SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

مبادئ أخلاقيات الذكاء الاصطناعي

سبتمبر 2023

الإصدار الأول

المحتويات

مقدمة

نظراً إلى النمو المتسارع الذي تشهده الممارسات والتقنيات المتعلقة بالذكاء الاصطناعي، فقد تنوعت استخدامات الذكاء الاصطناعي لتشمل عديداً من القطاعات، مثل: الصحة والتعليم والترفيه وغيرها، مما أدى إلى تسريع وتيرة عمليات صنع القرار وجعلها أكثر كفاءة ودقة بفضل ما يتيحه من قدرات للتنبؤ بالأنماط المستقبلية، بالإضافة إلى ذلك يمكن استخدام تقنيات الذكاء الاصطناعي لتحليل البيانات، بما في ذلك البيانات الضخمة من خلال إنشاء وتشغيل أنظمة ذات نماذج وخوارزميات أكثر تطوراً تساعد على تحسين جودة العمليات، وفي ضوء الاهتمام المتزايد بهذه التقنيات، قامت جهات عدة في القطاعين العام والخاص، بالإضافة إلى الجهات غير الربحية، بتطوير وتبني حلول رقمية قائمة على الذكاء الاصطناعي تستخدم أساليب مبتكرة لمساعدتها في مواجهة تحدياتها الراهنة، وهو الأمر الذي عَظّم دور الذكاء الاصطناعي في تعزيز القدرات التنافسية لهذه الجهات.

وإشارة إلى الترتيبات التنظيمية للهيئة الصادرة بقرار مجلس الوزراء رقم (٢٩٢) وتاريخ ١٤٤١/٠٤/٢٧هـ، القاضي في الفقرة (١) من المادة "الرابعة" بأن للهيئة على وجه خاص تنظيم قطاعات البيانات والذكاء الاصطناعي من خلال وضع سياسات ومعايير وضوابط خاصة بها وكيفية التعامل معها، وتعميمها على الجهات ذوات العلاقة الحكومية وغير الحكومية، ومتابعة الالتزام بها؛ وفقاً للأحكام النظامية ذات الصلة، وانطلاقاً من التزام المملكة العربية السعودية بحقوق الإنسان وقيمها الثقافية، وتماشياً مع المعايير والتوصيات الدولية بشأن أخلاقيات الذكاء الاصطناعي، عليه فقد قامت الهيئة بالاستفادة من الممارسات والمعايير العالمية عند وضع مبادئ أخلاقيات الذكاء الاصطناعي التي تهدف إلى:

دعم وتعزيز جهود المملكة في تحقيق رؤيتها واستراتيجياتها الوطنية المتعلقة بتبني تقنيات الذكاء الاصطناعي وتشجيع البحث والابتكار وتعزيز النمو الاقتصادي.

وضع المبادئ التوجيهية المتعلقة بأخلاقيات الذكاء الاصطناعي.

حوكمة نماذج الذكاء الاصطناعي للحد من الآثار السلبية لها (اقتصادياً واجتماعياً وغير ذلك) والمخاطر المحتملة التي قد تنتج عنها.

مساعدة الجهات في تبني المعايير والأخلاقيات عند بناء وتطوير الحلول القائمة على الذكاء الاصطناعي لضمان الاستخدام المسؤول لها.

حماية خصوصية أصحاب البيانات وحقوقهم فيما يتعلق بمعالجة بياناتهم الشخصية.

التعريفات

يُقصد بالعبارات الواردة أدناه المعاني الموضحة أمام كل منها، ما لم يقتضِ سياق النص خلاف ذلك:

الهيئة

الهيئة السعودية للبيانات والذكاء الاصطناعي.

الأخلاقيات

مجموعة من القيم والمبادئ والأساليب لتوجيه السلوك الأخلاقي في تطوير تقنيات الذكاء الاصطناعي واستخدامها.

نظام أو نموذج الذكاء الاصطناعي

مجموعة من النماذج التنبؤية والخوارزميات المتقدمة التي يمكن استخدامها لتحليل البيانات والتنبؤ بالمستقبل أو تسهيل عملية صنع القرار للأحداث المستقبلية المتوقعة.

مطور نظام الذكاء الاصطناعي

أي شخص ذي صفة طبيعية أو اعتبارية يقوم بتطوير أنظمة الذكاء الاصطناعي.

مقيم نظام الذكاء الاصطناعي

أي شخص ذي صفة طبيعية أو اعتبارية يقوم بتدقيق أنظمة الذكاء االصطناعي لتحقيق أهداف معينة.

البيانات

مجموعة من الحقائق في صورتها الأولية أو في صورة غير منظّمة مثل الأرقام، أو الحروف، أو الصور، أو الفيديو، أو التسجيلات الصوتية أو الرموز التعبيرية.

الجهات المطبقة

أي جهة عامة أو خاصة أو فرد يتعين عليه الالتزام بهذه الأخلاقيات.

الذكاء الاصطناعي

مجموعة من التقنيات التي تمكن آلة أو نظاماً من التعلم، والفهم، والتصرف والاستشعار.

دورة نظام الذكاء الاصطناعي

العملية الدورية التي يتوقع من مطوري الذكاء الاصطناعي اتباعها لتصميم وبناء وإنتاج نظام قوي وآمن يقدم قيمة عملية ورؤى من خلال الالتزام بطريقة موحدة ومنظمة لإدارة تنفيذ وتسليم نموذج الذكاء لاصطناعي.

مسؤول نظام الذكاء الاصطناعي

أي شخص ذي صفة طبيعية أو اعتبارية يدير أو يطبق أنظمة الذكاء الاصطناعي أو يستخدمها لتحقيق أهداف معينة.

المدير التنفيذي للبيانات

مدير مكتب إدارة البيانات في الجهة الحكومية، أو المسؤول عن تطوير وإدارة البيانات وتنفيذ الحوكمة والإشراف على تنفيذ ممارسات إدارة البيانات في الجهة غير الحكومية، ويكون مسؤولاً عن وضع المعايير الأخلاقية ومعايير الالتزام التي يجب أن تتبعها الجهة وتحافظ عليها عند بناء أو تطوير أنظمة ذكاء اصطناعي واستخدامها.

عينة البيانات

جزء من البيانات المستخدمة في بناء النماذج التنبؤية وخوارزميات الذكاء الاصطناعي وتدريبها واختبارها للوصول إلى نتائج محددة.

المستخدم النهائي

أي شخص ذي صفة طبيعية أو اعتبارية يستهلك أو يستخدم السلع أو الخدمات التي تنتجها أنظمة الذكاء الاصطناعي.

أصحاب البيانات الشخصية

الفرد الذي تتعلق به البيانات الشخصية.

البيانات الشخصية

كل بيان -مهما كان مصدره أو شكله- من شأنه أن يؤدي إلى معرفة الفرد على وجه التحديد، أو يجعل التعرف عليه ممكناً بصفة مباشرة أو غير مباشرة، ومن ذلك: الاسم، ورقم الهوية الشخصية، والعناوين، وأرقام التواصل، وأرقام الرُّخص والسجلات والممتلكات الشخصية، وأرقام الحسابات البنكية والبطاقات الائتمانية، وصور الفرد الثابتة أو المتحركة، وغير ذلك من البيانات ذات الطابع الشخصي.

البيانات الحساسة

كل بيان شخصي يتضمن الإشارة إلى أصل الفرد العرقي، أو أصله الإثني، أو معتقده الديني، أو الفكري، أو السياسي. وكذلك البيانات الأمنية والجنائية، أو بيانات السمات الحيوية التي تحدد الهوية، أو البيانات الوراثية، أو البيانات الصحية، والبيانات التي تدل على أن الفرد مجهول الأبوين أو أحدهما.

الأطراف الخارجية

أي شخصية طبيعية أو اعتبارية، عامة أو خاصة بخلاف المشاركين الرئيسيين لدى المستخدم النهائي لنظام الذكاء الاصطناعي، ومسؤول نظام الذكاء الاصطناعي، ومطور نظام الذكاء الاصطناعي ومقيم نظام الذكاء الاصطناعي.

الموثوقية

تشير الموثوقية إلى اتساق المقياس المستخدم، أي ما إذا كان يمكن الوصول إلى النتائج ذاتها في ظل وجود ظروف مماثلة.

صحة القياس

تعني ما إذا كانت النتائج بالفعل تقيس ما يفترض أن تقيسه.

نطاق التطبيق

تطبق المبادئ على جميع الجهات العامة والخاصة وغير الربحية والأفراد الذين يقومون بتطوير أو تبني الحلول المعتمدة على تقنيات الذكاء الاصطناعي.

مخاطر أنظمة الذكاء الاصطناعي

تصنف فئات ومستويات المخاطر المرتبطة بتطوير و/أو استخدام تقنيات الذكاء الاصطناعي إلى كل من الآتي:

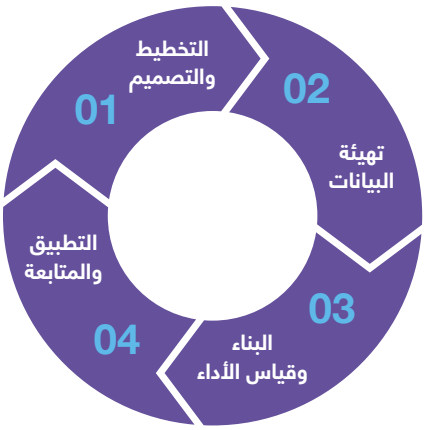
- **مخاطر بسيطة أو منعدمة:** لا يوجد أي قيود على أنظمة الذكاء الاصطناعي التي تشكل مخاطر بسيطة أو لا تنطوي على أي مخاطر مثل مرشحات البريد العشوائي غير المرغوب فيه، ولكن يوصى بأن تكون هذه الأنظمة متوافقة مع الأخلاقيات.
- **مخاطر محدودة:** تخضع أنظمة الذكاء الاصطناعي التي تشكل مخاطر محدودة مثل البرامج التقنية المتعلقة بالوظيفة والتطوير والأداء إلى تطبيق مبادئ الأخلاقيات المذكورة في هذه الوثيقة.
- **مخاطر عالية:** يتعين على أنظمة الذكاء الاصطناعي التي تشكل "مخاطر عالية" على الحقوق الأساسية للإنسان الخضوع لإجراء تقييمات ما قبل المطابقة وبعدها، وإضافة إلى الالتزام بالأخلاقيات يجب مراعاة المتطلبات النظامية ذات العلاقة.
- **مخاطر غير مقبولة:** لا يُسمح بأنظمة الذكاء الاصطناعي التي تشكل "خطراً غير مقبول" على سلامة الناس وسبل عيشهم وحقوقهم كتلك المتعلقة بالتصنيف الاجتماعي أو استغلال الأطفال أو تشويه السلوك الذي يحتمل أن تحدث نتيجة عنه أضرار جسدية أو نفسية وإضافة إلى الالتزام بالأخلاقيات يجب مراعاة المتطلبات النظامية ذات العلاقة.

هذا وينبغي أن تكون إدارة المخاطر مرتبطة ارتباطاً مباشراً بمبادرات الذكاء الاصطناعي، حيث تكون الرقابة متزامنة مع عمليات التطوير الداخلي لتقنيات الذكاء الاصطناعي، وتؤثر إدارة مخاطر أنظمة الذكاء الاصطناعي على مجموعة واسعة من أنواع المخاطر بما في ذلك البيانات والخوارزمية والالتزام والمخاطر التشغيلية والقانونية وتلك المتعلقة بالسمعة والمخاطر التنظيمية، ويتم بناء المكونات الفرعية لإدارة المخاطر، مثل: قابلية النموذج للتفسير، والكشف عن التحيز، ومراقبة الأداء، حيث تكون المراقبة ثابتة ومتسقة مع أنشطة تطوير الذكاء الاصطناعي.

دورة حياة نظام الذكاء الاصطناعي

دورة حياة نظام الذكاء الاصطناعي هي المنهجية التي يتم اتباعها عند تنفيذ مشاريع الحلول التقنية المعتمدة على تقنيات الذكاء الاصطناعي، والتي من خلالها يتم تحديد الخطوات التي يجب على الجهة اتباعها عند تنفيذ وإدارة نماذج أو أنظمة الذكاء الاصطناعي للاستفادة من هذه التقنيات وتحقيق القيمة منها، بشكل يضمن المواءمة مع أخلاقيات الذكاء الاصطناعي.

