأمن والمرونة هو أحد المعايير المستخدمة لتأهيل الموردين المحتملين.

سيساعد الموردون الملتزمون بالأمن المؤسسات القائمة على تطبيق الذكاء الاصطناعي على حماية الجهات من التهديدات المتمثلة بعدم الامتثال التنظيمي، والتقليل من تواتر الهجمات، من بين أمور أخرى.

في حال أن المؤسسة تعمل مع مقدمي حلول الذكاء الاصطناعي من جهات خارجية، فسيتم إجراء تقييم شامل لممارساتهم الأمنية وإجراءات التعامل مع البيانات، والتأكد من أنها تلبي معايير الأمن الخاصة بالمؤسسة، وتحديدًا، يجب على المؤسسات التأكد من أن البيانات المتبادلة مع جهات خارجية، بما في ذلك الموردين والشركاء، تخضع لسياسة تبادل البيانات، ويجب على الموردين / الشركاء التأكد من معالجة البيانات وفقًا للتصنيف الأمني المحدد للبيانات بما يتماشى مع سياسة تصنيف البيانات الخاصة بالمؤسسة.

كما ويجب على المؤسسات التأكد من أن الاتفاقيات القانونية مع الموردين والمتعاقدين تشير بوضوح إلى متطلبات الأمن والامتثال.

5.3.5 الوعي بالموقف

في النظام المعقد، فإنه من الضروري وجود فهم واضح للنظام، ترابطاته، سلوكه، وتأثيره في حالة تعطله. تحتاج المؤسسات القائمة على تطبيق الذكاء الاصطناعي إلى التأكد من إنشاء العمليات والأنظمة اللازمة وتحديد الأدوات التي من شأنها أن تزودهم بالوعي بالموقف اللازم لإدارة ومراقبة نظام الذكاء الاصطناعي. فيما يلي بعض الطرق الخاصة حول كيفية تحسين الوعي بالموقف لأنظمة الذكاء الاصطناعي:

1.5.3.5 إدارة الثغرات

نظرًا لأن أنظمة الذكاء الاصطناعي تتكون من وحدات مترابطة، فقد يكون للثغرات في النظام "تأثير الدومينو" على الأنظمة المترابطة الأخرى. وبالتالي، يجب على المؤسسات إنشاء خرائط توضح كيف سيؤثر اختراق أحد الأصول أو النظم على المكونات الأخرى لأنظمة الذكاء الاصطناعي.

2.5.3.5 برنامج مكافأة الأخطاء

توضح برامج مكافأة الأخطاء مدى استباقية المؤسسة في تحديد الأخطاء ونقاط الضعف المحتملة، وإذا تم تنفيذه بشكل صحيح، فإن ذلك يضمن تمكين المؤسسات من تحديد المشكلات المحتملة في أنظمتها التي تواجه الجمهور قبل أن تقوم بها أي جهة ضارة.

19 أصدرت الوكالة الوطنية للأمن السيبراني توجيهات بشأن الحماية من هجمات حجب الخدمة الموزعة.

3.5.3.5 استخبارات التهديدات

بالنظر إلى مدى ضخامة وديناميكية الهجمات، والتي تتضمن العديد من عوامل التهديد الضارة، سيكون من المستحيل للبشر تتبع التهديدات المحتملة. وبالتالي، يجب على المؤسسات الاستثمار في شراء أو بناء خدمات استخبارات التهديدات إذ أن وجود نظام جيد لاستخبارات التهديدات من شأنه أن يساعد المؤسسات على اكتساب رؤى ثاقبة عن التهديدات المحتملة في الفضاء السيبراني، مما يمنحها ميزة الاستباقية في التخفيف من هذه التهديدات.

كما و يجب على المؤسسات القائمة على تطبيق الذكاء الاصطناعي أن تسعى بشكل فعال إلى التعاون مع الوكالة الوطنية للأمن السيبراني، وزارة الداخلية، الهيئات التنظيمية القطاعية، وجهات إنفاذ القانون، حيث تتمتع هذه الجهات برؤية واضحة حول مشهد التهديدات الإقليمية والوطنية وتحصل على معلومات استخباراتية منظمة عن التهديدات.

6.3.5 نمذجة التهديدات لتحديدها

يجب أن تتبع المؤسسات القائمة على تطبيق الذكاء الاصطناعي نمذجة دقيقة للتهديدات لتحديد التهديدات المحتملة للأنظمة. إن استخدام نهج علمي سيضمن قيام المؤسسات بتحليل النظام بشكل شامل، وتحديد جميع التهديدات المحتملة، وتقييم الحلول المحتملة للتخفيف من التهديدات. كما ويمكن للمؤسسات تقييم نموذج التهديد المناسب لاتباعه. تتضمن بعض نماذج التهديد المعمول بها ما يلي:

