رؤيـــة VISION
2030
المملكة العربية السعودية
KINGDOM OF SAUDI ARABIA

مبادئ أخلاقيات الذكاء الاصطناعي

الإصدار 1
أغسطس 2022

المحتويات

مقدمة

نظراً إلى النمو المتسارع الذي تشهده الممارسات والتقنيات المتعلقة بالذكاء الاصطناعي، فقد تنوعت استخدامات الذكاء الاصطناعي لتشمل العديد 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 العديد من الجهات 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 من دور الذكاء الاصطناعي في تعزيز القدرات التنافسية لهذه الجهات.

انطلاقاً من التزام المملكة العربية السعودية بحقوق الإنسان وقيمها الثقافية، وتماشياً مع المعايير والتوصيات الدولية بشأن أخلاقيات الذكاء الاصطناعي، وبالإشارة إلى قرار مجلس الوزراء رقم (292) بتاريخ 27-04-1441هـ، القاضي في الفقرة (1) من المادة "العاشرة" بأن يتولى المكتب وضع السياسات وآليات الحوكمة والمعايير والضوابط الخاصة بالبيانات والذكاء الاصطناعي ومتابعة الالتزام بها بعد إقرارها، عليه فقد قام مكتب إدارة البيانات الوطنية بالاستفادة من الممارسات والمعايير العالمية عند وضع إطار أخلاقيات الذكاء الاصطناعي والذي يهدف إلى:

- دعم وتعزيز جهود المملكة في تحقيق رؤيتها واستراتيجياتها الوطنية المتعلقة باعتماد تقنيات الذكاء الاصطناعي وتشجيع البحث والابتكار وتعزيز النمو الاقتصادي لتحقيق الازدهار والتنمية المنشودة.
- وضع السياسات والمبادئ التوجيهية واللوائح والأطر المتعلقة بأخلاقيات الذكاء الاصطناعي.
- تمكين الجهات التنظيمية من وضع سياساتها الخاصة أو قواعد سلوكها وتنفيذ خططها واستخدام تقنية الذكاء الاصطناعي للتنبؤ بالمستقبل واتخاذ القرارات بما يتماشى مع الرؤية والاستراتيجيات الوطنية.
- حوكمة نماذج البيانات والذكاء الاصطناعي للحد من الآثار السلبية لأنظمة الذكاء الاصطناعي (اقتصادياً ونفسياً واجتماعياً وما إلى ذلك) والتهديدات المحتملة (الأمنية والسياسية وغيرها).
- مساعدة الجهات في اعتماد المعايير والأخلاقيات عند بناء وتطوير الحلول القائمة على الذكاء الاصطناعي لضمان الاستخدام المسؤول لها.
- حماية خصوصية أصحاب البيانات وحقوقهم فيما يتعلق بمعالجة بياناتهم الشخصية.
- التنسيق والعمل مع المنظمات الإقليمية والدولية لمواءمة واعتماد أفضل الممارسات المتبعة لتطبيق وحوكمة الذكاء الاصطناعي.

التعريفات

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

المصطلحات المستخدمة	التعريف
الجهات المُطبقة	أي جهة عامة أو شركة أو فرد يتعين عليه الالتزام بهذه السياسة.
الذكاء الاصطناعي	الذكاء الاصطناعي هو مجموعة من التقنيات التي تمكن آلة أو نظامًا من التعلم والفهم والتصرف و الإحساس.
نظام او نموذج الذكاء الاصطناعي	مجموعة من النماذج التنبؤية والخوارزميات المتقدمة التي يمكن استخدامها لتحليل البيانات والتنبؤ بالمستقبل أو تسهيل عملية صنع القرار للأحداث المستقبلية المتوقعة.
مقيّم نظام الذكاء الاصطناعي	أي شخص ذو صفة طبيعية أو اعتبارية يقوم بتدقيق أنظمة الذكاء الاصطناعي لتحقيق أهداف معينة.
مطور نظام الذكاء الاصطناعي	أي شخص ذو صفة طبيعية أو اعتبارية يقوم بتطوير أنظمة الذكاء الاصطناعي من خلال بناء نماذج تنبؤية باستخدام البيانات والخوارزميات لتحقيق أهداف معينة.
دورة نظام الذكاء الاصطناعي	العملية الدورية التي يتوقع من مشاريع الذكاء الاصطناعي اتباعها لتكون قادرة على تصميم وبناء وإنتاج نظام قوي وآمن يقدم قيمة عملية ورؤى من خلال الالتزام بطريقة موحدة ومنظمة لإدارة تنفيذ وتسليم نموذج الذكاء الاصطناعي.
مسؤول نظام الذكاء الاصطناعي	أي شخص ذو صفة طبيعية أو اعتبارية يطبق أنظمة الذكاء الاصطناعي أو يستخدمها لتحقيق أهداف معينة.

المصطلحات المستخدمة	التعريف
المستخدم المفوض	شخص مسموح له و لديه إمكانية الوصول إلى نظام معلومات لأداء أو المساعدة في أداء دور أو مسؤولية محددة مسبقًا على وظائف ومكونات النظام.
المدير التنفيذي للامتثال/ مسؤول الالتزام	يتبع المدير التنفيذي للامتثال أو مسؤول الالتزام رئيس الجهة أو المدير التنفيذي للبيانات ويكون مسؤولًا أمامه عن وضع المعايير الأخلاقية ومعايير الالتزام التي يجب أن تتبعها الجهات وتحافظ عليها عند توسيع العمليات، والتوظيف، والاستحواذ على المزيد من الأصول، وما إلى ذلك.
المدير التنفيذي للبيانات	يتولى المدير التنفيذي للبيانات مسؤولية تطوير وإدارة البيانات وتنفيذ الحوكمة بشأنها والإشراف على تنفيذ ممارسات إدارة البيانات في الجهات العامة.
مدونة قواعد الأخلاقيات والسلوكيات	مدونة قواعد الأخلاقيات والسلوكيات هي دليل المبادئ المصممة لمساعدة المهنيين على ممارسة الأعمال التجارية بكل نزاهة وصدق مع الحفاظ على حقوق الإنسان وحرياته.
البيانات	مجموعة من الحقائق في صورتها الأولية أو في صورة غير منظّمة مثل الأرقام أو الحروف أو الصور أو مقاطع الفيديو أو التسجيلات الصوتية أو الرموز.
حوكمة البيانات	عملية إدارة توفّر البيانات وإمكانية استخدامها وسلامتها ونزاهتها وأمنها في المؤسسات والأنظمة، وفق معايير وسياسات تتحكم في استخدام تلك البيانات.
عينة البيانات	البيانات المستخدمة في بناء النماذج التنبؤية وخوارزميات الذكاء الاصطناعي وتدريبها واختبارها للوصول إلى نتائج محددة.
صاحب البيانات	أي شخص تتعلق به البيانات الشخصية، أو من ينوب عنه، أو الشخص الذي له الولاية الشرعية عليه.
المستخدم النهائي	أي شخص ذو صفة طبيعية أو اعتبارية يستهلك أو يستخدم السلع أو الخدمات التي تنتجها أنظمة الذكاء الاصطناعي.
مسؤول الأخلاقيات	يتولى مسؤول الأخلاقيات مسؤولية الالتزام للأخلاقيات في الجهة العامة.
مؤشر الأداء الرئيسي	مؤشر الأداء الرئيسي
المملكة	المملكة العربية السعودية
الهيئة التنظيمية الوطنية	أي جهة حكومية أو عامة مستقلة تتولى مهام ومسؤوليات تنظيمية أو رقابية لقطاع معين في المملكة العربية السعودية بناءً على مستند نظامي.
مكتب إدارة البيانات الوطنية	مكتب إدارة البيانات الوطنية
البيانات الشخصية	كافة البيانات، بغض النظر عن مصدرها أو شكلها، التي قد تؤدي إلى الكشف عن هوية الشخص تحديدًا، أو تسهّل تحديد هوية الشخص بشكل مباشر أو غير مباشر، بما في ذلك الاسم أو رقم الهوية الشخصية أو العنوان أو أرقام الاتصال أو أرقام التراخيص أو الممتلكات الشخصية أو أرقام الحسابات البنكية أو بطاقات الائتمان أو الصور ومقاطع الفيديو الشخصية أو البيانات الشخصية الأخرى.
سدايا	الهيئة السعودية للبيانات والذكاء الاصطناعي
البيانات الحساسة	كل بيان شخصي يتضمن الإشارة إلى أصل الفرد ومن في حكمه العرقي أو القبلي، أو معتقده الديني أو الفكري أو السياسي، أو يدل على عضويته في جمعيات أو مؤسسات أهلية، وكذلك البيانات الجنائية والأمنية، أو بيانات السمات الحيوية التي تحدد الهوية، أو البيانات الوراثية، أو البيانات الائتمانية، أو البيانات الصحية، وبيانات تحديد الموقع، والبيانات التي تدل على أن الفرد مجهول الأبوين أو أحدهما.
الأطراف الخارجية	أي شخصية طبيعية أو اعتبارية، أو جهة عامة، أو وكالة أو هيئة بخلاف المشاركين الرئيسيين لدى المستخدم النهائي لنظام الذكاء الاصطناعي، ومالك نظام الذكاء الاصطناعي، ومطور نظام الذكاء الاصطناعي ومقيم نظام الذكاء الاصطناعي.

نطاق التطبيق

تطبق المبادئ المنصوص عليها في أخلاقيات الذكاء الاصطناعي على جميع الجهات (العامة والخاصة وغير الربحية) المعنية بتطوير أو تبني الحلول المعتمدة على تقنيات الذكاء الاصطناعي ويستثنى من نطاق التطبيق، تطوير أو تبني الحلول المعتمدة على تقنيات الذكاء الاصطناعي لأغراض المحافظة على الصحة والسلامة العامة، وكذلك المحافظة على المصالح الحيوية للأفراد، والمصالح العليا للمملكة.