تنقسم دورة حياة نظام الذكاء الاصطناعي إلى أربع مراحل رئيسية، وتتضمن كل مرحلة من المراحل عدداً من الأنشطة الرئيسية، وذلك على النحو التالي:



المرحلة الأولى: التخطيط والتصميم

- تحديد المشكلة
- وضع الحلول المقترحة
- اختيار تقنية الذكاء الاصطناعي بما يتناسب مع الحلول المقترحة
- دراسة المخاطر المرتبطة بالحلول المقترحة وجدوى البدائل المحتملة
- تطوير مؤشرات الأداء المناسبة

المرحلة الثانية: تهيئة البيانات

- جمع البيانات
- استكشاف وتقييم البيانات
- تنظيف البيانات والتحقق من صحتها
- تجويد البيانات
- تحويل البيانات إلى صيغة تناسب مدخلات نموذج الذكاء الاصطناعي

المرحلة الثالثة: البناء وقياس الأداء

- تدريب واختبار النموذج
- ضبط المتغيرات أو مدخلات النموذج
- التحقق من أداء النموذج
- تقييم المخاطر

المرحلة الرابعة: التطبيق والمتابعة

- تطبيق النموذج على نظام الذكاء الاصطناعي
- تعريف الإصدارات
- مراقبة أداء النموذج بشكل دوري
- تقييم مدى الحاجة إلى تغيير التصميم وفقاً لنتائج المراجعات الدورية

مبادئ أخلاقيات الذكاء الاصطناعي

المبدأ الأول - النزاهة والإنصاف

على مطور نظام الذكاء الاصطناعي عند تصميم أو جمع أو تطوير أو نشر أو استخدام أنظمة الذكاء الاصطناعي، اتخاذ الإجراءات اللازمة للتأكد من عدم وجود التحيز أو التمييز أو التنميط أو الحد منها التي يتعرض لها الأفراد أو الجماعات أو الفئات بسبب البيانات أو الخوارزميات ويمكن أن تؤدي إلى تمييز سلبي لفئة محددة.

عند تصميم واختيار وتطوير أنظمة الذكاء الاصطناعي، من الضروري اتخاذ ما يلزم لضمان تطوير معايير غير متحيزة، وعادلة ومنصفة وموضوعية وشاملة ومتنوعة وممثلة لجميع شرائح المجتمع أو الشرائح المستهدفة، ويجب ألا تقتصر وظيفة نظام الذكاء الاصطناعي على مجموعات محددة بناء على أساس الجنس أو العرق أو الدين أو العمر أو غيره. إضافة إلى ذلك، يجب عند استخدام البيانات الشخصية أن يكون الغرض من استخدامها مبرراً ومحدداً بشكل دقيق من قبل مطور نظام الذكاء الاصطناعي، مع التأكد من أن ذلك لا يخالف نظام حماية البيانات الشخصية ولوائحه التنفيذية، وأن يقوم مطور نظام الذكاء الاصطناعي بإخفاء هوية أصحاب البيانات الشخصية أو ترميزها ما أمكن ذلك.

لضمان بناء أنظمة ذكاء اصطناعي قائمة على الإنصاف والشمولية؛ يتم تدريب أنظمة الذكاء الاصطناعي على البيانات التي يتم تنقيحها من التحيز، كما يتم بناء وتطوير الخوارزميات بطريقة تجعل تكوينها خالياً من التحيز والمغالطات.

التخطيط والتصميم

١- في هذه المرحلة يتم تحديد الغرض والأهداف من نظام الذكاء الاصطناعي وكيفية تحقيقها، والخروج بتصميم يتسم بالنزاهة والإنصاف ويأخذ الاحتياطات المناسبة لمنع التحيز والتمييز والتنميط الذي قد يؤدي إلى إحداث أي أضرار.

٢- يبدأ التصميم المراعي للنزاهة والإنصاف من بداية دورة حياة نظام الذكاء الاصطناعي من خلال جهود تكاملية بين الأعضاء الفنيين وغير الفنيين وذلك لتحديد الفوائد والأضرار المحتملة والأفراد المحتمل تضررهم والفئات غير الممثلة في النظام وتقييم مدى تأثرهم بالنتائج وما إذا كان التأثير مبرراً في ظل الهدف العام من نظام الذكاء الاصطناعي.

٣- يعد تقييم نزاهة نظام الذكاء الاصطناعي خطوة بالغة الأهمية، ولضمان نزاهة النظام أو النموذج يتم اختيار المقاييس بناءً على ما يلي: نوع الخوارزمية (مبنية على قاعدة، أو تصنيف، إلخ)، تأثير القرار والضرر أو الفائدة التي ستعود على العينات المتوقعة بشكل صحيح أو غير صحيح.

٤- كما يتم في هذه المرحلة تحديد وتعريف البيانات المتعلقة بالأشخاص أو الفئات التي تتأثر من النظام بصورة منهجية، ووضع الحد الذي يكون عنده التقييم عادلاً أو غير عادل، وتحديد مقاييس تقييم النزاهة التي سيتم تطبيقها على البيانات والنظام خلال الخطوات المستقبلية.

تهيئة البيانات

١- يتم اتباع أفضل الممارسات في الحصول على البيانات والتعامل معها وتصنيفها وإدارتها لضمان توافق النتائج مع الأهداف والغايات المحددة لنظام الذكاء الاصطناعي، وتتحقق جودة البيانات من خلال ضمان حداثتها وسلامة مصدرها، كما يتم التأكد من عدم وجود استبعاد منظّم للفئات غير الممثلة، ويجب أن تكون كمية ونوعية مجموعات البيانات كافية ودقيقة لخدمة الغرض من النظام، حيث يؤثر حجم عينة البيانات التي تم جمعها أو الحصول عليها تأثيراً كبيراً في دقة وعدالة مخرجات النموذج تحت التطوير.

٢- يجب ألا تدرج البيانات الحساسة المحددة في مرحلة التخطيط والتصميم في خصائص النموذج، وذلك لكي لا تزيد التحيز ضدها؛ لذا يجب تحليل خصائص البيانات الحساسة وعدم إدراجها في بيانات المدخلات، ويستثنى من ذلك بعض الحالات التي قد لا يكون ذلك ممكناً بسبب دقة أو هدف نظام الذكاء الاصطناعي، وفي هذه الحالة يجب تقديم مبررات لاستخدام البيانات الحساسة وتحديد بديل عنها.

البناء وقياس الأداء

١- في هذه المرحلة تتم مراعاة النزاهة في التنفيذ باعتبارها عاملاً أساسياً عند بناء نظام الذكاء الاصطناعي واختباره وتنفيذه، ويتطلب بناء النموذج واختيار الخصائص أن يكون المهندسون والمصممون على دراية بأن الخيارات المتخذة بشأن تجميع أو فصل أو استبعاد الخصائص قد يكون له عواقب سلبية على الفئات الضعيفة أو غير الممثلة في البيانات.

٢- يجب عند اختيار النموذج النظر في مقاييس النزاهة والإنصاف، حيث تكون مقاييس النزاهة والإنصاف في النموذج ضمن الحد المحدد للخصائص الحساسة، كما يجب تحديد منهجية التقييم الخاص بالنزاهة والإنصاف ومقاييس الأداء بوضوح خلال هذه المرحلة، وأن يتم حصر الأسباب والمبررات إذا لم يجتاز النموذج التقييم.

٣- من الضروري أن يتم في هذه المرحلة التأكد من صلاحية المنهجية المتبعة في اختيار خصائص النموذج، كما يجب التحقق من الخصائص المختارة مع الأطراف أصحاب العلاقة.

٤- يجب وضع آليات لمنع أو الحد من النتائج السلبية أو غير المرغوب فيها عند استخدام الذكاء الاصطناعي في أتمتة أنظمة دعم القرار التي قد تؤدي إلى ضرر أو تمييز في هذه المرحلة.

التطبيق والمتابعة

١- يتم وضع آليات وبروتوكولات واضحة عند التطبيق الفعلي لنظام الذكاء الاصطناعي وذلك لقياس نزاهة النتائج وأدائها وكيفية تأثيرها في مختلف الأفراد والجماعات وعند تحليل نتائج النموذج التنبؤي، كما يتم تقييم المجموعات الممثلة في عينة البيانات والتأكد من أنها تتأثر بشكل متساوٍ أو مماثل، أو توضيح ما إذا كان نظام الذكاء الاصطناعي قد يضر بفئة محددة وذلك لضمان تحقيق العدالة في النتائج.

٢- تتم مراقبة مقاييس النزاهة والإنصاف المحددة مسبقاً، وإذا كان هناك أي انحراف عن الحد المسموح به فيجب التحقق ما إذا كانت هناك حاجة إلى تعديل النموذج.

٣- يتم قياس الضرر العام، والمنفعة المتحققة من النظام، واتخاذ الإجراءات اللازمة وفقاً لذلك.

المبدأ الثاني - الخصوصية والأمن

يتم تطوير أنظمة الذكاء الاصطناعي لتكون محمية بطريقة آمنة وتراعي المتطلبات النظامية ذات العلاقة، ومن ذلك المتطلبات النظامية المتعلقة بحماية خصوصية أصحاب البيانات الشخصية، ومعايير الأمن السيبراني ذات العلاقة؛ بهدف منع الوصول غير المشروع إلى البيانات والنظام مما قد يؤدي إلى الإضرار بالسمعة أو الأضرار النفسية أو المالية أو المهنية.

ويتم تصميم أنظمة الذكاء الاصطناعي باستخدام آليات وضوابط توفر إمكانية إدارة ومراقبة النتائج والتقدم الذي يتم طوال دورتها لضمان امتثالها بقواعد وضوابط الخصوصية والأمن ذات العلاقة.

التخطيط والتصميم

١- يتم تطوير وتصميم نظام الذكاء الاصطناعي والخوارزمية المرتبطة به بطريقة يمكن من خلالها حماية خصوصية الأفراد، وعدم إساءة استخدام البيانات الشخصية، وضمان عدم استناد معايير اتخاذ القرارات في التقنية الآلية إلى خصائص أو معلومات تحدد الهوية الشخصية، ويقتصر استخدام البيانات الشخصية على الحد الأدنى اللازم لتشغيل النظام بشكل سليم.

٢- يتم تصميم أنظمة الذكاء الاصطناعي التي تؤدي إلى تحديد سمات أفراد أو مجموعات محددة فقط في حالة الموافقة على ذلك من قبل المدير التنفيذي للبيانات أو وفقاً لمدونة قواعد السلوك المهني التي تطورها الجهة التنظيمية لقطاع معين.

٣- تتم مواءمة مخطط الأمن والحماية لنظام الذكاء الاصطناعي والبيانات التي تتم معالجتها والخوارزمية التي يتم استخدامها مع أفضل الممارسات لتكون هذه الأنظمة قادرة على التصدي للهجمات السيبرانية ومحاولات الوصول غير المشروع إلى البيانات أو النموذج أو أوزان النموذج.

٤- يتم اتباع الأطر والمتطلبات النظامية للخصوصية والأمن وتهيئتها بما يتناسب مع النظام أو الجهة المعنية.

٥- يتم التخطيط لتصنيف البيانات وتحديد خصائصها من أجل تحديد مستويات الحماية اللازمة وآلية استخدام البيانات الشخصية، كما يتم التخطيط للآليات الأنسب لإخفاء هوية أصحاب البيانات الشخصية أو ترميزها.

٦- يتم إجراء تقويم الأثر قبل معالجة البيانات الشخصية أو البيانات الحساسة وفقاً لما ينص عليه نظام حماية البيانات الشخصية ولوائحه.

تهيئة البيانات

١- عند جمع أو شراء أو إدارة أو تنظيم البيانات يجب الالتزام بالأطر والمتطلبات النظامية المتعلقة بخصوصية أصحاب البيانات الشخصية.