- نموذج MITRE ATLAS الذي طورته شركة مايكروسوفت بالتعاون مع شركة ميتري
- نموذج تهديدات الذكاء الصناعي الذي طورته شركة ETSI GR SAI 001
- أفضل 10 تقييمات جيدة للذكاء الاصطناعي (GAIA) الذي طورته شركة جوجل
- دليل أمن وخصوصية الذكاء الاصطناعي الذي طورته مؤسسة OWASP

7.3.5 الذكاء الاصطناعي المرن

يعد النظام القوي إحدى سمات نظام الذكاء الاصطناعي، إذ أن بناء المرونة هو الأساس في بناء نظام قوي للذكاء الاصطناعي. كما و يمكن استخدام الضوابط التالية لبناء وتقييم مرونة نظام الذكاء الاصطناعي:

1.7.3.5 النسخ الاحتياطي

إن التحدي الرئيسي في تحديد استراتيجية النسخ الاحتياطي لنظام الذكاء الاصطناعي هو البيانات الهائلة التي يتم التعامل معها، بالإضافة إلى القيود التنظيمية المفروضة، ونظرًا لأن أنظمة الذكاء الاصطناعي تستخدم مجموعات ضخمة من البيانات، والتي قد تشمل بيانات شخصية تخضع لقانون حماية خصوصية البيانات الشخصية، فعلى المؤسسات النظر في اللوائح وتأثيرها على استراتيجية النسخ الاحتياطي للبيانات.

2.7.3.5 الاختبار

إن اختبار النظام وتأكيد صلاحيته والتحقق منه، مع الأخذ في الاعتبار مدى تعقيده وأهميته، من الأمور بالغة الأهمية، ولا يجب أن يشمل ذلك المكونات الأساسية مثل نموذج الذكاء الاصطناعي ومجموعات البيانات ومنطق المؤسسة فحسب، بل يجب أن يشمل أيضًا الأنظمة الأساسية والبنية التحتية، والأهم من ذلك هو أيضًا الاختبار الشامل للنظام بأكمله.

يجب على المؤسسات القائمة على تطبيق الذكاء الاصطناعي استخدام تقنيات مثل صناديق الاختبارات لاختبار التطبيقات، وإنشاء بيئة اختبار مخصصة، كما قد يكون إنشاء توأم رقمي فكرة جيدة لأنظمة الذكاء الاصطناعي المستخدمة في الوظائف الهامة.

3.7.3.5 المحاكاة

إلى جانب الاختبارات المنتظمة، فإن تنفيذ تمارين المحاكاة باستخدام النطاقات السيبرانية ومنصات المحاكاة، والتدريبات والتمارين هي أساليب جيدة لاختبار وتقييم وتأمين أنظمة الذكاء الاصطناعي. كما ويمكن أيضًا استخدام هذه الأنظمة والتقنيات لتقييم نظام الذكاء الاصطناعي في سياق المؤسسة، وفقًا للعمليات القائمة ومهارات الأفراد.

8.3.5 تأمين البيانات

كما تم مناقشته سابقًا، فإن البيانات هي شريان حياة نظام الذكاء الاصطناعي، إذ أن نظام الذكاء الاصطناعي النموذجي عبارة عن تعاون بين العديد من الجهات المعنية ويتضمن مشاركة البيانات والوصول إليها من جهات خارجية، وبالتالي يزيد سطح الهجوم وخطر الوصول غير المصرح به و إساءة استخدام البيانات الشخصية.

إن طبيعة وسياق وتعقيد أنظمة الذكاء الاصطناعي قد يجعل من ضمان أمن البيانات أمراً صعباً. يوضح القسم 4 من هذه المبادئ التوجيهية العديد من المخاطر والتحديات المتعلقة بتجميع البيانات والاحتفاظ بها وخصوصيتها.

من بين الأُمور الأُخرى، يجب على المؤسسات مراعاة ما يلي لضمان أمن البيانات:

1. التأكد من أن الأنظمة الأساسية والشبكات والبنية التحتية التي تخزن البيانات مؤمنة باستخدام أفضل الممارسات.
2. تسجيل ومراجعة حقوق الوصول والتكوينات وسجلات قاعدة البيانات باستمرار.
3. إعادة تقييم سياسات البيانات المفتوحة في مجال الذكاء الاصطناعي، لضمان تخفيف مخاطر فقدان السرية بسبب تجميع البيانات.
4. تقييم وتنفيذ نموذج التعلم الموحد[20] لتصميم قاعدة البيانات، حيث يتيح التعلم الموحد تدريب النماذج على كميات كبيرة من البيانات مع الحد من كشف أو حركة البيانات الأولية، وبالتالي يمكن اعتباره وسيلة خاصة لتبادل البيانات.
5. استخدام التكنولوجيا المتاحة لحماية وتأمين البيانات، (على سبيل المثال: أنظمة حماية نقطة النهاية، أنظمة منع التطفل، حلول منع تسرب البيانات وإدارة حقوق البيانات).
6. التأكد من النزاهة من خلال تطبيق وظائف التجزئة المشفرة على البيانات وتخزين قيم التجزئة الناتجة، ثم التوقيع على قيم التجزئة باستخدام خوارزمية التوقيع الرقمي إذ تتيح هذه الحماية إمكانية إثبات والتحقق من نزاهة وسلامة البيانات.