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 بموجبها يتم تحديد كل خطوة يتوقع من الجهة اتباعها للاستفادة من هذه التقنية لتحقيق قيمة عملية، وهي طريقة موحدة لتمثيل المهام استناداً إلى أفضل الممارسات في تنفيذ وإدارة نماذج الذكاء الاصطناعي، مما يجعلها أنسب الخيارات لتضمين أخلاقيات الذكاء الاصطناعي.

تنقسم دورة حياة نظام الذكاء الاصطناعي إلى أربع مراحل رئيسية، لها نفس المستوى من الأهمية، وتتضمن كل مرحلة من المراحل عدد من الأنشطة الرئيسية، على النحو التالي:

<u>التخطيط والتصميم:</u>

- تحديد المشكلة.
- دعم المشكلة من خلال نهج قائم على البيانات.
- اختيار تقنية الذكاء الاصطناعي بما يتناسب مع الحلول المقترحة
- دراسة جدوى البدائل المحتملة.
- تطوير مؤشرات الأداء المناسبة.

<u>تهيئة البيانات:</u>

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

<u>البناء وقياس الأداء:</u>

- تنفيذ طريقة العمل.
- تدريب واختبار النموذج.
- ضبط المتغيرات أو مدخلات النموذج.
- التحقق من أداء النموذج.

<u>التطبيق والمتابعة:</u>

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

تتأثر ممارسات أخلاقيات الذكاء الاصطناعي ببعضها البعض عبر دورة حياة نظام الذكاء الاصطناعي، لذلك من المهم التأكد من تضمين المبادئ الأخلاقية والضوابط المتعلقة بها لكل مرحلة من مراحل دورة حياة نظام الذكاء الاصطناعي المذكورة، ونتيجة لذلك تم وضع مبادئ أخلاقيات الذكاء الاصطناعي، المدعومة بالضوابط وتم تصنيفها أيضاً وفق مراحل دورة حياة نظام الذكاء الاصطناعي لعرضها من خلال نهج منظم وشامل.

مبادئ وضوابط أخلاقيات الذكاء الاصطناعي

تم تحديد مبادئ أخلاقيات الذكاء الاصطناعي في المملكة بما يتماشى مع المعايير العالمية والقيم الثقافية للمملكة. كما أن المبادئ الموضحة أدناه مدعومة بضوابط توجيهية عبر دورة حياة نظام الذكاء الاصطناعي. استندت الضوابط الواردة في هذه المبادئ، والتي بدورها يجب أن توجه أي تطور مستقبلي لبرنامج أعمال أخلاقيات الذكاء الاصطناعي في المملكة.

المبدأ الأول - النزاهة والإنصاف

يتطلب مبدأ النزاهة والإنصاف عند تصميم أو جمع أو تطوير أو نشر أو استخدام أنظمة الذكاء الاصطناعي، اتخاذ الإجراءات اللازمة للقضاء على التحيز أو التمييز أو الوصم الذي يتعرض له الأفراد أو الجماعات أو الفئات، وقد يحدث التحيز بسبب البيانات أو التمثيل أو الخوارزميات ويمكن أن يؤدي إلى تمييز فئة ضد أخرى.

عند تصميم واختيار وتطوير أنظمة الذكاء الاصطناعي، من الضروري ضمان معايير عادلة ومنصفة وغير متحيزة وموضوعية وشاملة ومتنوعة وممثلة لجميع شرائح المجتمع أو الشرائح المستهدفة منها، ويجب ألا تقتصر وظيفة نظام الذكاء الاصطناعي على مجموعة محددة على أساس الجنس أو العرق أو الدين أو العمر أو غير ذلك. بالإضافة إلى ذلك، يجب أن تكون المخاطر المحتملة، والفوائد العامة، والغرض من استخدام البيانات الشخصية مبررة ومحددة بشكل واضح ودقيق من قبل الجهة المسؤولة عن نظام الذكاء الاصطناعي.

لضمان تطابق أنظمة الذكاء الاصطناعي القائمة على الإنصاف والشمولية، يجب تدريب أنظمة الذكاء الاصطناعي على البيانات التي يتم تنظيفها من التحيز، كما يجب أن تمثل مجموعات الأقليات المتأثرة، وبناء وتطوير الخوارزميات بطريقة تجعل تكوينها خالياً من التحيز والمغالطات.

<u>التخطيط والتصميم:</u>

في المراحل الأولى من تحديد الغرض من نظام الذكاء الاصطناعي، يتعاون فريق التصميم لتحديد الأهداف وكيفية تحقيقها بطريقة فعالة ومحسنة، ويعد تخطيط وتصميم نظام الذكاء الاصطناعي مرحلة أساسية لترجمة الأهداف والنتائج المرجوة من النظام، ومن المهم خلال هذه المرحلة الخروج بتصميم يتسم بالنزاهة والإنصاف ويأخذ الاحتياطات المناسبة عبر نظام الذكاء الاصطناعي وعملياته وآلياته لمنع التحيزات وللحيلولة دون أن يكون لها تأثير تمييزي أو تؤدي إلى نتائج غير مرغوب فيها.

يبدأ التصميم المراعي للنزاهة والإنصاف من بداية دورة حياة نظام الذكاء الاصطناعي من خلال جهود تعاونية بين الأعضاء الفنيين وغير الفنيين وذلك لتحديد الأضرار والفوائد المحتملة والأفراد المتضررين والفئات غير الممثلة في النظام وتقييم مدى تأثرهم بالنتائج وما إذا كان التأثير مبرراً في ظل الهدف العام من نظام الذكاء الاصطناعي.

يعد تقييم نزاهة نظام الذكاء الاصطناعي أمراً بالغ الأهمية، لذا يجب اختيار المقاييس في هذه المرحلة من دورة حياة نظام الذكاء الاصطناعي، واختيار المقاييس بناء على نوع النظام (قائمة على القاعدة، التصنيف، الانحدار، إلخ)، تأثير القرار (عقابي، انتقائي، إلخ)، والضرر أو الفائدة التي ستعود على العينات المتوقعة بشكل صحيح أو غير صحيح.

يتم في هذه المرحلة تحديد وتعريف سمات البيانات الشخصية المتعلقة بالأشخاص أو الفئات وبصورة منهجية، ويتم تحديد الحد الذي يكون عنده التقييم عادلاً أو غير عادل، كما يجب تحديد مقاييس تقييم النزاهة التي سيتم تطبيقها على البيانات الحساسة خلال الخطوات المستقبلية.

<u>تهيئة البيانات:</u>

يُعد اتباع أفضل الممارسات في الحصول على البيانات والتعامل معها وتصنيفها وإدارتها من الأولويات لضمان توافق النتائج مع الأهداف والغايات المحددة لنظام الذكاء الاصطناعي، وتتحقق فعالية وسلامة جودة البيانات من خلال ضمان سلامة مصدر البيانات ودقتها في تمثيل جميع الملاحظات لتجنب أي حرمان منظّم للفئات غير الممثلة تمثيلاً كافياً أو الأقل حظاً، ويجب أن تكون كمية ونوعية مجموعات البيانات كافية ودقيقة لخدمة الغرض من النظام. يؤثر حجم عينة البيانات التي تم جمعها أو الحصول عليها تأثيراً كبيراً على دقة وعدالة مخرجات النموذج تحت التطوير.

ويجب اعتماد دقة البيانات عند إدارة البيانات وتصنيفها (التوسيم والتعليق والتنظيم) لتجنب إدخال التقديرات البشرية ولمنع التحيز وضمان نزاهة البيانات.

كما يجب ألا تدرج خصائص البيانات الشخصية المحددة في مرحلة التخطيط والتصميم في بيانات النموذج، كي لا تزيد التحيز الموجود ضدها. لذا يجب تحليل خصائص البيانات الحساسة وعدم إدراجها في بيانات المدخلات، ولكن في بعض الحالات قد لا يكون ذلك ممكناً بسبب دقة أو هدف نظام الذكاء الاصطناعي، وفي هذه الحالة يجب تقديم مبررات لاستخدام خصائص البيانات الشخصية وتحديد بديل عنها.

<u>البناء وقياس الأداء:</u>

في مرحلة البناء وقياس الأداء من دورة حياة نظام الذكاء الاصطناعي، من الضروري مراعاة النزاهة في التنفيذ باعتبارها عاملاً مهماً عند بناء نظام الذكاء الاصطناعي واختباره وتنفيذه، ويتطلب بناء النموذج واختيار الميزات من المهندسين والمصممين أن يكونوا على دراية بأن الخيارات المتخذة بشأن تجميع أو فصل، أو استبعاد الميزات، بالإضافة إلى أن الأحكام العامة المتخذة بشأن موثوقية وأمن المجموعة الإجمالية من الميزات، قد يكون لها عواقب وخيمة على الفئات الضعيفة أو غير الممثلة في البيانات. لذا، عند اختيار النموذج يجب النظر في تقييم مقاييس النزاهة والإنصاف، إذ يجب أن تكون مقاييس النزاهة والإنصاف في النموذج ضمن الحد المحدد للخصائص الحساسة، كما يجب تحديد نهج التقييم الخاص بالنزاهة والإنصاف ومقاييس الأداء بوضوح خلال هذه المرحلة، ويجب أن يكون تقييم النزاهة مبرراً إذا لم يجتاز النموذج الرائد التقييم.

من الضروري التأكد من اختيار الخصائص السببية، كما يجب التحقق من الخصائص المختارة مع ممثلي بيانات الأعمال والفرق غير الفنية.