٢- يتم حصر الوصول إلى المعلومات على الأشخاص المصرح لهم بذلك، ووضع ضوابط محددة لإدارة تفويض صلاحيات الوصول إلى المعلومات والبيانات.

٣- يجب أن يتمتع مطور نظام الذكاء الاصطناعي بالنزاهة والمعرفة، وذلك لضمان اتباع الأطر والمعايير النظامية للخصوصية والأمن، كما يجب التأكد من توفير أنظمة قواعد بيانات آمنة.

٤- يتم تصنيف جميع البيانات المعالجة لضمان حصولها على المستوى المناسب من الحماية وفقاً لحساسيتها أو تصنيفها، ويجب أن يكون مطور نظام الذكاء الاصطناعي والمسؤول عنه على دراية بتصنيف أو حساسية البيانات التي يتم التعامل معها والمتطلبات المرتبطة بها للحفاظ على أمنها وسريتها وخصوصية أصحابها، وتُصنف جميع البيانات حسب متطلبات الأعمال وأهميتها وحساسيتها لمنع الإفصاح غير المصرح به عنها أو تعديلها بشكل غير صحيح، وتُصنف البيانات بطريقة لا تؤدي إلى استنتاج المعلومات الشخصية.

٥- يتم اتخاذ إجراءات النسخ الاحتياطي للبيانات وتشفيرها وأرشفتها في هذه المرحلة لضمان استمرارية الأعمال والحد من الكوارث وتخفيف المخاطر.

البناء وقياس الأداء

١- تتم حماية الأبعاد التصميمية المختلفة لنموذج الذكاء الاصطناعي من المخاطر المحتملة، ويعمل مطور نظام الذكاء الاصطناعي على حماية هيكل ووحدات النظام من التلف أو التعديل أو الدخول غير المصرح به لأي من مكوناته.

٢- يعمل مطور نظام الذكاء الاصطناعي على أن يكون النظام آمناً حيث يظل فعالاً وجاهزاً للاستخدام من قبل المستخدمين المصرح لهم ومحافظاً على أمن المعلومات في كافة الأحوال، بالإضافة إلى ذلك يجب وضع ضوابط حماية مناسبة لضمان تقيّد أنظمة اتخاذ القرار بالذكاء الاصطناعي بمتطلبات خصوصية الأفراد وأمن البيانات ذات الصلة، وينبغي اختبار نظام الذكاء الاصطناعي للتأكد من أن البيانات المتاحة لا تفصح عن البيانات الشخصية أو الحساسة بشكل غير نظامي أو تنتهك قواعد إخفاء الهوية أو الترميز.

التطبيق والمتابعة

١- بعد تشغيل نظام الذكاء الاصطناعي، يجب أن تكون هناك متابعة مستمرة لضمان الحفاظ على الخصوصية في النظام وضمان سلامته وأمنه، وتتم إعادة النظر في تقييم أثر الخصوصية وتقييم إدارة المخاطر باستمرار لضمان التقييم المنتظم للاعتبارات النظامية والاجتماعية والأخلاقية.

٢- يكون مسؤول نظام الذكاء الاصطناعي مسؤولاً عن تصميم وتنفيذ أنظمة الذكاء الاصطناعي بما يضمن حماية البيانات الشخصية طوال دورة نظام الذكاء الاصطناعي، ويتم تحديث عناصر نظام الذكاء الاصطناعي بناءً على تقارير المتابعة المستمرة.

المبدأ الثالث - الإنسانية

يسلط مبدأ الإنسانية الضوء على ضرورة بناء أنظمة الذكاء الاصطناعي باستخدام منهجية عادلة وأخلاقية تستند إلى حقوق الإنسان والقيم الثقافية الأساسية وذلك لإحداث أثر إيجابي على الأطراف المعنية والمجتمعات المحلية والمساهمة في تحقيق الأهداف والغايات طويلة وقصيرة الأجل من أجل مصلحة البشرية وازدهارها، ومن الضروري أن يتم تصميم أنظمة الذكاء الاصطناعي، حيث لا تخدع، أو تتلاعب، أو تضع سلوكاً لا يقصد به تمكين المهارات البشرية، أو تعزيزها، أو زيادتها، بل ينبغي لها أن تتبنى نهجاً تصميمياً أكثر تركيزاً على إتاحة الاختيار واتخاذ القرار لمصلحة الإنسان.

التخطيط والتصميم

١- من الضروري تصميم وبناء نموذج قائم على حقوق الإنسان الأساسية والقيم والمبادئ الثقافية وتطبيقه على قرارات وعمليات ووظائف نظام الذكاء الاصطناعي، ويتعين ذلك على مصممي نموذج الذكاء الاصطناعي.

٢- تحديد الكيفية التي سيتوافق بها نظام الذكاء الاصطناعي مع حقوق الإنسان الأساسية والقيم الثقافية والاجتماعية في المملكة العربية السعودية، إضافة إلى تحديد التقنيات اللازمة واختبارها، مع تحديد الآلية التي سيسعى من خلالها نظام الذكاء الاصطناعي ونتائجه إلى تعزيز المهارات والقدرات البشرية.

تهيئة البيانات

١- لضمان تبني نماذج الذكاء الاصطناعي لهيكل وتصميم يركز على مبدأ الإنسانية، يجب الالتزام بممارسات إدارة البيانات بشكل أخلاقي وكذلك المعايير والضوابط الخاصة بإدارة البيانات في المملكة.

٢- يتم الحصول على البيانات وتصنيفها ومعالجتها وإتاحتها بشكل صحيح لضمان احترام حقوق الإنسان والقيم الثقافية والاجتماعية في المملكة العربية السعودية.

البناء وقياس الأداء

١- يجب على المصممين والمهندسين إعطاء الأولوية لبناء أنظمة وخوارزميات الذكاء الاصطناعي التي تتمحور حول الإنسان وتسمح وتسهل عملية صنع القرار وتراعي التوافق مع حقوق الإنسان والقيم الثقافية للمملكة، حيث لا تعمل القرارات المؤتمتة الناتجة عن أنظمة الذكاء الاصطناعي بطريقة مستقلة دون مراعاة حقوق الإنسان والقيم الاجتماعية والثقافية.

٢- يجب على المصممين والمهندسين تطوير أنظمة الذكاء الاصطناعي باستخدام المعايير الأخلاقية وتدريب الخوارزميات لتحقيق النتائج التي تنهض بالإنسانية.

التطبيق والمتابعة

١- يتم إجراء تقييمات دورية لنظام الذكاء الاصطناعي لضمان عدم تعارض نتائجه مع حقوق الإنسان والقيم الاجتماعية والثقافية، وللتأكد من دقة مؤشرات الأداء الرئيسية، ولرصد تأثيره على الأفراد أو الجماعات وذلك لضمان التحسين المستمر للتقنية.

٢- على مصممي نماذج الذكاء الاصطناعي أن يضعوا آليات لتقييم أنظمة الذكاء الاصطناعي من حيث القيم الثقافية وحقوق الإنسان الأساسية للحد من أي نتائج سلبية وضارة ناتجة عن استخدام النظام، وفي حال رصد أي نتائج سلبية وضارة يجب على مسؤول نظام الذكاء الاصطناعي تحديد المجالات التي تحتاج إلى معالجة وتطبيق التدابير التصحيحية لتحسين أداء نظام الذكاء الاصطناعي ونتائجه ومتابعة ذلك بشكل دوري ومستمر.

المبدأ الرابع - المنافع الاجتماعية والبيئية

يسعى مبدأ المنافع الاجتماعية والبيئية إلى تعزيز الأثر الإيجابي والمفيد للأولويات الاجتماعية والبيئية التي يجب أن تفيد الأفراد والمجتمع ككل والتي تركز على الأهداف والغايات المستدامة، لا ينبغي لأنظمة الذكاء الاصطناعي أن تسبب أو تسرع الضرر أو تؤثر سلباً على البشر، بل يجب أن تسهم في تمكين واستكمال التقدم التقني والاجتماعي والبيئي مع السعي إلى معالجة التحديات المرتبطة بها.

التخطيط والتصميم

١- تؤثر أنظمة الذكاء الاصطناعي تأثيراً كبيراً في المجتمعات والمنظومات الموجودة لديها، وبالتالي يجب أن يكون لدى مطوري ومسؤولي أنظمة الذكاء الاصطناعي وعي كافٍ بأن هذه التقنيات قد تكون لها آثار ضارّة أو تحولية في المجتمع والبيئة، ويجب التعامل مع تصميم أنظمة الذكاء الاصطناعي بطريقة أخلاقية لمنع الضرر على البشر والبيئة.

٢- عند تخطيط وتصميم أنظمة الذكاء الاصطناعي، يجب الأخذ بعين الاعتبار المسائل الاجتماعية والبيئية ذات العلاقة والعمل على معالجتها بطريقة مسؤولة.

تهيئة البيانات

١- يتم اتباع الإجراءات والسياسات المنظمة لإدارة البيانات عند تصنيف وهيكلة البيانات التي ستغذي نظام الذكاء الاصطناعي.

٢- ينبغي أن تكون البيانات المتعلقة بالمواضيع الاجتماعية والبيئية متاحة للهياكل الأساسية للبيانات ويجب أن تبين بوضوح المنفعة الاجتماعية للبيانات المعروضة.

البناء وقياس الأداء

١- يجب أن يكون الهدف النهائي أو الآثار الاجتماعية أو البيئية للنماذج والخوارزميات قادرة على إظهار ارتباطها بالنتائج المتوقعة والفوائد الانتقالية والمؤثرة.

٢- كما يمكن تحديد الأسلوب الذي ستسعى من خلاله أنظمة الذكاء الاصطناعي إلى معالجة المخاوف المتعلقة بالقضايا الاجتماعية والبيئية.

التطبيق والمتابعة

١- بعد تشغيل نظام الذكاء الاصطناعي، يجب على مسؤول نظام الذكاء الاصطناعي أن يضمن إجراء تقييم مستمر للأثر الاجتماعي والثقافي والاقتصادي والبيئي لتقنيات الذكاء الاصطناعي، مع الإدراك الكامل لآثار نظام الذكاء الاصطناعي على الاستدامة كهدف يجب متابعته وتطويره باستمرار عبر مجموعة من الأهداف ذات الأولوية التي تم وضعها في مرحلة التخطيط والتصميم.

٢- الحرص على تعزيز وتشجيع قدرة حلول الذكاء الاصطناعي في معالجة المجالات المتعلقة بأهداف التنمية المستدامة.

المبدأ الخامس - الموثوقية والسلامة

يسعى مبدأ الموثوقية والسلامة إلى ضمان التزام نظام الذكاء الاصطناعي بالمواصفات المحددة وأن نظام الذكاء الاصطناعي يعمل بشكل كامل وفق الآلية التي كان يقصدها ويتوقعها مصممو، وتمثل الموثوقية مقياساً للمصداقية والاعتمادية التي يتمتع بها النظام من الناحية التشغيلية مع وظائفه المحددة والنتائج التي يسعى إلى تحقيقها.

من ناحية أخرى تمثل السلامة مقياساً للكيفية التي لا يشكل بها نظام الذكاء الاصطناعي خطراً على المجتمع والأفراد، ومن ذلك على سبيل المثال يمكن لأنظمة الذكاء الاصطناعي مثل المركبات ذاتية القيادة أن تشكل خطراً على حياة الناس في حال عدم التعرف عليهم ككائنات حية أو في حالة عدم تدريب هذه المركبات على بعض السيناريوهات أو حالات تعطل النظام. وعليه يجب أن يكون النظام موثوقاً وآمناً من خلال عدم تعريض المجتمع للخطر ويجب أن تكون لديه آليات مدمجة لمنع وقوع الضرر، لذا يرتبط إطار الحد من المخاطر ارتباطاً وثيقاً بهذا المبدأ، وينبغي على مسؤول نظام الذكاء الاصطناعي العمل على تقليل المخاطر المحتملة والأضرار غير المقصودة إلى أدنى حد ممكن.

التخطيط والتصميم

١- يتم في هذه المرحلة العمل على تصميم وتطوير نظام ذكاء اصطناعي يمكنه تحمل عدم الاستقرار والتقلبات التي قد يتعرض لها.