6 توصيات خاصة بشأن الذكاء الاصطناعي التوليدي

كما هو الحال مع أي تكنولوجيا جديدة واعدة، تحرص المؤسسات على تسخير واستخدام براعة الذكاء الاصطناعي التوليدي لزيادة الإنتاجية وتعزيز عائدات الاستثمار في المؤسسة. وكما هو الحال مع أي تكنولوجيا أخرى، يأتي الذكاء الاصطناعي التوليدي مع مجموعة من المخاطر والتحديات خاصة به و التي يجب على المؤسسات مراعاتها.

كحد أدنى، يجب على المؤسسات القائمة على تطبيق الذكاء الاصطناعي مراعاة ما يلي:

خصوصية وأمن البيانات: تتضمن بيئة المؤسسات عادةً معالجة بيانات حساسة، مثل المعلومات المالية أو بيانات العملاء أو الأسرار التجارية. ومن المهم التأكد من أن أدوات الذكاء الاصطناعي التوليدي مصممة لتلبية متطلبات الأمن والخصوصية للمؤسسة والدولة، وأن التدابير المناسبة مطبقة لحماية البيانات (مثل الامتثال لقانون حماية خصوصية البيانات الشخصية).

20 . المرجع: ETSI GR SAI 002 V1. 1. 1 (2021-08) تأمين الذكاء الاصطناعي (SAI)؛ أمن سلسلة توريد البيانات على الرغم من عدم خلوه من التهديدات الأمنية، فقد ثبت أن هذا النهج يقلل من فعالية هجمات تسميم البيانات في بعض الحالات حيث يسمح بإدخال المزيد والمزيد من بيانات التدريب المتنوعة، التي تساعد على زيادة قوة النموذج، وتقلل من سيطرة المهاجم على مجموعة البيانات التي يرغب في تسميمها.

التكامل مع الأنظمة الحالية: قد تحتاج أدوات الذكاء الاصطناعي التوليدي إلى التكامل مع الأنظمة الحالية، مثل برنامج إدارة علاقات العملاء (CRM) لتوفير تجربة سلسة للمستخدمين، ويتطلب هذا تخطيطاً وتنسيقاً دقيقاً لضمان التكامل الآمن والموثوق.

التدريب والمساندة: تتطلب أدوات الذكاء الاصطناعي الابتكاري بيانات التدريب والمراقبة والمساندة المستمرة للتأكد من أنها تقدم استجابات دقيقة ومفيدة. إذ يجب على المؤسسات القائمة على تطبيق الذكاء الاصطناعي الاستثمار في الموارد اللازمة لضمان تدريب وصيانة أدوات الذكاء الاصطناعي التوليدي بشكل صحيح. علاوة على ذلك، يجب على المؤسسات أيضًا الاستثمار في تطوير مهارات محددة (مثل هندسة التنبيه الفوري) داخل المؤسسة لإدارة هذه الأدوات.

الوعي الأمني: يجب تزويد عامة المستخدمين بالوعي الأمني بشأن كيفية استخدام هذه الأنظمة، كما و يجب تدريب المستخدمين على عدم الاعتماد فقط على نصائح أو معلومات أدوات الذكاء الاصطناعي التوليدي و التحقق دائماً من المعلومات مع مصادر أخرى موثوقة.

تجربة المستخدم: قد يتردد بعض المستخدمين في التفاعل مع برنامج دردشة آلي وربما يفضلون التحدث مع ممثل بشري. لذا فإنه من المهم التعريف بفوائد الذكاء الاصطناعي التوليدي وتوفير التدريب والمساندة المناسبة لضمان قبول المستخدم. علاوة على ذلك، و لتعزيز الشفافية، فإنه من المستحسن إبلاغ المستخدمين عند تفاعلهم مع الوسائل الآلية.

الصحة الأمنية: لحماية بيانات المستخدمين، يجب عليهم اتباع أفضل ممارسات الأمن عبر الإنترنت، كاستخدام كلمات مرور قوية وفريدة من نوعها، وتجنب النقر على الروابط المشبوهة أو تنزيل مرفقات غير معروفة، والحفاظ على تحديث أجهزتهم وبرامجهم بأحدث التصحيحات والتحديثات الأمنية. كما و ينبغي توعية المستخدمين بعدم مشاركة المعلومات الحساسة أو الشخصية مثل كلمات المرور أو تفاصيل بطاقة الائتمان أو أرقام الضمان الاجتماعي مع أدوات الذكاء الاصطناعي التوليدي.

مراقبة الأداء والاستخدام: يجب على المؤسسات القائمة على الذكاء الاصطناعي مراقبة أداء واستخدام أدوات الذكاء الاصطناعي التوليدي للتأكد من أنها تقدم قيمة للمستخدمين وتلبي أهداف المؤسسة مما يساعد في تحديد مجالات تحسين الأداء.