تتضمن التقنيات المؤتمتة لدعم القرار العديد من المخاطر الكبيرة المتمثلة في التحيز والتطبيق غير المرغوب فيه في مرحلة التشغيل الفعلي، لذلك من المهم وضع آليات لمنع النتائج الضارة والتمييزية في هذه المرحلة.

التطبيق والمتابعة:

يجب وضع آليات وبروتوكولات واضحة عند التطبيق الفعلي لنظام الذكاء الاصطناعي وذلك لقياس نزاهة النتائج وأدائها وكيفية تأثيرها على مختلف الأفراد والجماعات. عند تحليل نتائج النموذج التنبؤي، يجب تقييم ما إذا كانت المجموعات الممثلة في عينة البيانات تتلقى مزايا بشكل متساوٍ أو مماثل، أو إذا كان نظام الذكاء الاصطناعي يضر بفئة محددة على نحو غير متناسب دون مراعاة الفروق الديموغرافية وذلك لضمان تحقيق العدالة في النتائج.

كما يجب مراقبة مقاييس النزاهة والإنصاف المحددة مسبقاً، وإذا كان هناك أي انحراف عن الحد المسموح بها، فيجب التحقق فيما إذا كانت هناك حاجة لتجديد النموذج.

يجب تحديد حجم الضرر العام والمنفعة المتحققة من النظام وتوزيعه على فئات معينة من المستخدمين

المبدأ الثاني - الخصوصية والأمن

يمثل مبدأ الخصوصية والأمن القيم والمبادئ الشاملة التي يُطلب بموجبها من أنظمة الذكاء الاصطناعي، طوال دورتها أن تكون مبنية بطريقة آمنة وتراعي خصوصية أصحاب البيانات الشخصية التي يتم جمعها بحيث تكون على أعلى مستويات الأمن في جميع العمليات والإجراءات المتعلقة بالبيانات وسريتها، الأمر الذي يفضي بدوره إلى منع اختراق البيانات والنظام بما قد يؤدي إلى الإضرار بالسمعة أو الأضرار النفسية أو المالية أو المهنية أو غيرها، ويجب تصميم أنظمة الذكاء الاصطناعي باستخدام آليات وضوابط توفر إمكانية إدارة ومراقبة نتائجها والتقدم المحرز طوال دورتها لضمان امتثالها دائماً بقواعد وبرتوكولات الخصوصية والأمن.

التخطيط والتصميم:

يتم إعداد وتصميم نظام الذكاء الاصطناعي والخوارزمية المرتبطة به بطريقة من خلالها يمكن مراعاة حماية خصوصية الأفراد، وعدم إساءة استخدام البيانات الشخصية واستغلالها، وضمان عدم استناد معايير اتخاذ القرارات في التقنية الآلية إلى خصائص أو معلومات تحدد الهوية الشخصية، ويقتصر استخدام البيانات الشخصية على ما هو ضروري لتشغيل النظام بشكل سليم. يتم تصميم أنظمة الذكاء الاصطناعي التي تؤدي إلى تحديد سمات الأفراد أو الجماعات فقط في حالة الموافقة على ذلك من قبل مسؤول الالتزام والأخلاقيات أو إذا تم ذلك وفقاً لمدونة قواعد السلوك المهني التي تطورها الجهة التنظيمية لقطاع معين. تتم مواءمة مخطط الأمن والحماية لنظام الذكاء الاصطناعي، والبيانات التي تتم معالجتها والخوارزمية التي يتم استخدامها، مع أفضل الممارسات حتى تكون هذه الأنظمة قادرة على تحمل الهجمات السيبرانية ومحاولات اختراق البيانات.

يجب اتباع الأطر والمعايير القانونية للخصوصية والأمن وتهيئتها بما يتناسب مع حالة الاستخدام أو الجهة المعنية. من الجوانب المهمة في الخصوصية والأمن هي بنية البيانات، وبالتالي يجب التخطيط لتصنيف البيانات وتحديد خصائصها من أجل تحديد مستويات الحماية واستخدام البيانات الشخصية. يجب التخطيط لآليات أمنية لإلغاء التعريف بالبيانات الحساسة أو الشخصية في النظام، كما يجب اعتماد إجراءات القراءة/ الكتابة/ التحديث للمجموعات ذات الصلة.

تهيئة البيانات:

عند الحصول أو إدارة أو تنظيم البيانات يجب الالتزام بالأطر والمعايير القانونية لخصوصية البيانات؛ لحماية خصوصية الأفراد وأمن البيانات المعلومات من مجموعة واسعة من التهديدات.

تضمن سرية البيانات اقتصار الوصول إلى المعلومات على الأشخاص المصرح لهم بالوصول إلى المعلومات ووجود ضوابط محددة لإدارة تفويض صلاحيات الوصول إلى المعلومات والبيانات. يجب أن يتمتع مصمم ومهندس نظام الذكاء الاصطناعي بالمستويات المناسبة من النزاهة لحماية دقة واكتمال المعلومات وطرق المعالجة وذلك لضمان اتباع الأطر والمعايير القانونية للخصوصية والأمن، كما يجب التأكد من حماية إتاحة وتخزين البيانات من خلال توفير أنظمة قواعد بيانات آمنة. يتم تصنيف

جميع البيانات المعالجة لضمان حصولها على المستوى المناسب من الحماية وفقاً لحساسيتها أو تصنيفها الأمني، ويجب أن يكون مطورو نظام الذكاء الاصطناعي ومالكوه على دراية بتصنيف أو حساسية المعلومات التي يتعاملون معها والمتطلبات المرتبطة بها للحفاظ على أمنها، وتُصنف جميع البيانات من حيث متطلبات الأعمال وأهميتها وحساسيتها لمنع الإفصاح غير المصرح به عنها أو تعديلها، وتُصنف البيانات بطريقة سياقية لا تؤدي إلى استخلاص المعلومات الشخصية، بالإضافة إلى ذلك يجب استخدام آليات إلغاء التحديد بناءً على تصنيف البيانات ووفقاً للمتطلبات المتعلقة بأنظمة وقوانين حماية البيانات.

يتم اتخاذ إجراءات النسخ الاحتياطي للبيانات وأرشفتها في هذه المرحلة للتوافق مع سياسات استمرارية الأعمال والتعافي من الكوارث وتخفيف المخاطر.

البناء وقياس الأداء:

يُطبق مبدأ الخصوصية والأمن خلال عملية تصميم وبناء نظام الذكاء الاصطناعي، وتتضمن آليات الأمن حماية الأبعاد التصميمية المختلفة لنموذج الذكاء الاصطناعي من الهجمات التخريبية، وتتم حماية هيكل ووحدات نظام الذكاء الاصطناعي من التلف أو التعديل غير المصرح به لأي من مكوناته، ويجب تأمين نظام الذكاء الاصطناعي للحفاظ على سلامة المعلومات التي يعالجها، ويجب كذلك أن يكون نظام الذكاء الاصطناعي آمناً بحيث يظل فعالاً وجاهزاً للاستخدام من قبل المستخدمين المصرح لهم ومحافظاً على أمن المعلومات السرية والخاصة حتى في الظروف العدائية أو التخريبية، بالإضافة إلى ذلك يجب وضع ضوابط حماية مناسبة لضمان تقيد أنظمة اتخاذ القرار بالذكاء الاصطناعي بمتطلبات خصوصية وأمن البيانات ذات الصلة، وينبغي اختبار نظام الذكاء الاصطناعي للتأكد من أن البيانات المتاحة لا تكشف عن البيانات الحساسة أو تنتهك قواعد إخفاء الهوية.

التطبيق والمتابعة:

بعد تشغيل نظام الذكاء الاصطناعي، وبعد تحقيق النتائج المرجوة، يجب أن تكون هناك متابعة مستمرة لضمان الحفاظ على الخصوصية في نظام الذكاء الاصطناعي وضمان سلامته وأمنه، وتتم إعادة النظر في تقييم أثر الخصوصية وتقييم إدارة المخاطر باستمرار لضمان التقييم المنتظم للاعتبارات الاجتماعية والأخلاقية. يجب أن يكون مسؤولو نظام الذكاء الاصطناعي مسؤولين عن تصميم وتنفيذ أنظمة الذكاء الاصطناعي بما يضمن حماية المعلومات الشخصية طوال دورة نظام الذكاء الاصطناعي، ويتم تحديث مكونات نظام الذكاء الاصطناعي بناءً على تقارير المتابعة المستمرة، كما يجب تقييم أثر الخصوصية.

المبدأ الثالث - الإنسانية

يسلط مبدأ الإنسانية الضوء على ضرورة بناء أنظمة الذكاء الاصطناعي باستخدام منهجية عادلة مسموح بها أخلاقياً تستند إلى حقوق الإنسان والقيم الثقافية الأساسية وذلك لإحداث أثر مفيد على الأطراف المعنية والمجتمعات المحلية والمساهمة في تحقيق الأهداف والغايات طويلة وقصيرة الأجل من أجل صالح البشرية، ومن الضروري أن يتم تصميم النماذج التنبؤية بحيث لا تخدع، أو تتلاعب، أو تضع سلوكاً لا يقصد به تمكين، أو تعزيز، أو زيادة المهارات البشرية، بل ينبغي لها أن تتبنى نهجاً تصميمياً أكثر تركيزاً على الإنسان يتيح له الاختيار واتخاذ القرار.

التخطيط والتصميم:

من الضروري تصميم وبناء نموذج قائم على حقوق الإنسان الأساسية والقيم والمبادئ الثقافية وتطبيقه على قرارات وعمليات ووظائف نظام الذكاء الاصطناعي. يتعين على مصممي نموذج الذكاء الاصطناعي تحديد الكيفية التي سيتوافق بها نظام الذكاء الاصطناعي مع حقوق الإنسان الأساسية والقيم الثقافية للمملكة العربية السعودية، فضلاً عن تحديد التقنيات اللازمة واختبارها، مع تحديد الآلية التي سيسعى من خلالها نظام الذكاء الاصطناعي ونتائجه إلى تعزيز المهارات والقدرات البشرية.