٢- يعد وضع نظام ذكاء اصطناعي قوي وموثوق يعمل مع مجموعات مختلفة من المدخلات والمواقف أمراً ضرورياً لمنع الضرر غير المقصود والحد من المخاطر التي قد تعطل النظام عند مواجهة أحداث جديدة غير معروفة وغير متوقعة.

٣- من الضروري وضع مجموعة من المعايير والبروتوكولات التي تقيم موثوقية نظام الذكاء الاصطناعي لضمان سلامة خوارزمية النظام ومخرجات البيانات والحفاظ على النفقات الفنية المستدامة ونتائج النظام للحفاظ على ثقة المستخدمين في نظام الذكاء الاصطناعي.

٤- تعد معايير التوثيق ضرورية لتتبع تطور النظام وتوقع المخاطر المحتملة ومعالجة الثغرات، ويجب أن تخضع جميع نقاط القرار المهمة في تصميم النظام لموافقة أصحاب العلاقة للحد من المخاطر ووضع المسؤولية على متخذي تلك الموافقات.

٥- تتم مراعاة مستويات المخاطر الخاصة بنظام الذكاء الاصطناعي، واتخاذ الإجراءات والضوابط اللازمة وفقاً لمستوى المخاطر المذكور سابقاً.

تهيئة البيانات

١- يتم اتخاذ الخطوات والإجراءات المناسبة لقياس عينة البيانات وجودتها ودقتها وملاءمتها وموثوقيتها عند التعامل مع مجموعات البيانات لنموذج الذكاء الاصطناعي، ويعد ذلك ضرورياً لضمان اتساق البيانات ودقة تفسيرها من قبل نظام الذكاء الاصطناعي، وتجنب حدوث قياسات مضللة، وضمان ارتباط نتائج نظام الذكاء الاصطناعي بغرض النموذج.

٢- يجب أن يتم وضع خطوة للتحقق من كيفية عمل النظام في ظل الأحداث الطارئة والسيناريوهات غير المتوقعة.

البناء وقياس الأداء

١- لتطوير نظام ذكاء اصطناعي سليم وظيفياً وآمن وموثوق في الوقت نفسه، يجب أن يكون الهيكل الفني لنظام الذكاء الاصطناعي مصحوباً بمنهجية شاملة لاختبار جودة الأنظمة والنماذج التنبؤية القائمة على البيانات وفقاً لسياسات وبروتوكولات موحدة لدى مطور النظام.

٢- لضمان الاعتمادية الفنية لنظام الذكاء الاصطناعي، يجب اختباره والتحقق منه وإعادة تقييمه بشكل دقيق، بالإضافة إلى دمج آليات الإشراف والضوابط المناسبة لذلك ضمن عملية تطويره، وتلزم الموافقة على اختبار تكامل النظام من قبل أصحاب العلاقة المعنيين للحد من المخاطر وتحديد المسؤولية.

٣- يجب أن يتوفر إشراف بشري على أنظمة الذكاء الاصطناعي التي تتضمن أعمالها اتخاذ قرارات لها تأثير لا يمكن تداركه أو تنطوي على قرارات تتعلق بالحفاظ على الحياة، إضافة إلى ذلك، لا ينبغي استخدام أنظمة الذكاء الاصطناعي لأغراض التقييم الاجتماعي أو المراقبة واسعة النطاق.

التطبيق والمتابعة

١- تتم مراقبة مصداقية واعتمادية نظام الذكاء الاصطناعي بطريقة دورية ومستمرة لقياس وتقييم أي مخاطر تتعلق بالجوانب الفنية لنظام الذكاء الاصطناعي (من منظور داخلي)، بالإضافة إلى قياس حجم المخاطر التي يشكلها النظام وقدراته (من منظور خارجي).

٢- كما يجب أن تتم مراقبة النموذج بطريقة دورية ومستمرة للتحقق مما إذا كانت عملياته ووظائفه متوافقة مع الهيكل والأطر المصممة، كما يجب أن يكون نظام الذكاء الاصطناعي سليماً وقوياً ومتطوراً من الناحية الفنية لمنع الاستخدام التخريبي لاستغلال بياناته ونتائجه لإلحاق الضرر بجهات أو أفراد أو مجموعات، ومن الضروري العمل بشكل مستمر على التنفيذ والتطوير لضمان موثوقية النظام.

المبدأ السادس - الشفافية والقابلية للتفسير

يؤسس مبدأ الشفافية والقابلية للتفسير لبناء الثقة في أنظمة وتقنيات الذكاء الاصطناعي، لذا يجب بناء أنظمة الذكاء الاصطناعي بدرجة عالية من الوضوح والقابلية للتفسير، مع وجود ميزات لتتبع مراحل اتخاذ القرارات المؤتمتة، ولا سيما تلك التي قد تؤدي إلى آثار ضارة على الأفراد، وهذا يعني أن البيانات والخوارزميات والقدرات والعمليات والغرض من نظام الذكاء الاصطناعي جميعها تحتاج إلى أن تكون شفافة وقابلة للتفسير للمتأثرين بها بشكل مباشر وغير مباشر، وتعتمد الدرجة التي يكون فيها النظام قابلاً للتتبع والتدقيق والشفافية والقابلية للتفسير على سياق نظام الذكاء الاصطناعي والغرض منه والنتائج التي قد تنتج، ويجب أن تكون أنظمة الذكاء الاصطناعي ومطوروها قادرين على تبرير أسس تصميمها وممارساتها وعملياتها وخوارزمياتها وقراراتها وسلوكياتها المسموح بها أخلاقياً وغير الضارة للعامة.

التخطيط والتصميم

١- عند تصميم نظام ذكاء اصطناعي شفاف وموثوق من المهم التحقق من أن أصحاب العلاقة المتوقع تأثرهم بالنظام على دراية تامة بكيفية خروج النظام بالنتائج، كما يجب منحهم إمكانية الوصول إلى الأساس المنطقي للقرارات التي تتخذها تقنية الذكاء الاصطناعي لشرحها بطريقة مفهومة وواضحة، ويجب أن تكون القرارات قابلة للتتبع بشكل واضح.

٢- ينبغي على مطور نظام الذكاء الاصطناعي تحديد مستوى الشفافية لمختلف أصحاب العلاقة، ويلزم تصميم نظام الذكاء الاصطناعي، حيث يتضمن قسماً للمعلومات يتيح إلقاء نظرة عامة على قرارات نموذج الذكاء الاصطناعي كجزء من تطبيق الشفافية الشاملة للتقنية، ويجب الالتزام بمشاركة المعلومات مع المستخدمين النهائيين وأصحاب العلاقة في نظام الذكاء الاصطناعي، وذلك بناءً على طبيعة نظام الذكاء الاصطناعي والسوق المستهدف، ويجب أن يحدد النموذج آلية عمل لتسجيل ومعالجة المشاكل والشكاوى التي تنشأ وكيفية حلها بطريقة شفافة ونظامية.

تهيئة البيانات

١- يتم توثيق مجموعات البيانات والعمليات التي توضح قرارات نظام الذكاء الاصطناعي وفقاً لأفضل المعايير للسماح بإمكانية التتبع وزيادة مستوى الشفافية.

٢- يجب تقييم مجموعات البيانات من حيث دقتها وملاءمتها وصحتها ومصدرها، نظراً إلى كون ذلك يحدث تأثيراً مباشراً على تدريب وبناء هذه الأنظمة، بما يضمن أن تكون هذه الأنظمة ممتثلة للمتطلبات التنظيمية في المملكة.

البناء وقياس الأداء

١- تتم مراعاة الشفافية في الذكاء الاصطناعي من منظورين، الأول هو العمليات المسبقة (ممارسات التصميم والبناء التي تؤدي إلى نتيجة مدعومة خوارزمياً) والثاني من حيث النتيجة (محتوى وتبرير النتيجة)، ويتم تطوير الخوارزميات بطريقة شفافة لضمان وضوح المدخلات وشرحها للمستخدمين النهائيين لنظام الذكاء الاصطناعي ليتمكنوا من تقديم الأدلة والمعلومات حول البيانات المستخدمة في معالجة القرارات التي تمت معالجتها.

٢- تضمن الخوارزميات التي تتسم بالشفافية والقابلية للتفسير أن أصحاب العلاقة المتأثرين بأنظمة الذكاء الاصطناعي سواء الأفراد أو المجموعات على اطلاع تام عندما تتم معالجة النتيجة من قبل نظام الذكاء الاصطناعي، وذلك من خلال إتاحة الفرصة لطلب معلومات توضيحية من مسؤول أو مطور نظام الذكاء الاصطناعي، ويتيح ذلك تحديد قرار الذكاء الاصطناعي وتحليله، الأمر الذي يسهل إمكانية مراجعته بالإضافة إلى إمكانية تفسيره.

٣- إذا تم بناء نظام الذكاء الاصطناعي من قبل طرف خارجي، فيجب على الجهات المسؤولة عن نظام الذكاء الاصطناعي التأكد من الاهتمام بتطبيق أخلاقيات الذكاء الاصطناعي وإمكانية الوصول إلى جميع الوثائق وتتبعها قبل الشراء أو الاعتماد.

التطبيق والمتابعة

١- عند تطبيق نظام الذكاء الاصطناعي، يجب توثيق مقاييس الأداء المتعلقة بمخرجات نظام الذكاء الاصطناعي والتحقق من دقتها وتوافقها مع الأولويات والأهداف، فضلاً عن قياس أثرها على الأفراد والفئات المستهدفة.

٢- ينبغي تسجيل توثيق معلومات عن أي أعطال في النظام أو خرق أو تسرب للبيانات أو غير ذلك، وإبلاغ الجهات المعنية وأصحاب العلاقة والجهات المختصة بها وفق ما تقرره الأنظمة ذات العلاقة، مع الحفاظ على شفافية أداء نظام الذكاء الاصطناعي، ويلزم إجراء اختبار دوري لواجهة وتجربة المستخدم لتجنب مخاطر التحيز أو صعوبة التعامل مع نظام الذكاء الاصطناعي أو أي مخاطر أخرى.

المبدأ السابع - المساءلة والمسؤولية

يُحمِّل مبدأ المساءلة والمسؤولية المصممين والمطورين ومسؤولي ومقيمي أنظمة الذكاء الاصطناعي المسؤولية الأخلاقية عن القرارات والإجراءات التي قد تؤدي إلى مخاطر محتملة وآثار سلبية على الأفراد والمجتمعات، ويجب تطبيق الإشراف البشري والحوكمة والإدارة المناسبة عبر دورة حياة نظام الذكاء الاصطناعي بأكملها لضمان وجود آليات مناسبة لتجنب الأضرار وإساءة استخدام هذه التقنية، وينبغي ألا تؤدي أنظمة الذكاء الاصطناعي إلى خداع الناس أو الإضرار بحرية اختيارهم دون مبرر، وأن يكون المصممون والمطورون والأشخاص الذين ينفذون نظام الذكاء الاصطناعي مذكورين ويمكن لأصحاب المصلحة التواصل معهم.

على الأطراف المسؤولين اتخاذ الإجراءات الوقائية اللازمة ووضع استراتيجية تقييم المخاطر والتخفيف منها للحد من الضرر الناجم عن نظام الذكاء الاصطناعي، ويجب على الأطراف المسؤولين عن نظام الذكاء الاصطناعي ضمان الحفاظ على عدالة النظام واستدامة هذه العدالة من خلال آليات الرقابة، وعلى جميع الأطراف المشاركة في دورة حياة نظام الذكاء الاصطناعي مراعاة هذه المبادئ عند اتخاذهم للقرارات.

التخطيط والتصميم

١- تعد هذه الخطوة بالغة الأهمية لتصميم أو شراء نظام ذكاء اصطناعي بطريقة مسؤولة وخاضعة للمساءلة، وينبغي إسناد المسؤولية الأخلاقية عن نتائج نظام الذكاء الاصطناعي إلى أصحاب العلاقة المسؤولين عن الإجراءات والأعمال الرئيسية في دورة حياة نظام الذكاء الاصطناعي، ومن الضروري وضع هيكل حوكمة ناضج يحدد مجالات التفويض والمسؤولية لدى الجهات المعنية الداخلية والخارجية بشكل واضح ومحدد، ويجب أن يراعي النهج المتبع في تصميم نظام الذكاء الاصطناعي حقوق الإنسان ومصالح الأفراد، بالإضافة إلى الأنظمة والقيم الاجتماعية والثقافية للمملكة.