1.6 تسريب البيانات الحساسة

يمكن للموظفين و بسهولة الكشف عن بيانات المؤسسة الحساسة والخاصة في الأسئلة والاستعلامات أثناء استخدام أدوات الذكاء الاصطناعي التوليدي مثل: ChatGPT, Bard, Bing AI, Dall-E, WordTune Read, Whisper, Eleven Labs. في حالة نظام " ChatGPT (المحول الابتكاري المدرب مسبقًا للدردشة)" في الوقت الحالي، يتم تخزين هذه الأسئلة إلى أجل غير مسمى في البنية التحتية لمنظمة " أوبن أيه آي OpenAI" ويمكن تخزينها بالمثل في إصدارات ChatGPT الأخرى التي يوفرها البائع. بالإضافة إلى ذلك، يمكن استخدام هذه الأسئلة لتدريب نماذج GPT التابعة لجهات خارجية (من خلال مقدمي حلول الذكاء الاصطناعي) في المستقبل مما يزيد من كشف سرية معلومات المؤسسة.

لمنع كشف المعلومات الحساسة، يمكن للمؤسسات القائمة على تطبيق الذكاء الاصطناعي القيام بما يلي:

- عدم السماح بأي قص ولصق لمحتوى خاص بأعمال المؤسسة (مثل: رسائل البريد الإلكتروني والتقارير وسجلات الدردشة)
- إيقاف تشغيل سجل الدردشة والتدريب على البيانات؛
- عدم السماح بأي مدخلات في منتج الذكاء الاصطناعي التوليدي تتضمن بيانات العميل أو أي بيانات شخصية؛
- إجراء مراجعة بشرية لجميع المخرجات المتولدة من الذكاء الاصطناعي المستخدمة في التفاعلات مع العملاء؛
- استخدام واجهات برمجة تطبيقات آمنة لدمج النماذج اللغوية الكبيرة (LLMs) أو نماذج الأساس متعددة الوسائط (MfMs) مع أنظمتها وتطبيقاتها. يجب تأمين واجهات برمجة التطبيقات بآليات تفويض قوية، مثل التفويض المفتوح (OAuth) أو مفاتيح واجهات برمجة التطبيقات، لمنع الوصول غير المصرح به؛

- تقوية الإعدادات وتعطيل خيارات استخدام بيانات المؤسسة للتدريب والتحسين إذا أمكن ذلك؛
- استخدام صناديق الاختبارات أو بوابات إدارة المحتوى لتصفية البيانات المرسلة إلى أدوات الذكاء الاصطناعي التوليدي.

2.6 فهم قيود الذكاء الاصطناعي التوليدي

نظرًا لكون أدوات الذكاء الاصطناعي التوليدي غير قابلة للتفسير ولا يمكن التنبؤ بها إلى حد كبير، فإنها يمكن أن تُقدم باستمرار معلومات غير دقيقة ومُلفقة، وتنتج مخرجات مسيئة، وتكون عرضة للتحيز. و بالتالي قد لا تكون مناسبة لكل حالة استخدام بالمؤسسة. علاوة على ذلك، ينبغي تقييم مخرجات الذكاء الاصطناعي الابتكاري للتأكد من دقتها وملاءمتها وفائدتها قبل قبولها.

قبل نشر النماذج اللغوية الكبيرة أو نماذج الأساس متعددة الوسائط، يجب على الإدارة النظر فيما إذا كان نظام/منتج الذكاء الاصطناعي مناسبًا لغرض المؤسسة، وتقييم الأداء على نطاق واسع من المدخلات المحتملة، لتحديد الحالات التي قد ينخفض فيها الأداء.

يتسم البشر بالتميز والتفرد. بينما نماذج الذكاء الاصطناعي، بغض النظر عن مدى ذكاء الأنظمة الحالية، لا يمكنها فهم الغموض والتناقضات، وتفتقر إلى السياق والنطاق والوعي بالموقف والثقافة التنظيمية لمشكلة معينة ولا تتمتع بالوعي. وأخيرًا، فإن جميع نماذج وأنظمة الذكاء الاصطناعي محدودة في معرفتها. لذلك، يجب على إدارة المؤسسة القائمة على تطبيق الذكاء الاصطناعي أن تأخذ في الاعتبار قاعدة العملاء/الوكلاء ونطاق المدخلات التي سيستخدمونها، لضمان معايرة توقعاتهم بشكل مناسب.

3.6 تطوير الذكاء الاصطناعي التوليدي

يسمح الذكاء الاصطناعي التوليدي بدورات تطوير أسرع من مشاريع الذكاء الاصطناعي التقليدية. ولهذا السبب، يتطلب التجريب دورة بسيطة من الابتكار - تجارب قصيرة لاختبار كيف يمكن للتكنولوجيا أن تضيف قيمة استراتيجية مع تخفيف المخاطر المحتملة التي تأتي معها.