تهيئة البيانات:

لضمان تجسيد نماذج الذكاء الاصطناعي لهيكل وتصميم يركزان على الإنسان، يجب الالتزام بممارسات إدارة البيانات المسؤولة والأخلاقية التي يجب اتباعها وفقاً لأفضل الممارسات وكذلك المعايير والضوابط الخاصة بإدارة البيانات في المملكة، ويتم الحصول على البيانات وتصنيفها ومعالجتها وإتاحتها بشكل صحيح لضمان احترام حقوق الإنسان والقيم الثقافية للمملكة العربية السعودية.

البناء وقياس الأداء:

عند إنشاء أنظمة الذكاء الاصطناعي، يجب على المصممين والمهندسين إعطاء الأولوية لبناء أنظمة وخوارزميات الذكاء الاصطناعي التي تسمح وتسهل عملية صنع القرار والتي تراعي التوافق مع حقوق الإنسان والقيم الثقافية للمملكة. يجب ألا تعمل القرارات المؤتمتة الناتجة عن أنظمة الذكاء الاصطناعي بطريقة جزئية ومستقلة دون مراعاة حقوق الإنسان والقيم الثقافية الأوسع نطاقاً في نتائجها النهائية، ولتحقيق ذلك يجب على المصممين تمكين أنظمة الذكاء الاصطناعي باستخدام المعايير المناسبة وتدريب الخوارزميات لتحقيق النتائج التي تنهض بالإنسانية.

التطبيق والمتابعة:

يتم إجراء تقييمات دورية لنظام الذكاء الاصطناعي المستخدم لضمان مواءمة نتائجه مع حقوق الإنسان والقيم الثقافية، ولضمان دقة مؤشرات الأداء الرئيسية، ولرصد تأثيره على الأفراد أو الجماعات وذلك لضمان التحسين المستمر للتقنية.

ينبغي 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 نظام الذكاء الاصطناعي. في حال العثور على أي نتائج سلبية وضارة، يجب على مسؤول نظام الذكاء الاصطناعي تحديد المجالات التي تحتاج إلى معالجة وتطبيق تدابير تصحيحية لتحسين أداء نظام الذكاء الاصطناعي ونتائجه بشكل متكرر.

المبدأ الرابع - المنافع الاجتماعية والبيئية

يعزز مبدأ المنافع الاجتماعية والبيئية 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 الاجتماعي والبيئي مع معالجة التحديات الاجتماعية والبيئية المرتبطة بها، وهذا يستلزم حماية المنفعة الاجتماعية والاستدامة البيئية.

التخطيط والتصميم:

تؤثر أنظمة الذكاء الاصطناعي تأثيراً كبيراً على المجتمعات والمنظومات المتواجدة بها، وبالتالي يجب أن يكون لدى مسؤولي أنظمة الذكاء الاصطناعي شعور عالٍ بالوعي بأن هذه التقنيات قد يكون لها آثار ضارّة أو تحولية على المجتمع والبيئة، كما يجب التعامل مع تصميم أنظمة الذكاء الاصطناعي بطريقة أخلاقية وحساسة بما يتماشى مع قيم منع الضرر لكل من البشر والبيئة. عند تخطيط وتصميم أنظمة الذكاء الاصطناعي، يجب الاهتمام بمنع المشاكل الاجتماعية والبيئية والمساعدة في معالجتها بطريقة تكفل المسؤولية الاجتماعية والبيئية المستدامة.

تهيئة البيانات:

يتم اتباع العمليات والسياسات التي تحكم إدارة البيانات عند إعداد تصنيف وهيكلة البيانات التي ستغذي نظام الذكاء الاصطناعي، وينبغي أن تكون البيانات المتعلقة بالمواضيع الاجتماعية والبيئية متاحة للهياكل الأساسية للبيانات العامة ويجب أن تبين بوضوح المنفعة الاجتماعية للبيانات المعروضة.

البناء وقياس الأداء:

تكون للنماذج والخوارزميات هدف نهائي ونتيجة اجتماعية أو بيئية، مع القدرة على إظهار ارتباط النتائج المتوقعة بذلك الغرض الاجتماعي أو البيئي من خلال فوائد تحويلية ومؤثرة، على سبيل المثال تحقيق مستويات مقبولة من استهلاك الموارد واستهلاك الطاقة والمحافظة عليها كما يمكن تحديد الأسلوب الذي ستسعى من خلاله أنظمة الذكاء الاصطناعي على معالجة المخاوف العالمية المتعلقة بالقضايا الاجتماعية والبيئية، وممارسة مسؤوليات مستدامة وبيئية.

التطبيق والمتابعة:

بعد تشغيل نظام الذكاء الاصطناعي، يجب على الجهة المسؤولة عن نظام الذكاء الاصطناعي أن تضمن إجراء تقييم مستمر للأثر البشري والاجتماعي والثقافي والاقتصادي والبيئي لتقنيات الذكاء الاصطناعي. مع الإدراك الكامل لآثار نظام الذكاء الاصطناعي على الاستدامة كهدف يجب متابعته و تطويره باستمرار عبر مجموعة من الأهداف ذات الأولوية التي تم وضعها في مرحلة التخطيط والتصميم. مع الحرص على تعزيز وتشجيع قدرة حلول الذكاء الاصطناعي في معالجة المجالات ذات الاهتمام العالمي التي تتماشى مع أهداف التنمية المستدامة.

المبدأ الخامس - الموثوقية والسلامة

يضمن مبدأ الموثوقية والسلامة التزام نظام الذكاء الاصطناعي بالمواصفات المحددة وأن نظام الذكاء الاصطناعي يعمل بشكل كامل وفق الآلية التي كان يقصدها ويتوقعها مصممو. تمثل الموثوقية مقياساً للثبات وتبعث الثقة بمدى قوة النظام كما تمثل مقياساً للاعتمادية التي يتوافق بها النظام من الناحية التشغيلية مع وظائفه المرجوة والنتائج التي يحققها، من ناحية أخرى تمثل السلامة مقياساً للكيفية التي لا يشكل بها نظام الذكاء الاصطناعي خطراً على المجتمع والأفراد. على سبيل التوضيح، 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 تعطل النظام. يجب أن يكون نظام العمل الموثوق آمناً من خلال عدم تعريض المجتمع للخطر ويجب أن تكون لديه آليات مدمجة لمنع الضرر.

لذا يرتبط إطار الحد من المخاطر ارتباطاً وثيقاً بهذا المبدأ، وينبغي تقليل المخاطر المحتملة والأضرار غير المقصودة إلى أدنى حد.

وتتم مراقبة النموذج التنبؤي ومراقبته 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 بالجهات أو الأفراد أو الجماعات، ومن الضروري اتباع نهج مستمر للتنفيذ والتطوير لضمان الموثوقية.

التخطيط والتصميم:

هناك حاجة كبيرة لتصميم وتطوير نظام ذكاء اصطناعي يمكنه تحمل عدم دقّة وعدم الاستقرار والتقلب التي قد يواجهها، ويعد 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 غير معروفة وغير متوقعة، كما أنه من الضروري وضع مجموعة من المعايير والبروتوكولات التي تقيم موثوقية نظام الذكاء الاصطناعي لضمان سلامة خوارزمية النظام ومخرجات البيانات. من الضروري الحفاظ على النفقات الفنية المستدامة ونتائج النظام للحفاظ على ثقة الجمهور بنظام الذكاء الاصطناعي. تعد معايير التوثيق ضرورية لتتبع تطور النظام وتوقع المخاطر المحتملة ومعالجة الثغرات. يجب أن تخضع جميع نقاط القرار المهمة في تصميم النظام لموافقة الجهات المعنية للحد من المخاطر وتحميل الجهات المعنية مسؤولية القرارات.

تهيئة البيانات:

يتم اتخاذ الخطوات والإجراءات المناسبة لقياس جودة ودقة وملاءمة وموثوقية عينة البيانات عند التعامل مع مجموعات البيانات الخاصة بنموذج الذكاء الاصطناعي، ويعد ذلك ضرورياً لضمان دقة تفسير البيانات من قبل نظام الذكاء الاصطناعي واتساقها، وتجنب القياسات المضللة، فضلاً عن ضمان صلة نتائج نظام الذكاء الاصطناعي بالغرض من النموذج.

ومن الضروري وضع خطوة للتحقق من كيفية عمل النظام في ظل الأحداث الطارئة والسيناريوهات غير الطبيعية، وفي هذه الخطوة يجب إعداد بيانات اختبارات التحمل في السيناريوهات غير الطبيعية.

البناء وقياس الأداء:

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

لضمان القوة الفنية لنظام الذكاء الاصطناعي، يجب اختباره والتحقق منه وإعادة تقييمه بشكل دقيق، بالإضافة إلى دمج آليات الإشراف والضوابط المناسبة في تطويره، وتلزم الموافقة على اختبار تكامل النظام من قبل الجهات المعنية ذات الصلة للحد من المخاطر والمسؤولية.

ويجب أن تعود أنظمة الذكاء الاصطناعي الآلية التي تتضمن سيناريوهات يفهم فيها أن القرارات لها تأثير لا رجعة فيه أو قد تنطوي على قرارات تتعلق بالحياة والموت، إلى العنصر البشري لاتخاذ هذه القرارات، علاوة على ذلك لا ينبغي استخدام أنظمة الذكاء الاصطناعي لأغراض التقييم الاجتماعي أو المراقبة الجماعية.

التطبيق والمتابعة:

تتم مراقبة قوة 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).