٢- على الجهات وضع أدوات إضافية مثل تقييم الأثر، وأطر التخفيف من المخاطر، وآليات التدقيق والتقييم الشامل، والتصحيح، وخطط الحد من الكوارث.

٣- يتم بناء وتصميم نظام ذكاء اصطناعي تتم فيه مراقبة القرارات المتعلقة بعمليات ووظائف التقنية وتنفيذها، وتكون خاضعة للتدخل من قبل المستخدمين المصرح لهم، وتحدد الحوكمة والإشراف البشري الرقابة اللازمة ومستويات الاستقلالية من خلال وضع آليات محددة لذلك.

تهيئة البيانات

١- جودة البيانات من الجوانب المهمة في مبدأ المساءلة والمسؤولية لكونها تؤثر في نتائج نموذج الذكاء الاصطناعي والقرارات ذات الصلة، لذلك من المهم إجراء اختبارات جودة البيانات وفرز البيانات وضمان سلامتها للحصول على نتائج دقيقة للوصول إلى السلوك المقصود في النماذج الخاضعة للإشراف والنماذج غير الخاضعة للإشراف.

٢- يجب الموافقة على مجموعات البيانات واعتمادها قبل البدء في تطوير نموذج الذكاء الاصطناعي، بالإضافة إلى ذلك يجب تنقيح البيانات من التحيزات، وعدم إدراج السمات الحيوية في بيانات النموذج، وفي حال الحاجة إلى إدراج سمات حساسة، يجب توضيح الأساس المنطقي أو أهداف من قرار الإدراج بوضوح.

٣- من المهم توثيق عملية إعداد البيانات والتحقق من جودتها وصحتها من قبل المخولين بذلك، إذ يعد توثيق العملية ضرورياً للتدقيق والحد من المخاطر، ويجب الحصول على البيانات وتصنيفها ومعالجتها وإتاحتها بسهولة لتسهيل التدخل والسيطرة البشرية في مراحل لاحقة عند الحاجة.

البناء وقياس الأداء

١- يتكون تطوير نموذج نظام الذكاء الاصطناعي والخوارزمية من اختيار الخصائص وتهيئة مدخلات ضبط النموذج واختياره، ولتحقيق ذلك يجب أن يكون أصحاب العلاقة الفنيين الذين يقومون ببناء النماذج والتحقق منها مسؤولين عن هذه القرارات.

٢- إن تحديد المسؤوليات فيما يتعلق بالملكية والاعتمادات من شأنه أن يحدد آلية المساءلة التي تساعد في توجيه تطوير نظام الذكاء الاصطناعي من حيث الأسباب والتداخل والسماح بتدخل الاجتهادات البشرية.

٣- يجب دعم القرارات بمؤشرات كمية (مقاييس الأداء على مجموعات بيانات التدريب/ الاختبار، واتساق الأداء على المجموعات الحساسة المختلفة، ومقارنة الأداء لكل مجموعة مدخلات الضبط، وما إلى ذلك) ومؤشرات نوعية (القرارات اللازمة للتخفيف من المخاطر غير المقصودة الناتجة عن التنبؤات غير الدقيقة وتصحيحها).

٤- على الجهات المعنية والجهات المسؤولة عن تقنية الذكاء الاصطناعي مراجعة النموذج واعتماده بعد الاختبارات الناجحة وبعد جولات التحقق من قبول المستخدم قبل تطبيق نماذج الذكاء الاصطناعي.

التطبيق والمتابعة

١- يجب أن يتم تحديد المسؤولية والالتزامات المرتبطة بها في خطوة التطبيق والمتابعة بوضوح، ويجب مراقبة النتائج والقرارات المحددة في خطوة البناء والتحقق من صحتها بشكل مستمر، وينبغي أن تؤدي إلى إعداد تقارير أداء دورية، ويتم تحديد التنويه والتنبيه المناسب مسبقاً لهذه الخطوة بناء على البيانات ومقاييس الأداء.

٢- يمكن تحديد التنويه/ التنبيه كجزء من إجراءات تخفيف المخاطر أو التعافي من الكوارث وقد تحتاج إلى إشراف بشري.

الأدوار والمسؤوليات

يحدد إطار أخلاقيات الذكاء الاصطناعي الأدوار والمسؤوليات التالية على المستوى الوطني ومستوى الجهات.

على المستوى الوطني

الهيئة السعودية للبيانات والذكاء الأصطناعي

تعمل الهيئة على مراجعة وتحديث مبادئ أخلاقيات الذكاء الاصطناعي ومتابعة الالتزام بها، كما تقوم الهيئة بإعداد الأدلة والمعايير والتوجيهات الوطنية التي تضمن إدارة ونشر أخلاقيات الذكاء الاصطناعي بفاعلية على مستوى المملكة وتحقيق الهدف المنشود، وللهيئة في سبيل تنفيذ اختصاصاتها، القيام بالمهام التالية:

- **إعداد ومراجعة وتحديث أخلاقيات الذكاء الاصطناعي:** العمل على تحديث الأخلاقيات -بشكل دوري أو عندما تدعو الحاجة إلى ذلك- لمعالجة أي تغييرات أو تطورات مستقبلية.
- **وضع خطة تبني أخلاقيات الذكاء الاصطناعي:** إعداد الخطط الداعمة للتوعية بأخلاقيات الذكاء الاصطناعي وتقديم التوجيه المستمر للجهات المشمولة ضمن نطاق التطبيق لتسهيل تبني الأخلاقيات.
- **تقديم المشورة بشأن أخلاقيات الذكاء الاصطناعي:** دعم الجهات المشمولة ضمن نطاق التطبيق للالتزام بهذه المبادئ والإجابة عن أي استفسارات أو استشارات تتعلق بأخلاقيات الذكاء الاصطناعي.
- **قياس الالتزام بالأخلاقيات:** قياس التزام الجهات المطبقة بشكل منتظم وذلك من خلال التسجيل الاختياري للجهات المطبقة.
- **التنسيق ومتابعة التطبيق:** للهيئة التنسيق مع الجهات الحكومية ذوات العلاقة لتفعيل مبادئ أخلاقيات الذكاء الاصطناعي في القطاعات التي تشرف عليها تلك الجهات.

على مستوى الجهات المطبقة

تتحمل جميع الجهات المشمولة بنطاق التطبيق المسؤولية الأساسية عن ضمان نشر وثائق أخلاقيات الذكاء الاصطناعي الخاصة بها وفقاً لمبادئ أخلاقيات الذكاء الاصطناعي هذه وبالتالي يجب على الجهات تعيين أشخاص يتولون مسؤولية تنفيذ الأنشطة المتعلقة بأخلاقيات الذكاء الاصطناعي على النحو المنصوص عليه أدناه:

١- رئيس الجهة / مسؤول إدارة البيانات: مسؤول عن ممارسات أخلاقيات الذكاء الاصطناعي داخل الجهة، وتشمل المسؤوليات ما يلي:

- الموافقة على خطة أخلاقيات الذكاء الاصطناعي والإشراف عليها
- توزيع الأدوار المتعلقة بأخلاقيات الذكاء الاصطناعي
- اعتماد التقرير السنوي الخاص بأخلاقيات الذكاء الاصطناعي
- حل المشكلات التي يثيرها مسؤول الالتزام أو تفويضها أو تصعيدها إلى الهيئة إذا لزم الأمر
- التنسيق مع الهيئة والعمل كحلقة وصل بين الجهة والهيئة

٢- مسؤول الالتزام: هو القائد الاستراتيجي لممارسات أخلاقيات الذكاء الاصطناعي ويكون مسؤول الالتزام تحت إشراف مسؤول إدارة البيانات ويتبعه بشكل مباشر بالجهات العامة، وتشمل المسؤوليات ما يلي:

- تطوير استراتيجية أخلاقيات الذكاء الاصطناعي ومراجعة خطة الأخلاقيات وتقديم التحسينات المحتملة
- الإشراف على أخلاقيات الذكاء الاصطناعي ومراقبة وتحديد أولويات الأنشطة المتعلقة بالأخلاقيات وضمان صونها
- الالتزام بأخلاقيات الذكاء الاصطناعي وضمان تطبيق سياسات حوكمة البيانات والتزام الأطراف الخارجية بمبادئ حماية البيانات

٣- مسؤول الذكاء الاصطناعي: هو القائد التشغيلي المسؤول عن الذكاء الاصطناعي في الجهة ويتعاون بشكل وثيق مع فريق إدارة البيانات، وتشمل المسؤوليات ما يلي:

- التعاون مع الموظفين الآخرين العاملين في إدارة وحوكمة البيانات وحماية البيانات الشخصية
- وضع ومراجعة خطة أخلاقيات الذكاء الاصطناعي، مع تحديد الأهداف ومؤشرات الأداء بالتنسيق مع الجهات القيادية
- تحديد أولويات وإجراءات أخلاقيات الذكاء الاصطناعي، وتحديثها، وصيانتها ومراجعتها
- التوعية وتثقيف موظفي الجهة حول معايير وقيم أخلاقيات الذكاء الاصطناعي، والمشاركة في حملات التوعوية الوطنية بالتنسيق مع مسؤول الالتزام

٤- مقيم نظام الذكاء الاصطناعي: مسؤول عن تدقيق أنظمة الذكاء الاصطناعي لتحقيق أهداف معينة، وتشمل المسؤوليات ما يلي:

- مراجعة قنوات التواصل وأوجه التفاعل مع الأطراف المعنية للإفصاح عنها وكذلك قنوات تقديم الملاحظات الفعالة
- إجراء مراجعات دورية لتوثيق عمل إجراءات أخلاقيات الذكاء الاصطناعي
- المراجعة المستمرة لمؤشرات الأداء الرئيسية لأخلاقيات الذكاء الاصطناعي
- إصدار تقارير المراجعة والتدقيق بشأن تقييم أخلاقيات الذكاء الاصطناعي في الهيئة التي تشمل عملية تطوير الذكاء الاصطناعي وتنفيذه، وعمليات شراء منتجات الذكاء الاصطناعي الخاصة بالجهات الخارجية

التسجيل الاختياري

يهدف التسجيل الاختياري ومزاياه إلى تحفيز الجهات المستهدفة بتبنّي هذه الأخلاقيات عند بناء وتطوير الحلول القائمة على الذكاء الاصطناعي لضمان الاستخدام المسؤول لها.

الالتزام

- للهيئة متابعة وقياس مستوى التزام الجهات المسجلة بالأخلاقيات ودعمها في تقييم الالتزام بتطبيق أخلاقيات الذكاء الاصطناعي الخاصة بها وتقديم التقارير الاختيارية.
- يتم قياس مستوى الالتزام وفق الآتي:
  - عرض تقدم المنتج أو الجهة في الالتزام بالقائمة المرجعية لأخلاقيات الذكاء الاصطناعي المذكورة في مرفقات هذا المستند
  - نتائج التقييم الداخلي أو الخارجي لأخلاقيات الذكاء الاصطناعي
  - أهداف المنتج أو الجهة ومؤشرات قياس أداء أخلاقيات الذكاء الاصطناعي
  - مستوى الالتزام بأخلاقيات الذكاء الاصطناعي وتحقيق متطلباتها والوسوم التي حصل عليها المنتج أو الجهة

ويمكن للهيئة مساعدة الجهات في مراجعة التقارير ورفع التوصيات المتعلقة بالالتزام بأخلاقيات الذكاء الاصطناعي.

الوسوم التحفيزية

لتحفيز الجهات المطبقة للتسجيل والعمل بمبادئ أخلاقيات الذكاء الاصطناعي، للهيئة تقديم وسوم تحفيزية في أخلاقيات الذكاء الاصطناعي وتعكس هذه الوسوم مستوى نضج الممارسات المعمول بها على مستوى المنتجات والخدمات ومدى تبنيها لأخلاقيات الذكاء الاصطناعي وستصدر الهيئة دليلاً يوضح فئات الوسوم التحفيزية وآلية منحها.