يتطلب النجاح في تجارب الذكاء الاصطناعي التوليدي الاختبار السريع والتحسين، وفي كثير من الأحيان، استبعاد حالات الاستخدام التي ليس لها التأثير المتوقع على قيمة المؤسسة. يجب على المؤسسات القائمة على تطبيق الذكاء الاصطناعي تحديد حالات الاستخدام المحددة لتطبيق الذكاء الاصطناعي الابتكاري ومصادر البيانات التي سيستخدمها. وسيضمن ذلك أن الأدوات مصممة لتلبية الاحتياجات المحددة للمؤسسة ومتكاملة مع مصادر البيانات المناسبة.

يجب على المؤسسات توفير بيانات تدريب كافية لأدوات الذكاء الاصطناعي الابتكاري مثل النماذج اللغوية الكبيرة (LLMs) ونماذج الأساس متعددة الوسائط (MfMs) لضمان تدريبها بشكل صحيح وتمتلك قدرة على تقديم استجابات دقيقة وذات صلة. كما يجب على المؤسسات مراقبة أداء الأدوات وتعديلها حسب الضرورة لتحسين دقتها.

كما وينبغي على المؤسسات القائمة على تطبيق الذكاء الاصطناعي التأكد من حماية بيانات المستخدم بشكل صحيح وأن الوصول إلى البيانات يقتصر على الموظفين المصرح لهم فقط. كما وينبغي تشفير البيانات أثناء النقل وأثناء السكون ووضع ضوابط الوصول وآليات التدقيق المناسبة لضمان خصوصية البيانات وأمنها.

1.3.6 اختبار الخصوم (التحقق من وجود خصم)

يتم فيه اختبار منتج الذكاء الاصطناعي على نطاق واسع من المدخلات وسلوكيات المستخدم، سواء كانت مجموعة تمثيلية أو تلك التي تعكس شخصًا لديه نوايا خبيثة يحاول "اختراق" التطبيق.

تعتمد الاستجابة المقدمة من النماذج اللغوية الكبيرة أو نماذج الأساس متعددة الوسائط على البيانات التي تم تدريبها عليها، إذ أنه من الضروري أن يتحقق المستخدمون من صحة الاستجابات التي يتلقونها من أدوات الذكاء الاصطناعي التوليدي ولا يثقون بشكل أعمى في النظام.

2.3.6 إدارة المحتوى

تصفية المحتوى: لتجنب المحتوى غير المرغوب فيه، فإنه من الجيد دمج نظام تصفية للمحتوى مصمم خصيصًا لنظام الذكاء الاصطناعي الذي تم تطبيقه وحاجة المؤسسة له.

هندسة التنبيه الفوري: يمكن أن تساعد في تقييد الموضوع وأسلوب النص الناتج، الأمر الذي من شأنه أن يقلل من فرصة إنتاج محتوى غير مرغوب فيه، حتى لو حاول المستخدم إنتاجه. إن توفير سياق لنموذج الذكاء الاصطناعي من خلال تقديم بعض الأمثلة عالية الجودة للسلوك المرغوب قبل الإدخال، يمكن أن يسهل توجيه مخرجات النموذج في الاتجاهات المطلوبة.

تسمح هندسة التنبيه الفوري للمؤسسات القائمة على تطبيق الذكاء الاصطناعي باستخدام خدمات الذكاء الاصطناعي الابتكاري العامة، مع حماية الملكية الفكرية للمؤسسة والاستفادة من البيانات الخاصة لإنشاء استجابات دقيقة ومفيدة ومحددة.

قيود التنبيه الفوري: لتقليل احتماليات سوء الاستخدام، يمكن للمؤسسات القائمة على تطبيق الذكاء الاصطناعي إدخال قيود التنبيه الفوري كما يلي:

- الحد من عدد رموز المخرجات.
- تحديد كمية النص وحجم الملف/الصورة/الفيديو/التسجيل الذي يمكن للمستخدم إدخالها في التنبيه الفوري.
- تضييق نطاقات المدخلات أو المخرجات.
- التنبيه بالمدخلات من خلال قائمة الخيارات التي تم التحقق منها، بدلاً من مدخلات النص/الملف/الصورة/الفيديو/التسجيل المفتوحة.

7 الامتثال والتنفيذ

تم إصدار هذه الوثيقة كتوجيهات لمساعدة المؤسسات القائمة على تطبيق الذكاء الاصطناعي على فهم المخاطر المرتبطة بأنظمة الذكاء الاصطناعي وكيفية التخفيف منها من أجل استخدام فعال وآمن.

تكمل هذه المبادئ التوجيهية اللوائح والسياسات والمعايير الوطنية الحالية. كما يجب قراءة الوثيقة في سياق سياسة تصنيف البيانات الوطنية ومعيار تأمين المعلومات الوطنية.