المبدأ السادس - الشفافية والقابلية للتفسير

يعد مبدأ الشفافية والقابلية للتفسير عاملاً مهماً لبناء الثقة في أنظمة وتقنيات الذكاء الاصطناعي والحفاظ عليها، 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 تجاه أصحاب البيانات، وهذا يعني أن البيانات والخوارزميات والقدرات والعمليات والغرض من نظام الذكاء الاصطناعي تحتاج إلى أن تكون شفافة ومعممة وقابلة للتفسير للمتأثرين 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 عن هذه التقنية، ويجب أن تكون أنظمة الذكاء الاصطناعي ومصمموها قادرين على تبرير أسس تصميمها وممارساتها وعملياتها وخوارزمياتها وقراراتها أو سلوكياتها المسموح بها أخلاقياً وغير ضار للعامة.

التخطيط والتصميم:

عند تصميم نظام ذكاء اصطناعي شفاف وموثوق، من المهم التأكد من أن الجهات المعنية المتأثرة بأنظمة الذكاء الاصطناعي يجب أن تكون على دراية تامة بكيفية معالجة النتائج وتقديم التقرير بشأنها، كما يجب منحهم إمكانية الوصول إلى الأساس المنطقي للقرارات التي تتخذها تقنية الذكاء الاصطناعي لشرحه بطريقة مفهومة وسياقية، ويجب أن تكون القرارات قابلة

للتتبع بشكل واضح. ينبغي على الجهات المسؤولة عن أنظمة الذكاء الاصطناعي تحديد مستوى الشفافية لمختلف الجهات المعنية بالتقنية استناداً إلى خصوصية البيانات وتصاريح الجهات المعنية، ويلزم تصميم نظام الذكاء الاصطناعي بحيث يتضمن قسماً للمعلومات في المنصة يتيح إلقاء نظرة عامة على قرارات نموذج الذكاء الاصطناعي كجزء من تطبيق الشفافية الشاملة للتقنية، ويجب الالتزام بمشاركة المعلومات كمبدأ فرعي مع المستخدمين النهائيين والجهات المعنية في نظام الذكاء الاصطناعي عند الطلب أو فتحها للجمهور، وذلك اعتماداً على طبيعة نظام الذكاء الاصطناعي والسوق المستهدف، ويجب أن يحدد النموذج آلية عمل لتسجيل ومعالجة المشاكل والشكاوى التي تنشأ لتتمكن من حلها بطريقة شفافة وقابلة للتفسير.

تهيئة البيانات:

يتم توثيق مجموعات البيانات والعمليات التي تسفر عن قرار نظام الذكاء الاصطناعي وفقاً لأفضل المعايير الممكنة للسماح بإمكانية التتبع وزيادة مستوى الشفافية، ويجب تقييم مجموعات البيانات من حيث دقتها وملاءمتها وصحتها ومصدرها، وهذا له تأثير مباشر على تدريب وبناء هذه الأنظمة نظراً لأن طريقة تنظيم البيانات، والهيكلة يجب أن تكون شفافة وقابلة للتفسير عند الاستحواذ على البيانات وجمعها وأن تكون في امتثال تام لأنظمة خصوصية البيانات ومعايير وضوابط الملكية الفكرية.

البناء وقياس الأداء:

يتم التفكير في الشفافية في الذكاء الاصطناعي من منظورين، الأول هو العملية الكامنة وراءها (ممارسات التصميم البناء التي تؤدي إلى نتيجة مدعومة خوارزمياً) والثاني من حيث منتجها (محتوى وتبرير النتيجة)، ويتم تطوير الخوارزميات بطريقة شفافة لضمان وضوح شفافية المدخلات وشرحها للمستخدمين النهائيين لنظام الذكاء الاصطناعي ليتمكنوا من تقديم الأدلة والمعلومات حول البيانات المستخدمة في معالجة القرارات التي تمت معالجتها. تضمن الخوارزميات التي تتسم بالشفافية والقابلية للتفسير أن الجهات المعنية المتأثرة بأنظمة الذكاء الاصطناعي، سواء الأفراد أو المجتمعات، على اطلاع تام عندما تتم معالجة النتيجة من قبل نظام الذكاء الاصطناعي من خلال إتاحة الفرصة لطلب معلومات توضيحية من مسؤول نظام الذكاء الاصطناعي. ويتيح ذلك تحديد قرار الذكاء الاصطناعي وتحليله، الأمر الذي يسهل إمكانية مراجعته بالإضافة إلى إمكانية تفسيره، وإذا 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:

عند تطبيق نظام الذكاء الاصطناعي، يجب توثيق مقاييس الأداء المتعلقة بمخرجات نظام الذكاء الاصطناعي ودقتها وتوافقها مع الأولويات والأهداف، فضلاً عن قياس أثرها على الأفراد والمجتمعات، وإتاحتها للجهات المعنية بتقنية الذكاء الاصطناعي. ينبغي تسجيل معلومات عن أي أعطال في النظام أو خرق للبيانات أو غير ذلك، وإبلاغ الجهات المعنية بها، مع الحفاظ على شفافية أداء نظام الذكاء الاصطناعي، ويلزم إجراء اختبار دوري لواجهة المستخدم وتجربة المستخدم لتجنب خطر التحيز أو صعوبة التعامل مع نظام الذكاء الاصطناعي.

المبدأ السابع - المساءلة والمسؤولية

يُحمّل مبدأ المساءلة والمسؤولية المصممين والموردين والقائمين على المشتريات والمطورين ومسؤولي ومقيمي أنظمة الذكاء الاصطناعي والتقنية نفسها المسؤولية الأخلاقية والمسؤولية 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 الضرر وإساءة استخدام هذه التقنية، وينبغي ألا تؤدي أنظمة الذكاء الاصطناعي إلى خداع الناس أو الإضرار بحرية اختيارهم دون مبرر، ويكون المصممون والمطورون والأشخاص الذين ينفذون نظام الذكاء الاصطناعي قابلين للتعرف عليهم وأن يتحملوا المسؤولية عن أي أضرار محتملة للتقنية على الأفراد أو المجتمعات، حتى لو كان التأثير السلبي غير مقصود. على الأطراف المسؤولة اتخاذ الإجراءات الوقائية اللازمة بالإضافة إلى وضع استراتيجية تقييم المخاطر والتخفيف منها للحد من الضرر الناجم عن نظام الذكاء الاصطناعي، ويرتبط مبدأ المساءلة والمسؤولية ارتباطاً وثيقاً بمبدأ العدالة، ويجب على الأطراف المسؤولة عن نظام الذكاء الاصطناعي ضمان الحفاظ على عدالة النظام واستدامتها من خلال آليات الرقابة، وعلى جميع الأطراف المشاركة في دورة حياة نظام الذكاء الاصطناعي مراعاة هذه القيم عند اتخاذهم للقرارات.

التخطيط والتصميم:

تعد هذه الخطوة بالغة الأهمية لتصميم أو شراء نظام ذكاء اصطناعي بطريقة مسؤولة وخاضعة للمساءلة، وينبغي إسناد المسؤولية والمسؤولية الأخلاقية عن نتائج نظام الذكاء الاصطناعي إلى الجهات المعنية المسؤولة عن إجراءات معينة في دورة حياة نظام الذكاء الاصطناعي، ومن الضروري وضع هيكل حوكمة قوي يحدد مجالات التفويض والمسؤولية لدى الجهات المعنية الداخلية والخارجية دون ترك أي ثغرات من عدم اليقين تحول دون تحقيق هذا المبدأ، ويجب أن يراعي النهج المتبع في تصميم نظام الذكاء الاصطناعي حقوق الإنسان والحريات الأساسية، بالإضافة إلى الأنظمة والقوانين الوطنية والقيم الثقافية للمملكة. من المهم أيضاً للجهات وضع أدوات إضافية مثل تقييمات الأثر، وأطر التخفيف من المخاطر، وآليات التدقيق والتقييم الشامل، والتصحيح، وخطط التعافي من الكوارث، ومن الضروري 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.

تهيئة البيانات:

جودة البيانات من الجوانب المهمة في مبدأ المساءلة والمسؤولية لأنها تؤثر على نتائج نموذج الذكاء الاصطناعي والقرارات ذات الصلة، لذلك من المهم إجراء اختبارات جودة البيانات وفرز البيانات وضمان سلامة البيانات للحصول على نتائج دقيقة للوصول إلى السلوك المقصود في النماذج الخاضعة للإشراف والنماذج غير الخاضعة للإشراف، وتلزم الموافقة على مجموعات البيانات واعتمادها قبل البدء في تطوير نموذج الذكاء الاصطناعي، بالإضافة إلى ذلك يجب تنظيف البيانات من التحيزات، كما يجب عدم إدراج السمات الحساسة في بيانات النموذج كما ذكر في مبدأ العدالة، وفي حال الحاجة إلى إدراج سمات حساسة، يجب توضيح الأساس المنطقي أو أهداف من قرار الإدراج بوضوح، ويتم توثيق عملية إعداد البيانات والتحقق من جودتها والتحقق من صحتها من قبل الأطراف المسؤولة، 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:

يتكون تطوير نموذج نظام الذكاء الاصطناعي والخوارزمية من اختيار الخصائص وتهيئة مدخلات ضبط النموذج واختياره. ولتحقيق ذلك، يجب أن تكون الجهات المعنية الفنية التي تقوم ببناء النماذج والتحقق منها مسؤولة عن هذه القرارات. إن تحديد المسؤوليات المناسبة فيما يتعلق بالملكية والتواصل من شأنه أن يحدد وتيرة المساءلة التي من شأنها أن تساعد في توجيه تطوير نظام الذكاء الاصطناعي من حيث الأسباب والتداخل القوي والسماح بتدخل الاجتهادات الإنسانية، ويجب 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) ونوعية (القرارات اللازمة للتخفيف من المخاطر غير المقصودة الناتجة عن التنبؤات غير الدقيقة وتصحيحها)، وعلى الجهات المعنية والجهات المسؤولة عن تقنية الذكاء الاصطناعي مراجعة النموذج واعتماده بعد الاختبارات الناجحة وبعد جولات التحقق من قبول المستخدم قبل إمكانية إنتاج نماذج الذكاء الاصطناعي.