الملاحق

الملحق أ: أدوات أخلاقيات الذكاء الاصطناعي

- **تقرير عدالة الذكاء الاصطناعي:** يسمح تقرير العدالة والإنصاف للجهة المسؤولة عن تقنية الذكاء الاصطناعي بتحديد معايير العدالة المستخدمة في نظام الذكاء الاصطناعي بشكل واضح وكذلك توضيح الأساس المنطقي والمبرر وراء ذلك بلغة مباشرة وسهلة الفهم وغير تقنية، ولتنفيذ نظام ذكاء اصطناعي يتسم بالنزاهة والإنصاف على نحو مستدام، فإن اختيار أهداف النزاهة المواتية يعد أمراً أساسياً لتحديد أسلوب نموذج الذكاء الاصطناعي من حيث معاييره الأخلاقية ومتطلباته التنظيمية، ويتم ذلك من خلال مشاركة الأسباب والقيم الأساسية للإنصاف الواردة في النموذج بالإضافة إلى عملية صنع القرار في نموذج الذكاء الاصطناعي للتواصل مع الجمهور الأوسع نطاقاً والوصول إليه، ويكون هذا التقرير متاحاً ويمكن الوصول إليه من قبل كل من الجمهور والأفراد والأوساط المتأثرة على حد سواء.

- **تقييم الأثر الأخلاقي:** سرّعت أنظمة الذكاء الاصطناعي الابتكار في سبل إجراء الأعمال وتنفيذها من قبل الممارسين، وبالتالي أصبح من الضروري تطوير تقييمات الأثر الأخلاقي لنظام الذكاء الاصطناعي لتحديد المجالات التي تحتاج إلى تعديل وإعادة المعايرة لتصميم نموذج الذكاء الاصطناعي في تقنية مقبولة أخلاقياً لتعظيم تأثيرها الإيجابي في تعزيز القدرات والمهارات البشرية، ويكمن الهدف من تقييمات الأثر في تقييم وتحليل مستوى التأثير الأخلاقي لتقنية الذكاء الاصطناعي على كل من الأفراد أو المجتمعات في السلوكيات المباشرة وغير المباشرة على حد سواء، مما يساهم بدوره في تمكين الجهة المسؤولة لنظام الذكاء الاصطناعي من القدرة على معالجة المشكلات المحددة وتعزيز المجالات التي تستلزم إدخال تحسينات وتعديلات، والقدرة على تقييم المخاطر الأخلاقية التي يتعرض لها نظام الذكاء الاصطناعي، وتحليل أثر الضرر التمييزي، والتمثيل الدقيق للأثر الأخلاقي للنظام من خلال تحليل متنوع للمتأثرين بالنظام، فضلاً عن العمل لتحديد ما إذا كان النموذج يجب أن ينتقل إلى الإنتاج أو النشر والتنفيذ الفعلي، ويتمثل أحد أهداف تقييم الأثر الأخلاقي في المساعدة على بناء ثقة الجمهور في نظام الذكاء الاصطناعي وإظهار الاهتمام والاجتهاد تجاه الجمهور العام.

معايير الخصوصية والأمن: تساعد معايير الخصوصية والأمن الشركات على تحسين استراتيجيتها لأمن المعلومات من خلال توفير التوجيهات وأفضل الممارسات بناءً على مجال الشركة ونوعية البيانات التي تحتفظ بها، يتضمن الجدول التالي بعض الأمثلة لمعايير الأمن والخصوصية:

اسم المعيار	الرابط
معايير آيزو (المنظمة الدولية للمعايير) للذكاء الاصطناعي مثل معيار المخاطر والمعايير التي تصدرها الهيئة السعودية للمواصفات و المقاييس	https://www.iso.org/standard/77304.html
معايير منظمة مهندسي الكهرباء والإلكترونيات	https://www.standards.ieee.org
إطار عمل الأمن السيبراني للمعهد الوطني للمعايير والتقنية	https://www.nist.gov/cyberframework/framework
إطار عمل إدارة مخاطر الذكاء الاصطناعي الصادر عن المعهد الوطني للمعايير والتقنية - قيد التنفيذ	https://www.nist.gov/itl/ai-risk-management-framework
ضوابط مركز ضوابط أمن الإنترنت	https://www.cisecurity.org/controls/
معيار أمان بيانات صناعة بطاقات الدفع	https://www.pcisecuritystandards.org/pci_security/
أهداف التحكم بالمعلومات والتقنيات ذوات الصلة	http://www.isaca.org/resources/cobit

- **موثوقية بنية الذكاء الاصطناعي:** ينبغي أن تنعكس المتطلبات الواردة في قسم مبادئ وأخلاقيات الذكاء الاصطناعي في تصميم بنية نظام الذكاء الاصطناعي، وينبغي أن تحدد بنية نظام الذكاء الاصطناعي القواعد والقيود عبر مراحل عمل نظام الذكاء الاصطناعي، إذ تُعد المبادئ والضوابط عامة وينبغي التعامل معها في حالات الاستخدام أو أنظمة الذكاء الاصطناعي المحددة من خلال نهج منطقي وواضح وسهل الشرح والفهم. نحتاج إلى الخطوات الثلاث التالية لمواءمة البنية مع أخلاقيات الذكاء الاصطناعي:

  - **الإحساس:** ينبغي تطوير نظام يتعرف على جميع العناصر البيئية اللازمة لضمان الالتزام بالمتطلبات
  - **الخطة:** ينبغي أن يراعي النظام الخطط التي تلتزم بالمتطلبات فقط
  - **التصرف:** ينبغي أن تقتصر إجراءات النظام على السلوكيات التي تحقق المتطلبات

يتم إعطاء الأولوية للأهداف التقنية المتمثلة في الدقة والموثوقية والسلامة والمتانة لضمان عمل نظام الذكاء الاصطناعي على نحو آمن، وينبغي على مطوري نظام الذكاء الاصطناعي بناء نظام يعمل بدقة وموثوقية -وفقاً لأغراض التصميم المخصصة- حتى في حال مواجهة أي تغيرات أو مخالفات أو اضطرابات غير متوقعة.

- **تقييم الخوارزميات:** الهدف من تقييم الخوارزمية هو ضمان اطلاع الأفراد أو المجتمعات على استخدام الخوارزميات والأوزان المختارة بها وطريقة إدارة استخدامها، وتختلف درجة جاهزية الجهات في مجال الذكاء الاصطناعي عن غيرها، ويُظهر هذا التقييم مجالات التحسين بالنسبة للجهات المعنية لتحسين الأوصاف المتعلقة بكيفية إبلاغ الخوارزميات أو التأثير على صنع القرار، خاصة في الحالات التي يكون فيها اتخاذ القرارات تلقائياً أو عندما تدعم الخوارزميات القرارات التي لها تأثير كبير على الأفراد أو المجموعات، ويوضح هذا التقييم أهمية ووزن الإشراف البشري على أنظمة الذكاء الاصطناعي المستخدمة.

تقييم العدالة: هو مجموعة من الطرق التشخيصية التي تساعدك على مقارنة أداء النماذج والعلامات العادلة لمجموعات محددة. إذ إنه يتحقق مما إذا كانت نتيجة النموذج مبالغاً فيها أو يستهان بها بشكل منتظم بالنسبة لمجموعة أو أكثر مقارنة بالمجموعات الأخرى، وبالإضافة إلى ذلك فإنه يقيِّم مدى تنوع البيانات الممثلة لكل مجموعة، ويتضمن الجدول التالي أمثلة على أدوات تقييم العدالة:

اسم الأداة	الرابط
Google Model Card Toolkit	https://github.com/tensorflow/model-card-toolkit
AI Fairness 360	https://github.com/Trusted-AI/AIF360
Microsoft Fairlearn	https://fairlearn.org/
Google What-if Tool	https://pair-code.github.io/what-if-tool/
Aequitas Bias and Fairness Audit Toolkit	http://aequitas.dssg.io/
Veritas Fairness Assessment Tool	https://github.com/mas-veritas2/veritastool
TensorFlow Fairness Indicators	https://www.tensorflow.org/tfx/guide/fairness_indicators
AI Explainability 360	https://github.com/Trusted-AI/AIX360

- **تقرير تفسير طريقة عمل الذكاء الاصطناعي:** كما هو موضح في إطار قسم الشفافية والقابلية للتفسير، ينبغي توضيح أسباب تصرف النظام على النحو الذي يقوم به وسبل اتخاذ القرارات، وعلى الرغم من أن بعض أساليب التدريب ذات أداء متفوق إلا أنه قد يظن البعض أنها تعمل بآلية غامضة كما لو كانت صندوقاً أسودَ مما يشكل بدوره تحدياً في كيفية تفسير النتائج، ويمكن أن تؤدي الانحرافات البسيطة في البيانات إلى حدوث انحرافات وتغييرات جذرية على المخرجات؛ لذلك يجب أن يوضح التقرير طريقة عمل الذكاء الاصطناعي ويفسر سلوك النظام مما يرفع موثوقية النظام ويزيد استخدام هذه التقنية، كما يساعد التقرير على تحسين فهم الآلية الأساسية لنظام الذكاء الاصطناعي بشكل أفضل إلى جانب تفسير المخرجات.

وينبغي على الأطراف المعنية بأنظمة الذكاء الاصطناعي مراعاة المفاضلة بين أساليب الأداء والقابلية للتفسير والتحقق، وفي بعض الحالات تزيد طرق التفسير من التعقيد أو تتطلب التضحية بنسبة من الأداء لتحقيق فهم وموثوقية أفضل، وينبغي إجراء تحليل للتكاليف والمزايا لتبرير مستوى القابلية للتفسير استناداً إلى هذا التحليل.

- **تدقيق الخوارزميات:** يمكن اكتشاف سلوكيات خوارزمية غير متوقعة من خلال عمليات تدقيق الخوارزميات، وبشكل عام يتم إجراء عمليات تدقيق الخوارزميات على نحو مخصص، كما أنه من الضروري معايرة عملية التدقيق الخوارزمية مع دعم خوارزميات الذكاء الاصطناعي. وينبغي أن تكون العملية منهجية ومستمرة. إن تنظيم وتدقيق أنظمة الذكاء الاصطناعي فيما يتعلق بالالتزام الأخلاقي أكثر تعقيداً من تنظيم ومراجعة عمليات صنع القرار أو العمليات البشرية، وينبغي تصميم أنظمة الذكاء الاصطناعي مع مراعاة مبادئ وضوابط أخلاقيات الذكاء الاصطناعي، كما ينبغي أن تتبع آليات التدقيق ذات المبادئ والضوابط بما يتماشى مع هذه المبادئ.

- **التقييم الذاتي للسلامة:** يجب مراعاة اعتبارات السلامة المتعلقة بالدقة والموثوقية والأمن والمتانة في كل خطوة من خطوات مراحل عمل نظام الذكاء الاصطناعي، وينبغي تسجيل التقييمات الذاتية لسلامة نظام الذكاء الاصطناعي وتوثيقها باستمرار بطريقة تسمح باستعراضها وإعادة تقييمها دورياً، وينبغي على الأطراف المعنية إجراء التقييمات الذاتية لأداء نظام الذكاء الاصطناعي في كل مرحلة من مراحل سير العمل، كما ينبغي عليهم تقييم مدى توافق ممارسات التصميم والتنفيذ مع أهداف السلامة المتمثلة في الدقة والموثوقية والأمن والمتانة.