8 الملحق

1.8 مبادئ الذكاء الاصطناعي الأخلاقية والعادلة

1.1.8 الشفافية وقابلية التفسير

في سياق الذكاء الاصطناعي، تشير الشفافية إلى القدرة على فهم وتفسير كيفية وصول نظام الذكاء الاصطناعي إلى قراراته أو تنبؤاته. علاوة على ذلك، تعد الشفافية متطلبًا في العديد من لوائح الخصوصية - كما هو الحال في قانون حماية خصوصية البيانات الشخصية. غالبًا ما تكون العديد من نماذج الذكاء الاصطناعي، وخاصة نماذج التعلم الذكي، معقدة وتشمل طبقات ومعلمات عديدة. تتدرب هذه النماذج على كميات هائلة من البيانات، مما يجعل من الصعب على البشر فهم الأسباب الدقيقة وراء قراراتهم. ويُعرف هذا بتأثير "الصندوق الأسود".

في المجالات عالية المخاطر مثل الرعاية الصحية، أو الأمن السيبراني، أو الدفاع، أو العدالة الجنائية، حيث يتم تطبيق أنظمة الذكاء الاصطناعي بشكل متزايد لاتخاذ قرارات حاسمة تؤثر على حياة الأفراد، فإن الافتقار إلى الشفافية يمكن أن يثير جوانب قلق جدية. قد يتردد الأفراد في الثقة بأنظمة الذكاء الاصطناعي إذا لم يتمكنوا من فهم سبب اتخاذ قرار معين، مما يؤدي إلى انعدام الثقة في التكنولوجيا. علاوة على ذلك، عندما تقع أنظمة الذكاء الاصطناعي في أخطاء

أو تنتج مخرجات متحيزة، يصبح من الصعب تحديد السبب الدقيق أو معالجة المشكلة. إن تزويد أنظمة الذكاء الاصطناعي بالشفافية وقابلية التفسير يعد من الأمور الضرورية لمعالجة هذه المخاوف. يجب أن يدرك مستخدمو الذكاء الاصطناعي بأنهم يتفاعلون مع نظام للذكاء الاصطناعي، ويجب أن يكونوا على علم بقدرات النظام وقيوده.

هناك مجال ناشئ من الذكاء الاصطناعي قابل للتفسير مع أدوات لفهم سبب القرارات المختلفة التي تتخذها أنظمة الذكاء الاصطناعي في مختلف النطاقات.

2.1.8 المساءلة

بخلاف البرامج التقليدية أو عمليات اتخاذ القرار اليدوية، تتمتع أنظمة الذكاء الاصطناعي بطبيعة مستقلة، ويمكن أن يتأثر سلوكها بعوامل مختلفة، منها بيانات التدريب والخوارزميات وتكوين النظام. قد يكون تحديد المساءلة أمرًا صعبًا بسبب الطبيعة الموزعة لمسؤولية تطوير وتطبيق أنظمة الذكاء الاصطناعي. يمكن أن تشارك أطراف متعددة في هذا الأمر، مثل مطوري الذكاء الاصطناعي، ومقدمي البيانات، وجهات تكامل الأنظمة، والمستخدمين النهائيين، وحتى خوارزمية الذكاء الاصطناعي.

فإذا تسبب نظام الذكاء الاصطناعي في حدوث ضرر، فإن عدم وجود أُطر واضحة للمساءلة يمكن أن يؤدي إلى صعوبات في تحديد من يجب أن يتحمل مسؤولية العواقب. تعني المساءلة أنه يجب على المؤسسات القائمة على تطبيق الذكاء الاصطناعي وضع آليات حوكمة تحدد المسؤوليات في كل مرحلة من مراحل تطوير وتطبيق الذكاء الاصطناعي، ويشمل ذلك التوثيق الاستباقي للسياسات والعمليات والتدابير المنفذة.

3.1.8 السلامة

يجب أن تكون أنظمة الذكاء الاصطناعي مرنة وآمنة، تعني السلامة أنه يجب تقييم أنظمة الذكاء الاصطناعي بشكل استباقي لتحديد الأضرار التي قد تنجم عن استخدام النظام، بما في ذلك سوء الاستخدام. كما ويجب اتخاذ التدابير اللازمة للتخفيف من الأضرار.

تصبح السلامة إلزامية، عند استخدام أنظمة الذكاء الاصطناعي في البنية التحتية الحيوية حيث يكون لديها القدرة على الإضرار بحياة البشر والبيئة المحيطة والاقتصاد الوطني بشكل كبير.

4.1.8 الخصوصية

إلى جانب ضمان الامتثال الكامل للوائح حماية الخصوصية والبيانات، يجب أيضًا ضمان آليات مناسبة لحوكمة البيانات، مع مراعاة الجودة والنزاهة والوصول المشروع إلى البيانات.وأما بالنسبة لأنظمة الذكاء الاصطناعي المدمجة في عمليات المؤسسة، ينبغي مراجعة شروط وأحكام عمليات التكامل لضمان الخصوصية، والحصول على الموافقة لأغراض التدريب والتحليل في مجال الذكاء الاصطناعي.