التطبيق والمتابعة:

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 المحفزات والتنبيهات المحددة مسبقاً لهذه الخطوة على البيانات ومقاييس الأداء. يعد تحديد هذه المحفزات عملية صارمة ويجب إسناد كل محفز للجهة المعنية المناسبة. يمكن تحديد هذه المحفزات/ التنبيهات 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.

أولا: على المستوى الوطني

مكتب إدارة البيانات الوطنية

يعمل مكتب إدارة البيانات الوطنية بصفته الجهة التنظيمية للبيانات والذكاء الاصطناعي بمراجعة وتحديث مبادئ أخلاقيات الذكاء الاصطناعي ومتابعة الالتزام بها، كما يقوم المكتب بإعداد الأدلة والمعايير والتوجيهات الوطنية التي تضمن إدارة ونشر أخلاقيات الذكاء الاصطناعي بفعالية على مستوى المملكة وتحقيق الهدف المنشود.

وللمكتب في تنفيذ اختصاصاته، القيام بالمهام التالية:

- **إعداد ومراجعة وتحديث أخلاقيات الذكاء الاصطناعي**: إصدارها وتحديثها بانتظام لمعالجة التغييرات المحتملة التي تؤثر على أخلاقيات الذكاء الاصطناعي واللوائح المرتبطة بها والمجتمعات والبيئة.
- **وضع خطة اعتماد أخلاقيات الذكاء الاصطناعي**: إعداد المحتوى الداعم وتقديم التوجيه المستمر للجهات المطبقة لتسهيل اعتماد اخلاقيات الذكاء الاصطناعي.

- **تقديم المشورة بشأن أخلاقيات الذكاء الاصطناعي**: دعم الجهات المشمولة بنطاق التطبيق للالتزام بهذه المبادئ والإجابة عن أي استفسارات تتعلق بأخلاقيات الذكاء الاصطناعي.
- **قياس الالتزام لأخلاقيات الذكاء الاصطناعي**: قياس امتثال الجهات المطبقة على أساس منتظم بناءً على آلية الالتزام المحددة بشكل مباشر أو من خلال الجهات التنظيمية القطاعية (لمزيد من التفاصيل، يرجى الاطلاع على قسم "الالتزام") والتحقق من أنشطة أخلاقيات الذكاء الاصطناعي عند الحاجة.
- **التوعية بشأن أخلاقيات الذكاء الاصطناعي**: تنفيذ ومتابعة مبادرات التواصل والتدريب لتعزيز الوعي بأخلاقيات الذكاء الاصطناعي واعتمادها على المستوى الوطني.
- **تقييم أداء أخلاقيات الذكاء الاصطناعي**: تحليل مؤشرات الأداء الرئيسية لأخلاقيات الذكاء الاصطناعي وأثرها على المستوى الوطني وتجميع فرص التحسين لإبلاغ الجهات المعنية بها.
- **البوابة الوطنية للامتثال**: تصميم وبناء وصيانة بوابة الالتزام لضمان قدرة الجهات المشمولة بنطاق التطبيق على نشر تقاريرها وإدارتها وتحديثها، وإدراج المواد الإرشادية والداعمة لأخلاقيات الذكاء الاصطناعي. كما يجب استخدام البوابة للإبلاغ عن أي مخالفات أو إخلال في أنظمة الذكاء الاصطناعي المستخدمة بما قد يؤثر سلباً على المجتمعات أو البيئة.

الجهة التنظيمية

يجوز أن يفوض المكتب مهام الإشراف على مبادئ أخلاقيات الذكاء الاصطناعي وإنفاذها لقطاعات معينة إلى الجهات التنظيمية المشرفة على القطاعات الرئيسية. في حال تم تفويض الجهة التنظيمية بهذا الدور، تشمل مسؤوليات الجهة التنظيمية ما يلي:

- العمل كبديل لمكتب إدارة البيانات الوطنية فيما يتعلق بالمسؤوليات: وضع خطة اعتماد أخلاقيات الذكاء الاصطناعي، وتقديم المشورة بشأن أخلاقيات الذكاء الاصطناعي، والتوعية وقياس مدى الالتزام بأخلاقيات الذكاء الاصطناعي، وأداء أخلاقيات الذكاء الاصطناعي.
- نشر مبادئ توجيهية أو مدونات قواعد سلوك إضافية خاصة بالقطاع لدعم تنفيذ مبادئ أخلاقيات الذكاء الاصطناعي.

ثانيا: على مستوى الجهات

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.

1. **رئيس الجهة / مسؤول إدارة البيانات**: يكون رئيس الجهة أو من يعينهم مسؤولين عن ممارسات أخلاقيات الذكاء الاصطناعي داخل الجهة، في حين يكون مسؤول إدارة البيانات مسؤولاً عن ممارسات أخلاقيات الذكاء الاصطناعي داخل جهة عامة، **وتشمل المسؤوليات ما يلي:**

- **اعتماد خطة أخلاقيات الذكاء الاصطناعي**: الموافقة على تنفيذ خطة أخلاقيات الذكاء الاصطناعي داخل الجهة والإشراف عليها.
- **توزيع أدوار أخلاقيات الذكاء الاصطناعي**: توزيع الأدوار المختلفة المتعلقة بأخلاقيات الذكاء الاصطناعي.
- **اعتماد التقرير السنوي لأخلاقيات الذكاء الاصطناعي**: اعتماد التقرير السنوي لأخلاقيات الذكاء الاصطناعي الذي يعده المدير التنفيذي للامتثال أو مسؤول الالتزام.
- **حل المشكلات**: اتخاذ الإجراءات اللازمة أو تفويضها لحل المشكلات التي يثيرها مسؤول الامتثال/ الالتزام.
- **التنسيق مع المكتب**: العمل كحلقة وصل بين الجهة والمكتب، ويتولى رئيس الجهة أو مسؤول إدارة البيانات حل أي مشاكل معلقة تتعلق بأخلاقيات الذكاء الاصطناعي للجهة المعنية وتصعيدها إلى المكتب إذا لزم الأمر.

2. **مسؤول الامتثال / الالتزام:** يكون مسؤول الامتثال أو الالتزام هو القائد الاستراتيجي لممارسات أخلاقيات الذكاء الاصطناعي. وفي الجهات العامة، يقع مسؤول الالتزام تحت إشراف مسؤول إدارة البيانات ويتبعه بشكل مباشر. يتولى مسؤول الالتزام المسؤوليات بالتشاور مع المسؤولين المعنيين بإدارة البيانات وحوكمة البيانات وخصوصية البيانات والبيانات المفتوحة ووظائف التحليلات. **تشمل مسؤوليات مسؤول الالتزام ما يلي:**

- **استراتيجية أخلاقيات الذكاء الاصطناعي**: الإشراف على وضع خطة أخلاقيات الذكاء الاصطناعي وتقديمها إلى رئيس الجهة. يجب على مسؤول الالتزام مراجعة أداء أخلاقيات الذكاء الاصطناعي لتحديد فرص التحسين والاستفادة من خطة أخلاقيات الذكاء الاصطناعي.
- **الإشراف على أخلاقيات الذكاء الاصطناعي**: مراجعة أنشطة تحديد أخلاقيات الذكاء الاصطناعي وتحديد أولوياتها، ومراقبة مؤشرات الأداء الرئيسية لأخلاقيات الذكاء الاصطناعي للأنظمة الداخلية والخارجية، وضمان تنفيذ أنشطة الصيانة.

- **الالتزام بأخلاقيات الذكاء الاصطناعي**: ضمان التزام الجهة بسياسات حوكمة البيانات الوطنية وأنشطة أخلاقيات الذكاء الاصطناعي، بما في ذلك على سبيل المثال لا الحصر تصنيف البيانات وحماية البيانات الشخصية وحرية المعلومات. التأكد من التزام أنظمة الأطراف الخارجية بهذه المبادئ من خلال الضمانات التعاقدية.
- **استشارة مكتب الأخلاقيات**: يجب على مسؤول الالتزام التشاور مع مكتب الأخلاقيات في حال وجود قضايا أو تظلّم.
- **التنسيق مع المكتب**: العمل كحلقة وصل ثانية بين الجهة ومكتب إدارة البيانات الوطنية (الجهة التنظيمية).

3. **مسؤول الذكاء الاصطناعي:** هو القائد التشغيلي المسؤول عن الذكاء الاصطناعي في الجهة. يتم توزيع هذا الدور من قبل مسؤول إدارة البيانات في الجهات العامة. يجب أن يعمل موظف الذكاء الاصطناعي المسؤول بشكل تعاوني مع الموظفين الآخرين العاملين في إدارة البيانات، وحوكمة البيانات، وحماية البيانات الشخصية، والبيانات المفتوحة، والتحليلات. **وتشمل المسؤوليات ما يلي:**

- **التخطيط لأخلاقيات الذكاء الاصطناعي**: وضع خطة أخلاقيات الذكاء الاصطناعي، بما في ذلك منهجية تحديد أولويات الذكاء الاصطناعي، وتحديد الأهداف ومؤشرات الأداء الرئيسية التي سيتم الاتفاق عليها مع اللجنة التنفيذية أو رئيس الجهة أو المدير التنفيذي للبيانات.
- **إدارة أخلاقيات الذكاء الاصطناعي**: إدارة أنشطة أخلاقيات الذكاء الاصطناعي داخل الجهة، ولا سيما:
  - تحديد إجراءات أخلاقيات الذكاء الاصطناعي.
  - تحديد أولويات إجراءات أخلاقيات الذكاء الاصطناعي.
  - تحديث إجراءات أخلاقيات الذكاء الاصطناعي وصيانتها ومراجعتها.
- **التثقيف والتوعية بأخلاقيات الذكاء الاصطناعي**: تثقيف موظفي الجهة ورفع مستوى الوعي لديهم حول أخلاقيات الذكاء الاصطناعي ودعم حملات التوعية الوطنية بالتنسيق مع المدير التنفيذي للامتثال/ مسؤول الالتزام.