- **طرق حماية البيانات:** تساعد على حماية البيانات بعد تصنيفها، وخاصة البيانات الحساسة، ويتم تقديم طرق حماية البيانات التالية على سبيل المثال:

  - **إلغاء تحديد البيانات** هو التخلص من البيانات المحددة للهوية الشخصية من أي مستند أو وسائط أخرى، بما في ذلك المعلومات الصحية المحمية للأفراد

  - **إخفاء هوية البيانات** وعدم الكشف عنها هو إزالة العناصر التي تمكّن من تحديد الهوية الشخصية من مجموعات البيانات للحفاظ على عدم الكشف عن هوية وسرية الأفراد الذين تصفهم البيانات، وغالباً ما تكون الطريقة المفضلة لجعل مجموعات البيانات الطبية المنظمة آمنة لتبادل المعلومات

  - **إخفاء البيانات** هو التخلص من المعلومات أو إخفائها، واستبدالها ببيانات بديلة واقعية أو حتى معلومات وهمية زائفة، والهدف من ذلك إنشاء نسخة لا يمكن فك شفرتها أو هندستها على نحو عكسي، وهناك مجموعة من الطرق لتغيير البيانات، بما في ذلك التشفير أو خلط الأحرف أو استبدال الكلمات أو الحروف

  - **إخفاء البيانات بهوية مستعارة** هي طريقة لإخفاء البيانات التي تضمن عدم إمكانية إسناد البيانات الشخصية إلى شخص معين، دون استخدام معلومات إضافية خاضعة للتدابير الأمنية، وهي جزء التي (GDPR) لا يتجزأ من اللائحة العامة لحماية البيانات في النظام الأوروبي العام لحماية البيانات تحتوي على عديد من الحيثيات والتي تحدد كيفية استخدام البيانات المستعارة والتوقيت المناسب لذلك

  - **تشفير البيانات** هو عبارة عن آلية لإخفاء البيانات وحجبها إذ تُستخدم هذه الآلية لحمايتها من الجرائم السيبرانية أو حتى من الحوادث العرضية غير المتوقعة، وقد تكون البيانات عبارة عن محتويات قاعدة بيانات أو رسالة بريد إلكتروني أو رسالة فورية أو ملف محفوظ على الحاسوب

  - **ترميز البيانات** هي عملية استبدال البيانات الشخصية برمز عشوائي، وغالباً ما يتم الاحتفاظ بالرابط بين المعلومات الأصلية والرمز المميز (مثل معالجة العمليات المالية في المواقع)، ويمكن أن تكون عبارة عن أرقام عشوائية تماماً أو يتم إنشاؤها بوظائف أحادية أو متعددة الاتجاهات (Tokens) الرموز (hashes)

  - **منع فقدان البيانات** يستخدم منع فقدان البيانات للكشف ومنع انتهاكات البيانات. ويتضمن ذلك مراقبة نشاط الشبكة وتحديد وحجب السلوك المشبوه وتنفيذ التشفير وضوابط الوصول

  - **حوكمة البيانات** تتضمن حوكمة البيانات جميع جوانب إدارة البيانات طوال دورة حياتها، بما في ذلك الأمن والاستخدامية والتوفر والخصوصية، ويتضمن ذلك تحديد سياسات وعمليات معالجة البيانات

  - **تقليل البيانات** هي عملية جمع البيانات الشخصية المطلوبة فقط دون المعلومات الإضافية لتخفيف المخاطر المرتبطة بانتهاكات البيانات وسوء استخدام المعلومات الشخصية

الملحق ب: ربط أدوات أخلاقيات الذكاء الاصطناعي بمراحل عمل نظام الذكاء الاصطناعي

الأداة	التخطيط والتصميم	إعداد بيانات المدخلات	البناء والتحقق	التطبيق والمتابعة
تقرير عدالة الذكاء الاصطناعي	●			
تقييم الأثر الأخلاقي	●			●
معايير الخصوصية والأمن	●	●	●	●
موثوقية بنية الذكاء الاصطناعي	●			
تقييم الخوارزميات			●	●
تقييم العدالة			●	
تقرير تفسير طريقة عمل الذكاء الاصطناعي			●	●
تدقيق الخوارزميات			●	●
التقييم الذاتي للسلامة	●	●	●	●
طرق حماية البيانات	●	●	●	●

الملحق ج: القائمة المرجعية لأخلاقيات الذكاء الاصطناعي

المرحلة الأولى لدورة عمل نظام الذكاء الاصطناعي: التخطيط والتصميم.

المرحلة	السؤال	ملزمة لأي طرف ثالث؟	المبادئ
التخطيط والتصميم ١	هل تم تصميم مستوى إشرافي مناسب لنظام الذكاء الاصطناعي وحالة الاستخدام؟	نعم	المساءلة والمسؤولية
التخطيط والتصميم ٢	هل يمنع تصميم نظام الذكاء الاصطناعي لديكم الثقة المفرطة في نظام الذكاء الاصطناعي أو الاعتماد المفرط عليه باستخدام آلية التدخل البشري اللازمة؟	نعم	المساءلة والمسؤولية
التخطيط والتصميم ٣	هل تم تحديد عمليات الإشراف البشري باستخدام مؤشرات الأداء الرئيسية المناسبة وتحديد المسؤولية للأطراف ذات الصلة؟	لا	المساءلة والمسؤولية
التخطيط والتصميم ٤	هل تم وضع استراتيجية التشغيل والحوكمة المواتية لإيقاف النظام أو التدخل في آلية عمله عندما لا يعمل على النحو المطلوب؟	لا	المساءلة والمسؤولية
التخطيط والتصميم ٥	هل تمت مراعاة متطلبات حماية المسؤولية والجهة صاحبة البيانات وأخذها بعين الاعتبار؟	نعم	المساءلة والمسؤولية
التخطيط والتصميم ٦	هل تم وضع الحدود القصوى لمؤشرات الأداء الرئيسية، وهل تم وضع إجراءات حوكمة أو إجراءات مستقلة لتنفيذ خطط بديلة أو احتياطية؟	لا	المساءلة والمسؤولية
التخطيط والتصميم ٧	هل تم توفير التدريب والتعليم اللازم للمساعدة في تطوير ممارسات المساءلة؟	لا	المساءلة والمسؤولية
التخطيط والتصميم ٨	هل تم التأكد من توافق هيكل حوكمة أخلاقيات الذكاء الاصطناعي مع آلية الحوكمة المقترحة في المبادئ الوطنية لأخلاقيات الذكاء الاصطناعي؟	لا	المساءلة والمسؤولية
التخطيط والتصميم ٩	هل تم التأكد من أن هيكل حوكمة أخلاقيات الذكاء الاصطناعي يتضمن آليات تدقيق داخلية أو خارجية؟	لا	المساءلة والمسؤولية
التخطيط والتصميم ١٠	هل تم وضع استراتيجية أو مجموعة من الإجراءات لتجنب وجود أو تعزيز التحيز غير العادل في نظام الذكاء الاصطناعي، بما يشمل كلاً من بيانات المدخلات وكذلك تصميم الخوارزمية؟	نعم	النزاهة والإنصاف
التخطيط والتصميم ١١	هل تم تحديد سمات البيانات الشخصية الحساسة المتعلقة بالأفراد أو المجموعات المحرومة بشكل منهجي أو تاريخي؟ وفي حال كانت الأمور على هذا النحو، فيجب تحديد الحد المسموح به الذي يجعل التقييم عادلاً أو غير عادل	لا	النزاهة والإنصاف
التخطيط والتصميم ١٢	هل تم تحديد مؤشرات الأداء الرئيسية لتقييم النزاهة؟	لا	النزاهة والإنصاف
التخطيط والتصميم ١٣	هل تم النظر في وضع آلية تشمل مشاركة مختلف الأطراف المعنية في تطوير واستخدام نظام الذكاء الاصطناعي؟	لا	النزاهة والإنصاف
التخطيط والتصميم ١٤	هل تم إجراء تحليل للأثر حول كيفية تأثير نظام الذكاء الاصطناعي على حقوق الإنسان الأساسية والقيم الثقافية؟ هل تم التنويه عن احتمالية وجود أي آثار سلبية على حقوق الإنسان الأساسية والقيم الثقافية والحلول أو آليات التعافي؟	نعم	القيم الإنسانية
التخطيط والتصميم ١٥	هل تم اتخاذ تدابير لضمان ألا يؤدي نظام الذكاء الاصطناعي إلى خداع الناس أو الإضرار بحرية اختيارهم دون مبرر؟	نعم	القيم الإنسانية
التخطيط والتصميم ١٦	هل تمت مواءمة نظام الذكاء الاصطناعي لديكم مع المعايير أو السياسات وقانون خصوصية أصحاب البيانات) أو EEE ذات الصلة (مثل معيار الآيزو ومعيار البروتوكولات المعتمدة على نطاق واسع لإدارة وحوكمة البيانات اليومية؟	نعم	الخصوصية والأمن

المرحلة	السؤال	ملزمة لأي طرف ثالث؟	المبادئ
التخطيط والتصميم ١٧	هل تم اتباع البروتوكولات والعمليات والإجراءات المعمول بها لإدارة وضمان الحوكمة المناسبة للبيانات؟ هل تم التأكد من اتباع معايير إدارة البيانات الوطنية وحماية البيانات الشخصية؟	نعم	الخصوصية والأمن
التخطيط والتصميم ١٨	هل تم التأكد من أن التحكم في الوصول إلى البيانات يلبي متطلبات الأمن والخصوصية والالتزام؟ هل تم تصميم آلية تسجيل لأغراض التدقيق والتصحيح؟	نعم	الخصوصية والأمن
التخطيط والتصميم ١٩	هل تم وضع استراتيجية إدارة المخاطر لنظام الذكاء الاصطناعي لديكم؟ هل تم إدراج مستويات المخاطر ومؤشرات الأداء الرئيسية وتقييم المخاطر وإجراءات التخفيف من حدتها؟	لا	الموثوقية والسلامة
التخطيط والتصميم ٢٠	هل تم تقييم مدى احتمالية أن يتسبب نظام الذكاء الاصطناعي في إلحاق الضرر أو الأذى بكل من المستخدمين أو الجهات الخارجية على حد سواء؟ هل تم تقييم الأضرار المحتملة والجمهور المتأثر ولا سيما درجة الخطورة المتوقعة؟	نعم	الموثوقية والسلامة
التخطيط والتصميم ٢١	هل تم تقييم مدى احتمالية أن يقوم نظام الذكاء الاصطناعي عن غير قصد بتحقيق نتائج خاطئة أو توقعات غير دقيقة أو فشل أو تغذية التحيزات المجتمعية؟	نعم	الموثوقية والسلامة
التخطيط والتصميم ٢٢	هل تم النظر في التأثير المحتمل أو مخاطر السلامة على البيئة أو الكائنات الحية أو المجتمع وكذلك أصحاب البيانات؟	نعم	المزايا الاجتماعية والبيئية
التخطيط والتصميم ٢٣	هل تم تقييم كم يتماشى نموذج عمل النظام مع رؤية الجهة ورسالتها وكذلك مدونة قواعد السلوك؟	نعم	الشفافية والقابلية للتفسير
التخطيط والتصميم ٢٤	هل تم تقييم تصميم نظام ذكاء اصطناعي قابل للتفسير حيث تكون البيانات والخوارزميات والمخرجات والقرارات شفافة وقابلة للتفسير لجميع الأطراف المعنية وذات الصلة؟	نعم	الشفافية والقابلية للتفسير
التخطيط والتصميم ٢٥	هل تم تصميم تجربة المستخدم مع مراعاة علم النفس البشري لتجنب خطر الارتباك أو التحيز التأكيدي أو التعب المعرفي؟	نعم	الشفافية والقابلية للتفسير
التخطيط والتصميم ٢٦	هل كان هناك افتراض بأن هناك مفاضلة تجارية؟	نعم	النزاهة والإنصاف
التخطيط والتصميم ٢٧	هل تم وضع آلية لقياس أو تقييم أثر الخصوصية؟	نعم	الخصوصية والأمن
التخطيط والتصميم ٢٨	هل تمت مراجعة منهجية إدارة البيانات بناءً على القيم الإنسانية ووفقاً للوائح التنظيمية للبيانات في المملكة؟	نعم	القيم الإنسانية

المرحلة الثانية لدورة عمل نظام الذكاء الاصطناعي: تهيئة البيانات.