5.1.8 الرقابة والمراقبة البشرية

تعني الرقابة البشرية أنه يجب تصميم وتطوير أنظمة الذكاء الاصطناعي عالية التأثير بطريقة تمكن الأفراد الذين يديرون عمليات النظام من ممارسة رقابة هادفة. تعد المراقبة، من خلال قياس وتقييم أنظمة الذكاء الاصطناعي ومخرجاتها، من الأمور بالغة الأهمية لمساندة الرقابة البشرية الفعالة.

ومن الممكن أن يكون هذا أيضًا متطلبًا تنظيميًا كما هو واضح في اللوائح كاللائحة العامة لحماية البيانات (GDPR)، والتي تمنح أصحاب البيانات الحق في الحصول على التدخل البشري في حالات اتخاذ القرار الآلي.

6.1.8 المتانة

ينبغي أن تكون أنظمة الذكاء الاصطناعي مستقرة ومرنة في مجموعة متنوعة من الظروف، وأن تدافع بشكل فعال ضد الهجمات العدائية، وتقلل من المخاطر الأمنية وتُمكن الثقة في مخرجات الأنظمة.

7.1.8 العدالة والإنصاف

تشير العدالة والإنصاف إلى المعاملة المنصفة للأفراد أو مجموعات الأفراد من خلال نظام الذكاء الاصطناعي. ويتطلب ذلك بناء أنظمة الذكاء الاصطناعي مع الوعي بإمكانية حدوث نتائج تمييزية. كما ويجب اتخاذ الإجراءات المناسبة، بما في ذلك إدخال الفحوصات والتوازنات الكافية في النظام، للتخفيف من النتائج التمييزية للأفراد والمجموعات. ويعد هذا عنصرًا أساسيًا لضمان الاستفادة من فوائد الذكاء الاصطناعي دون تحيز أو مخرجات أخرى غير عادلة قد تحدث دون قصد.

8.1.8 الرفاهية المجتمعية والبيئية

يجب أن تفيد أنظمة الذكاء الاصطناعي جميع البشر، بما في ذلك الأجيال القادمة. ويجب التأكد من أنها مستدامة وصديقة للبيئة. علاوة على ذلك، يجب أن تراعي الكائنات الحية الأخرى، والنظر بعناية في تأثيرها الاجتماعي والمجتمعي.

أما بالنسبة إلى جانب آخر من الرفاهية المجتمعية فهو القدرة على تسخير براعة التكنولوجيا من أجل الرفاهية العامة للبشر والبيئة والمجتمع.

9.1.8 قابلية التدقيق

إن الشفافية والقدرة على اتخاذ القرار المدعوم بالذكاء الاصطناعي يعزز الثقة بشكل كبير في نظام الذكاء الاصطناعي إذ يجب أن تسعى المؤسسات القائمة على تطبيق الذكاء الاصطناعي إلى التأكد من أن أنظمة الذكاء الاصطناعي الخاصة بها قابلة للتدقيق. تتيح مراجعة نظام الذكاء الاصطناعي للجهات الخارجية المعنية استكشاف وفهم ومراجعة سلوك الخوارزمية من خلال الكشف عن المعلومات التي تتيح المراقبة أو الفحص أو النقد.

تشتمل قابلية تدقيق الذكاء الاصطناعي على ما يلي:

1. تقييم النماذج والخوارزميات وتدفقات البيانات
2. تحليل العمليات والنتائج والانحرافات المرصودة
3. الجوانب الفنية لأنظمة الذكاء الاصطناعي من حيث دقة النتائج
4. الجوانب الأخلاقية لأنظمة الذكاء الاصطناعي من حيث العدالة والشرعية والخصوصية

2.8 الاختصارات

ADM	اتخاذ القرار الآلي
AI	الذكاء الاصطناعي
AUP	سياسات المستخدم المقبولة
AS	الأنظمة المستقلة
CRM	إدارة علاقات العملاء
DoS	رفض الخدمة
GPU	وحدة معالجة الرسومات
ML	التعلم الآلي
MfM	النموذج الأساسي للوسائط المتعددة
NLP	المعالجة اللغوية الطبيعية
LLM	النموذج اللغوي الكبير
TPU	وحدة معالجة الموترات

3.8 المراجع المعيارية

القرار الأميري رقم (1) لسنة 2021 بإنشاءالوكالة الوطنية للأمن السيبراني

قرار رئيس الوكالة الوطنية للأمن السيبراني رقم (3) لسنة 2022

سياسة تصنيف البيانات الوطنية الإصدار 3. 0 (V3. 0)

مبادئ توجيهية للتأمين ضد هجمات حجب الخدمة الموزعة

قانون رقم (13) لسنة 2016 بشأن حماية خصوصية البيانات الشخصية

أصدر المكتب الوطني لخصوصية البيانات عدة توجيهات فيما يتعلق بالامتثال لقانون حماية خصوصية البيانات الشخصية، يمكن الاطلاع عليها من https://compliance. qcert. org/en/privacy/hub