4. **مطور نظم الذكاء الاصطناعي** - يتولى المسؤوليات التالية:

- تحديد المتطلبات وأهداف النظام بوضوح فيما يتعلق بمبادئ وضوابط أخلاقيات الذكاء الاصطناعي.
- تحديد الفوائد والمنافع المحتملة لأنظمة الذكاء الاصطناعي التي ستعود على الجهة والمستخدمين النهائيين.
- تصميم نظام ذكاء اصطناعي قوي وموثوق يمكنه تحمل المدخلات/ الأحداث غير المعروفة وغير المتوقعة.
- تصميم نظام ذكاء اصطناعي غير متحيز لأي مجموعات أو أفراد في عملية صنع القرار.
- تصميم نظام ذكاء اصطناعي يتوافق مع أنظمة وقوانين ومعايير حماية البيانات وكذلك الأحكام المرتبطة بعملية اتخاذ القرار المؤتمتة.
- تصميم آليات تقديم الملاحظات لتمكين عملية التواصل مع المستخدمين النهائيين وتمكين المستخدمين النهائيين من طلب الوصول إلى البيانات وتوفير خيارات تسجيل الدخول وإلغاء الاشتراك إذا كان نظام الذكاء الاصطناعي يستخدم البيانات الشخصية.
- التأكد من أن نظام الذكاء الاصطناعي يحترم ويحمي حقوق الإنسان الأساسية والقيم الثقافية للمملكة ويعزز الفوائد الاجتماعية والبيئية.
- تصميم نظام الذكاء الاصطناعي الذي تتماشى أهدافه مع أخلاقيات الجهة وقواعدها السلوكية.
- توثيق عملية التصميم بأكملها ووضع المعايير أو اتباع المعايير الحالية لعملية التوثيق لدعم إعادة إنتاج واستعادة نظام الذكاء الاصطناعي.
- تحديد الأدوار والمسؤوليات لمختلف الجهات المعنية وإدراج الموافقات اللازمة ونقاط اتخاذ القرار الحاسمة في جميع مراحل دورة حياة نظام الذكاء الاصطناعي.
- تصميم نظام يتضمن الإشراف والتوجيه البشري للآلة.
- تحديد مؤشرات الأداء الرئيسية للنظام فيما يتعلق بالدقة والأداء والمخاطر والعدالة والخصوصية والأمن وما إلى ذلك.
- وضع مواصفات واضحة تتناول المخاوف المتعلقة بمبادئ أخلاقيات الذكاء الاصطناعي أثناء عملية الشراء إذا كان نظام الذكاء الاصطناعي هو منتج تابع لجهة خارجية.
- تحديد السمات الحساسة التي قد يؤدي استخدامها في الخوارزميات إلى نتائج أو قرارات متحيزة.
- عدم استخدام السمات الحساسة كعينة للبيانات أثناء تطوير وتدريب أنظمة الذكاء الاصطناعي أو النماذج التنبؤية.
- اتخاذ الخطوات اللازمة لضمان خلو عملية أخذ عينات البيانات من التحيز وضمان تحققها من التحيز في السمات الحساسة.
- اتخاذ الخطوات اللازمة لضمان تنوع عينة البيانات وتمثيلها لجميع شرائح المجتمع أو الشرائح المستهدفة، وأن تكون عادلة ولن تؤدي إلى أي تمييز غير عادل.

- إعداد سجل مفصل لجميع أنشطة تحليل البيانات، بما في ذلك بيانات جميع العمليات والإجراءات التي تمت على كل مجموعة بيانات.
- إجراء تقييم النزاهة يجب أن تكون جميع القرارات التي تتعارض مع مبادئ النزاهة مبررة بأهداف الأعمال، وتحليل التكلفة والفائدة، ومفاضلة الأداء، وما إلى ذلك.
- بناء أنظمة ذكاء اصطناعي بسيطة وسهلة الاستخدام للمستخدمين النهائيين.
- اتخاذ الإجراءات اللازمة والتدابير الملائمة للتحقق من دقة تفسير البيانات.
- ضمان النزاهة في إصدار القرارات المهمة بشفافية عبر إتاحة إمكانية التحقق من العوامل الرئيسية التي تتم مراعاتها حال اتخاذ أي قرارات تنعكس على المصالح الأساسية للأفراد.
- إتاحة منهجية تدخل يدوي تمكن الأفراد من القدرة على تتبع مراحل اتخاذ القرارات الهامة ذات الصلة بمصالحهم الأساسية أو حتى سبل الاعتراض عليها.
- إعداد آلية تتضمن مجموعة من المعايير اللازمة لتقييم موثوقية أنظمة الذكاء الاصطناعي في التنبؤ واتخاذ القرارات المستقبلية.
- اعتماد منهجية شاملة لاختبار جودة الأنظمة والنماذج وخوارزميات الذكاء الاصطناعي القائمة على البيانات التنبؤية وفقاً للممارسات المعيارية.
- اتخاذ الخطوات اللازمة لضمان جودة ودقة عينة البيانات ومدى أهميته بالغرض من بناء النماذج التنبؤية وأنظمة الذكاء الاصطناعي.
- إعادة التقييم في مراحل عمل نظام الذكاء الاصطناعي ومراقبة مؤشرات الأداء الرئيسية على نظام الإنتاج وتحديد آليات المعالجة.

مهام الجهة المسؤولة عن نظام الذكاء الاصطناعي – تقوم بالمسؤوليات التالية:

- إعداد السياسات والتوجيهات لدعم وتمكين الاستخدام الأخلاقي للذكاء الاصطناعي وفقاً لأفضل ممارسات التطوير والتنفيذ وكذلك عمليات شراء المنتجات الخاصة بالجهات الخارجية.
- وضع البنود التعاقدية ذات الصلة لمزودي الخدمات من الجهات الخارجية للالتزام بمتطلبات هذه المبادئ.
- الالتزام بسياسات حوكمة البيانات الوطنية الصادرة عن مكتب إدارة البيانات الوطنية والمعتمدة من قبل مجلس إدارة الهيئة السعودية للبيانات والذكاء الاصطناعي (سدايا).
- الحصول على موافقة مكتب إدارة البيانات الوطنية -بعد التنسيق مع الجهة التنظيمية- قبل تحليل البيانات المصنفة باعتبارها سرية وفقاً لمبادئ تصنيف البيانات.
- ضمان اقتصار تحليل البيانات على مستويات التصنيف المقيدة والعامة -شريطة وجود حاجة لمعالجة البيانات لتحليلها- مع التحديد المناسب لها.
- اتخاذ الإجراءات اللازمة والتدابير الملائمة لضمان جودة ودقة وصلاحية البيانات المطلوب تحليلها، ومصداقية مصادرها، ومدى ملاءمة طرق جمعها.
- توفير قنوات مواتية تمكن الأفراد من الحصول على تفسيرات حول النتائج الهامة والقرارات التي تؤثر على مصالحهم الأساسية وكذلك الاعتراض أو حتى طلب إثبات مدى إنصاف تلك القرارات وعدالتها.
- إعداد إرشادات متعلقة بتوضيح آلية عمل النماذج التنبؤية أو خوارزميات الذكاء الاصطناعي المستخدمة والبيانات المطلوب تحليلها والشرائح المستهدفة والعوامل المؤثرة على النتائج والقرارات المهمة.
- اتخاذ الخطوات اللازمة لضمان عدم سيطرة الآلة ومباشرتها للأعمال وعدم اتخاذ أنظمة الذكاء الاصطناعي قرارات مهمة بالنيابة عن الأشخاص المخولين أو حتى التأثير على قراراتهم بدون موافقتهم المسبقة أو قدرتهم على الاستفسار عن النتائج أو الاعتراض عليها.
- إعداد وتوثيق سياسات وإجراءات حفظ وأرشفة البيانات وفقاً للأغراض المحددة والأنظمة والتشريعات ذات الصلة.
- التخلص من البيانات وإتلافها بأمان -بما في ذلك البيانات المؤرشفة والنسخ الاحتياطية -وفقاً لسياسة التخلص من البيانات المعتمدة لدى الجهة ووفقاً للأنظمة والسياسات المعمول بها ذات الصلة.
- إعداد دليل إجرائي يستعرض التدابير اللازمة لتقييم المخاطر الماثلة والتأثيرات المحتملة لتحليل البيانات باستخدام نماذج تنبؤية وخوارزميات ذكاء اصطناعي لقياس إنجاز الأهداف العامة بأقل تأثير ممكن على خصوصية الأفراد.
- إعداد دليل الإجراءات يستعرض تدابير تقييم أثر التحيز في النتائج للتأكد من أن مجموعة البيانات المطلوب تحليلها متنوعة وتمثل كافة الشرائح المستخدمة.