المرحلة	السؤال	ملزمة لأي طرف ثالث؟	المبادئ
تهيئة البيانات ١	هل توجد آلية ثابتة تحدد المشكلات المتعلقة بخصوصية أصحاب البيانات أو حمايتها في عملية جمع البيانات ومعالجتها؟	نعم	الخصوصية والأمن
تهيئة البيانات ٢	هل تمت مراجعة البيانات من حيث النطاق والتصنيف؟	لا	الخصوصية والأمن
تهيئة البيانات ٣	هل تمت مراجعة البيانات للتحقق من مدى وضوح البيانات الشخصية ضمن مجموعة البيانات؟ هل توجد آلية ثابتة تسمح لنموذج الذكاء الاصطناعي بالتدريب دون استخدام البيانات الشخصية أو الحساسة أو حتى من خلال إتاحة أقل قدر ممكن منها؟	لا	الخصوصية والأمن
تهيئة البيانات ٤	هل توجد آلية محددة للتحكم في استخدام البيانات الشخصية (مثل الموافقة الصحيحة وإمكانية الإلغاء عند الاقتضاء)؟	نعم	الخصوصية والأمن
تهيئة البيانات ٥	هل هناك عمليات لضمان أمن أنظمة الذكاء الاصطناعي والحفاظ على أمن المعلومات وسريتها وخصوصيتها، وكذلك سلامة المعلومات المعالجة حتى في ظل ظروف معادية أو عدائية؟	نعم	الخصوصية والأمن
تهيئة البيانات ٦	هل تم تقييم جودة ومصدر البيانات التي تم الحصول عليها من خلال عمليات محددة؟	لا	الخصوصية والأمن
تهيئة البيانات ٧	هل كان هناك تقييم حول مدى إمكانية إجراء التحليل بعد التدريب واختبار البيانات؟	لا	الشفافية والقابلية للتفسير
تهيئة البيانات ٨	هل تمت دراسة أو مراجعة التنوع وشمول مجموعة البيانات الحالية؟	لا	النزاهة والإنصاف
تهيئة البيانات ٩	هل هناك آلية محددة لقياس ما إذا تم تقييم سلامة وجودة ودقة جمع البيانات ومصادرها وتحديثها؟	لا	المساءلة والمسؤولية
تهيئة البيانات ١٠	هل تم تطوير عملية تحليل مزايا الخصائص الحساسة؟	نعم	النزاهة والإنصاف
تهيئة البيانات ١١	هل قام الفريق بتقييم تصنيف البيانات ومعالجتها والوصول إليها لضمان الحصول عليها بشكل صحيح؟	نعم	القيم الإنسانية
تهيئة البيانات ١٢	هل تم التحقق من صحة نماذج البيانات والذكاء الاصطناعي لتشمل احترام حقوق الإنسان والقيم والتفضيلات الثقافية في المملكة العربية السعودية؟	نعم	القيم الإنسانية
تهيئة البيانات ١٣	هل تم تصنيف البيانات استناداً إلى توصيات الهيئة؟ وفي حال استخدام معايير أخرى غير الواردة، يرجى ذكرها.	نعم	المزايا الاجتماعية والبيئية
تهيئة البيانات ١٤	هل هناك إجراءات مناسبة لقياس جودة ودقة وملاءمة وموثوقية عينة البيانات عند التعامل مع مجموعات البيانات لنموذج الذكاء الاصطناعي؟	نعم	الموثوقية والسلامة

المرحلة الثالثة لدورة عمل نظام الذكاء الاصطناعي: البناء وقياس الأداء.

المرحلة	السؤال	ملزمة لأي طرف ثالث؟	المبادئ
البناء وقياس الأداء ١	هل تم اختبار سلوك النظام مقارنة بالمواقف والبيئات غير المتوقعة؟ هل هناك خطة بديلة محددة في حال تعرض نموذج الذكاء الاصطناعي لهجمات عدائية أو غيرها من المواقف غير المتوقعة؟ هل تم اختبار الخطط البديلة وتأكيدها؟	نعم	الموثوقية والسلامة
البناء وقياس الأداء ٢	هل هناك عمليات محددة تحدد التدابير اللازمة لوصف الإجراءات التي يتعين أخذها في حال فشل نظام الذكاء الاصطناعي في سياقات مختلفة؟ هل تم اختبار العمليات؟	نعم	الموثوقية والسلامة
البناء وقياس الأداء ٣	هل هناك عمليات محددة تحدد التدابير اللازمة لوصف فشل نظام الذكاء الاصطناعي في سياقات مختلفة؟ هل تم اختبار العمليات؟	نعم	الموثوقية والسلامة
البناء وقياس الأداء ٤	هل هناك آلية تواصل ثابتة لضمان موثوقية النظام من قبل المستخدمين النهائيين؟	نعم	الموثوقية والسلامة
البناء وقياس الأداء ٥	هل هناك تعريفات واضحة ومفهومة لشرح سبب اتخاذ نظام الذكاء الاصطناعي قراراً محدداً؟	لا	الشفافية والقابلية للتفسير
البناء وقياس الأداء ٦	هل تم بناء النموذج بطريقة بسيطة وقابلة للتفسير؟	لا	الشفافية والقابلية للتفسير
البناء وقياس الأداء ٧	هل تم اختبار قابلية تفسير نموذج الذكاء الاصطناعي بنجاح بعد تدريب النموذج؟	لا	الشفافية والقابلية للتفسير
البناء وقياس الأداء ٨	هل تم إجراء بحث يتعلق باستخدام الأدوات التقنية المتاحة لتحسين فهم البيانات والنموذج وأدائه؟	لا	النزاهة والإنصاف
البناء وقياس الأداء ٩	هل هناك عمليات قائمة وتحليل كمي لاختبار ومراقبة التحيزات المحتملة والإنصاف العام للنظام أثناء مراحل تطوير النظام؟ هل هناك آليات مطبقة لحماية أي أفراد أو مجموعات قد تتأثر بشكل غير متناسب بالآثار السلبية؟	لا	النزاهة والإنصاف
البناء وقياس الأداء ١٠	هل هناك أي آليات محددة لتقييم ما إذا كان نظام الذكاء الاصطناعي يشجع البشر على تطوير الارتباط والتعاطف مع النظام؟ هل هناك آليات تضمن محاكاة التفاعل الاجتماعي لأنظمة الذكاء الاصطناعي وعدم قدرتها على (الشعور)؟	لا	المزايا الاجتماعية والبيئية
البناء وقياس الأداء ١١	هل وافقت الأطراف المعنية على الاختبارات الناجحة وجولات التحقق من قبول المستخدمين قبل إعداد نماذج الذكاء الاصطناعي؟	نعم	المساءلة والمسؤولية
البناء وقياس الأداء ١٢	هل تم استخدام أي بيانات أو سمات حساسة في النموذج؟ وفي حال كان الأمر كذلك، فهل هناك مبرر لاستخدام سمات البيانات الشخصية الحساسة أو خصائصها؟	نعم	النزاهة والإنصاف
البناء وقياس الأداء ١٣	هل هناك منهجيات وخوارزميات للذكاء الاصطناعي تسمح وتسهل مواءمة عمليات صنع القرار مع حقوق الإنسان والقيم الثقافية للمملكة العربية السعودية؟	نعم	القيم الإنسانية

المرحلة الرابعة لدورة عمل نظام الذكاء الاصطناعي: التطبيق والمتابعة.

المرحلة	السؤال	ملزمة لأي طرف ثالث؟	المبادئ
التطبيق والمتابعة ١	في حال وجود روبوتات الدردشة أو أنظمة التواصل الأخرى، هل يدرك المستخدمون النهائيون أنهم يتفاعلون مع طرف آلي غير بشري؟	نعم	المساءلة والمسؤولية
التطبيق والمتابعة ٢	هل أجرى الفريق تقييماً لنقاط ضعف نظام الذكاء الاصطناعي لمواجهة الهجمات السيبرانية المحتملة أو الكشف عن البيانات الحساسة أو خرق السرية؟	نعم	الخصوصية والأمن
التطبيق والمتابعة ٣	هل هناك آليات لقياس ما إذا كان النظام ينتج كمية غير مقبولة من التوقعات غير الدقيقة؟	لا	المساءلة والمسؤولية
التطبيق والمتابعة ٤	هل توجد استراتيجية محددة لمتابعة وقياس ما إذا كان نظام الذكاء الاصطناعي يحقق الأهداف والأغراض والتطبيقات على نحو المطلوب؟	لا	الموثوقية والسلامة
التطبيق والمتابعة ٥	هل يتمتع الأشخاص المخولين بالوصول إلى البيانات بالكفاءات اللازمة لفهم تفاصيل متطلبات حماية البيانات؟	لا	الخصوصية والأمن
التطبيق والمتابعة ٦	هل هناك آليات مطبقة لتقييم مستوى تأثير نظام الذكاء الاصطناعي على عمليات صنع القرار للمستخدمين النهائيين؟	لا	الشفافية والقابلية للتفسير
التطبيق والمتابعة ٧	هل هناك عملية محددة وواضحة وقابلة للتفسير لإبلاغ المستخدمين النهائيين بالأسباب والمعايير والمزايا الكامنة وراء نتائج ومخرجات نظام الذكاء الاصطناعي؟ هل هناك خطوات واضحة للتواصل بشأن كيفية إثارة القضايا التي يمكن طرحها وما الجهات التي يمكن طرحها عليها؟	لا	الشفافية والقابلية للتفسير
التطبيق والمتابعة ٨	هل هناك عملية محددة لجمع ملاحظات المستخدمين النهائيين ودراستها واعتمادها على النظام؟	نعم	الشفافية والقابلية للتفسير
التطبيق والمتابعة ٩	هل هناك عمليات قائمة وتحليل كمي لمتابعة التحيزات والإنصاف العام للنظام أثناء مراحل التنفيذ؟	نعم	النزاهة والإنصاف
التطبيق والمتابعة ١٠	في حال وجود تباين، هل تم وضع آلية لقياس أو تقييم الأثر المحتمل لمثل هذا التباين على الحقوق الأساسية؟	لا	النزاهة والإنصاف
التطبيق والمتابعة ١١	هل هناك آليات محددة لضمان العدالة والإنصاف في أنظمة الذكاء الاصطناعي الخاصة بك؟	لا	النزاهة والإنصاف
التطبيق والمتابعة ١٢	هل يمكن للمستخدمين النهائيين للتقنيات المساعدة الوصول إلى المعلومات المتعلقة بنظام الذكاء الاصطناعي؟	لا	النزاهة والإنصاف
التطبيق والمتابعة ١٣	هل هناك آليات محددة لقياس الأثر الاجتماعي والبيئي لتنفيذ واستخدام نظام الذكاء الاصطناعي؟	نعم	المزايا الاجتماعية والبيئية
التطبيق والمتابعة ١٤	هل هناك آليات محددة لضمان تطبيق حقوق الإنسان الأساسية؟	نعم	المساءلة والمسؤولية
التطبيق والمتابعة ١٥	هل هناك عمليات محددة للجهات الخارجية أو العاملين للإبلاغ عن نقاط الضعف أو المخاطر أو التحيزات المحتملة في نظام الذكاء الاصطناعي؟	نعم	المساءلة والمسؤولية
التطبيق والمتابعة ١٦	هل هناك آليات محددة لإثبات مدى التزامك بالمبادئ المنصوص عليها في هذا المستند؟	لا	المساءلة والمسؤولية
التطبيق والمتابعة ١٧	هل هناك آليات محددة تسمح بالتعويض في حالة حدوث أي ضرر أو تأثير سلبي؟	نعم	المساءلة والمسؤولية

المرحلة	السؤال	ملزمة لأي طرف ثالث؟	المبادئ
التطبيق والمتابعة ١٨	هل هناك آليات محددة توفر معلومات للمستخدمين النهائيين أو الجهات الخارجية بخصوص فرص التعويض المتاحة؟	نعم	المساءلة والمسؤولية
التطبيق والمتابعة ١٩	هل هناك تقنيات مراقبة مستمرة لضمان الحفاظ على الخصوصية والأمن في نظام الذكاء الاصطناعي؟	نعم	الخصوصية والأمن
التطبيق والمتابعة ٢٠	هل هناك تقييمات دورية لأنظمة الذكاء الاصطناعي المستخدمة لضمان تطبيق حقوق الإنسان الأساسية والقيم الثقافية للمملكة؟	نعم	القيم الإنسانية

SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

SDAIA.GOV.SA SDAIA_SA SDAIA.SAUDI SDAIA-KSA