قانون رقم (14) لسنة 2014 بإصدار قانون مكافحة الجرائم الإلكترونية

4.8 المراجع الإعلامية

1.4.8 التقارير والأطر

منظمة التعاون الاقتصادي والتنمية (OECD). (2019) توصية مجلس الذكاء الاصطناعي (OECD رقم المرجع OECD/LEGAL/0449)

وزارة الصناعة والعلوم والموارد الأسترالية (2023)؛ الذكاء الاصطناعي الآمن والمسؤول في أستراليا؛ ورقة مناقشة

هيئة تطوير وسائل الإعلام والمعلومات والاتصالات في سنغافورة (IMDA) ولجنة حماية البيانات الشخصية (PDPC). (2020) إطار عمل نموذجي لحوكمة الذكاء الاصطناعي، الطبعة الثانية

الوكالة الأوروبية للأمن السيبراني (ENISA). (2023) الأمن السيبراني للذكاء الاصطناعي والتوحيد القياسي

فريق الخبراء رفيع المستوى التابع للمفوضية الأوروبية المعني بالذكاء الاصطناعي (EU-HLEG). (2019) التوجيهات الأخلاقية للذكاء الاصطناعي الجدير بالثقة

فريق الخبراء رفيع المستوى التابع للمفوضية الأوروبية المعني بالذكاء الاصطناعي (EU-HLEG). (2019) تعريف الذكاء الاصطناعي: القدرات والتخصصات الرئيسية

جارتنر. (2023) تطبيق الذكاء الاصطناعي - الحوكمة وإدارة المخاطر.

2.4.8 المعايير

(2023) .NCSA معيار تأمين المعلومات الوطنية الإصدار 2. 1 (V2. 1)

المعهد الوطني للمعايير والتكنولوجيا. (2023) إطار إدارة مخاطر الذكاء الاصطناعي (AI RMF 1. 0) (معايير المعهد الوطني للمعايير والتكنولوجيا رقم NIST AI 100-1)

المنظمة الدولية للتقييس. (2022) تكنولوجيا المعلومات - حوكمة تكنولوجيا المعلومات - آثار الحوكمة على استخدام الذكاء الاصطناعي من قبل المنظمات (معيار المنظمة الدولية للتقييس رقم 38507:2022)

المنظمة الدولية للتقييس. (2023) تكنولوجيا المعلومات - الذكاء الاصطناعي - توجيهات بشأن إدارة المخاطر (معيار المنظمة الدولية للتقييس رقم 23894:2023)

المنظمة الدولية للتقييس. (2022) تكنولوجيا المعلومات - الذكاء الاصطناعي - مفاهيم ومصطلحات الذكاء

الاصطناعي (معيار المنظمة الدولية للتقييس رقم 22989:2022)

المنظمة الدولية للتقييس. (2023) الأمن والخصوصية في حالات استخدام الذكاء الاصطناعي - أفضل الممارسات (معيار المنظمة الدولية للتقييس رقم 27563:2023)

المعهد الأوروبي لمعايير الاتصالات. (2022) تأمين الذكاء الاصطناعي، وجودية تهديدات الذكاء الاصطناعي (تقرير المعهد الأوروبي لمعايير الاتصالات رقم ETSI GR SAI 001)

المعهد الأوروبي لمعايير الاتصالات. (2020) تأمين الذكاء الاصطناعي، بيان المشكلة (تقرير المعهد الأوروبي لمعايير الاتصالات رقم ETSI GR SAI 004)

المعهد الأوروبي لمعايير الاتصالات. (2021) تأمين الذكاء الاصطناعي، تقرير استراتيجية التخفيف (تقرير المعهد الأوروبي لمعايير الاتصالات رقم ETSI GR SAI 005)

المعهد الأوروبي لمعايير الاتصالات. (2023) تأمين الذكاء الاصطناعي، قابلية التفسير والشفافية لمعالجة الذكاء الاصطناعي (تقرير المعهد الأوروبي لمعايير الاتصالات رقم ETSI GR SAI 007)

3.4.8 الأوراق الأكاديمية

ستيفان لارسون (2020). حول حوكمة الذكاء الاصطناعي من خلال التوجيهات الأخلاقية. المجلة الآسيوية للقانون والمجتمع رقم 7، الصفحات من 437 إلى 451.

ثيلو هاجندورف (2020) أخلاقيات الذكاء الاصطناعي: تقييم التوجيهات. العقول والآلات رقم 30، ص 120-99.

دارون عاصم أوغلو (2021) أضرار الذكاء الاصطناعي؛ ورقة عمل المكتب الوطني للبحوث الاقتصادية رقم 29247.

5.8 الأشكال

www.ncsa.gov.qa
16555 @ info@ncsa.gov.qa