- حصر استخدام نتائج تحليل البيانات للغرض الذي أعدت من أجله والتي يتعين أن تتماشى مع الأنظمة واللوائح والسياسات المعمول بها ذات الصلة.
- ضمان الالتزام بأنظمة حماية البيانات.
- تدريب الموظفين على التعامل مع أنظمة الذكاء الاصطناعي.
- تمكين المستخدمين النهائيين من القدرة على طلب الوصول إلى البيانات أو الاعتراض على القرارات الصادرة وكذلك إمكانية الاستفسار عن بعض القرارات أو آلية عمل نظام الذكاء الاصطناعي.
- تحديد استراتيجية إدارة المخاطر للحد من المخاطر المتعلقة بتنفيذ إجراءات أخلاقيات الذكاء الاصطناعي.

الجهة المقيمة لنظام الذكاء الاصطناعي - تقوم بالمسؤوليات التالية:

- 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، والتي تشمل عملية تطوير الذكاء الاصطناعي وتنفيذه، وكذلك عمليات شراء منتجات الذكاء الاصطناعي الخاصة بالجهات الخارجية.

الالتزام

لتشجيع العمل بمبادئ اخلاقيات الذكاء الاصطناعي سيعلن مكتب إدارة البيانات الوطنية عن تقديم وسوم تقديرية في أخلاقيات الذكاء الاصطناعي. تعطى الوسوم للتطبيقات ذات المخاطر البسيطة والمطورة داخل أو خارج المملكة. ستعلن الوسوم وتسلم إلكترونياً عن طريق موقع مكتب إدارة البيانات الوطنية حيث يمكن لمطوري التطبيقات التسجيل الاختياري للحصول على الوسوم حسب مستوى الالتزام. تعتمد الوسوم على مستوى الالتزام في القائمة المرجعية المذكورة في مرفقات هذه المبادئ.

هناك أربعة مستويات من وسوم أخلاقيات الذكاء الاصطناعي (برونزي، فضي، ذهبي، بلاتيني). على الجهة التي ترغب في الحصول على الوسوم لمنتجاتها التسجيل أو تحديث التسجيل في الموقع أو البيئة التجريبية وعرض قدرات المنتج أو التطبيق العملية وتزويد المكتب بالمستندات التي تظهر مستوى الالتزام بمبادئ أخلاقيات الذكاء الاصطناعي حسب القائمة المرجعية. يحصل التطبيق أو المنتج على الوسم البرونزي في حال تسجيل المنتج أو التطبيق في الموقع وتزويد مكتب إدارة البيانات الوطنية بمعلومات المنتج الأساسية في القائمة المرجعية. يحصل المنتج على الوسم الفضي أو الذهبي أو البلاتيني وفقاً لمستوى الالتزام التي سيعلن عنها في الموقع.

ندعو الجهات العامة والخاصة على الاطلاع على القائمة المرجعية المذكورة في المبادئ والتسجيل في موقع مكتب إدارة البيانات الوطنية والبيئة التجريبية لدعمهم في تقييم الالتزام بتطبيق أخلاقيات الذكاء الاصطناعي الخاصة بهم وتقديم التقارير الاختيارية والحصول على الوسوم المميزة التي تثبت اهتمام المنتج بتطبيق معايير أخلاقيات الذكاء الاصطناعي. تحتوي التقارير المقترحة في القائمة المرجعية على:

- تقدُم المنتج أو الجهة في الالتزام بالقائمة المرجعية المذكورة في مرفقات هذا المستند.
- نتائج التقييم الداخلي أو الخارجي لأخلاقيات الذكاء الاصطناعي.
- أهداف المنتج أو الجهة ومؤشرات قياس أداء اخلاقيات الذكاء الاصطناعي.
- مستوى الالتزام بأخلاقيات الذكاء الاصطناعي وتحقيق متطلباتها والوسوم التي حصل عليها المنتج أو الجهة.

يمكن لمكتب إدارة البيانات الوطنية مساعدة الجهات في مراجعة التقارير السنوية ورفع التوصيات إلى مجلس إدارة الهيئة السعودية للبيانات والذكاء الاصطناعي بشأن الامتثال العام بأخلاقيات الذكاء الاصطناعي، كما يمكن للمكتب بناء على طلب الجهة المساعدة تسوية النزاعات المتعلقة بنشر تقرير أخلاقيات الذكاء الاصطناعي.

الملاحق

الملحق أ: أدوات أخلاقيات الذكاء الاصطناعي

الأدوات غير التقنية

- **إدارة المخاطر**: تعني إدارة مخاطر الذكاء الاصطناعي تقدير ومعالجة المخاطر الناتجة عن تطوير أو استخدام أنظمة الذكاء الاصطناعي، ويحدد التقييم المستويات القصوى للتعرض للمخاطر ويطلق عمليات وإجراءات التخفيف للحد من المخاطر الناشئة، ويمكن لإطار المخاطر المؤلف من أربعة مستويات التعرف على تباين مستويات المخاطر الناجمة عن أنظمة الذكاء الاصطناعي على صحة الفرد وسلامته و/أو حقوقه الأساسية، كما أنه يحدد المتطلبات والالتزامات المتناسبة مع كل مستوى من مستويات المخاطر الماثلة، وبالإضافة إلى ذلك تستخدم إدارة المخاطر لتصنيف فئات ومستويات المخاطر المرتبطة بتطوير و/أو استخدام تقنيات الذكاء الاصطناعي، وتصنف مستويات المخاطر هذه الى مخاطر بسيطة أو بلا مخاطر، ومخاطر محدودة، ومخاطر عالية، ومخاطر غير مقبولة، ولن يكون هناك أي قيود على أنظمة الذكاء الاصطناعي التي تشكل "مخاطر بسيطة أو لا تنطوي على أي مخاطر" -مثل مرشحات البريد العشوائي غير المرغوب فيها- ولكن سيتم تشجيع مزودي أنظمة الذكاء الاصطناعي على الالتزام بقواعد السلوك الطوعية. كما أنه سيسمح لأنظمة الذكاء الاصطناعي التي تشكل "مخاطر بسيطة أو لا تنطوي على أي مخاطر" -مثل مرشحات البريد العشوائي غير المرغوب فيها- بدون وضع أي قيود، ولكن سيتم تشجيع موردي أنظمة الذكاء الاصطناعي على الالتزام بالأخلاقيات. يجب أن تخضع أنظمة الذكاء الاصطناعي التي تشكل "مخاطر محدودة" -مثل روبوتات الدردشة- لالتزامات الشفافية (مثل: الوثائق التقنية المتعلقة بالوظيفة والتطوير والأداء) ويمكن أن تختار بالمثل الالتزام بالأخلاقيات. وتخدم التزامات الشفافية -من بين العديد من العوامل الأخرى- في السماح للجهات المسؤولة عن أنظمة الذكاء الاصطناعي باتخاذ قرارات حول سبل دمج برمجيات الذكاء الاصطناعي في منتجاتهم و/أو خدماتهم، ويتعين على أنظمة الذكاء الاصطناعي التي تشكل "مخاطر عالية" على الحقوق الأساسية الخضوع لإجراء تقييمات ما قبل المطابقة وبعدها، ولا يُسمح بأنظمة الذكاء الاصطناعي التي تشكل "خطراً غير مقبول" على سلامة الناس وسبل عيشهم وحقوقهم كتلك المتعلقة بالتسجيل الاجتماعي أو استغلال الأطفال أو تشويه السلوك البشري الذي يحتمل أن تحدث فيه أضرار جسدية أو نفسية. هذا، وينبغي أن تكون إدارة المخاطر مرتبطة ارتباطاً مباشراً بمبادرات الذكاء الاصطناعي، بحيث 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، بحيث تكون المراقبة ثابتة ومتسقة مع أنشطة تطوير الذكاء الاصطناعي. يتم تضمين المعايير والاختبارات والضوابط في هذا النهج في مراحل مختلفة من مراحل عمل نظام الذكاء الاصطناعي، بدءاً من مرحلة التصميم وصولاً إلى التطوير وبعد النشر والتنفيذ الفعلي.

- **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 فإنه 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 على تعزيز القدرات والمهارات البشرية، ويمكن 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 المالكة لنظام الذكاء الاصطناعي من القدرة على معالجة المشكلات المحددة وتعزيز المجالات التي تستلزم إدخال تحسينات وتعديلات، كما أنه من الضروري أيضاً القدرة 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 والأمان:** يساعد معيار الخصوصية والأمان 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 الأمان والخصوصية:

م	اسم المعيار	الرابط
1	معيار ISO/IEC 27000 Family (المنظمة الدولية للمعايير)	https://www.iso.org/isoiec-27001-information-security.html
2	معيار ISO 31000 Family (المنظمة الدولية للمعايير)	https://www.iso.org/iso-31000-risk-management.html
3	إطار عمل الأمن السيبراني للمعهد الوطني للمعايير والتقنية (NIST)	https://www.nist.gov/cyberframework/framework
4	إطار عمل إدارة مخاطر الذكاء الاصطناعي الصادر عن المعهد الوطني للمعايير والتقنية (NIST) - قيد التنفيذ	https://www.nist.gov/itl/ai-risk-management-framework
5	ضوابط مركز ضوابط أمن الإنترنت (CIS)	https://www.cisecurity.org/controls/
6	معيار أمان بيانات صناعة بطاقات الدفع (PCI-DSS)	https://www.pcisecuritystandards.org/pci_security/
7	أهداف التحكم بالمعلومات والتقنيات ذات الصلة (COBIT)	http://www.isaca.org/resources/cobit

الأدوات التقنية

- **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. تُعد المبادئ والضوابط عامة وينبغي التعامل معها في حالات الاستخدام أو أنظمة الذكاء الاصطناعي المحددة من خلال نهج منطقي واضح وسهل الشرح والفهم. نحتاج الخطوات الثلاث التالية لمواءمة البنية مع أخلاقيات الذكاء الاصطناعي:
  - **الإحساس**: ينبغي تطوير نظام بحيث 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 باستخدام 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 مبالغ فيها أو يستهان بها بشكل منتظم بالنسبة لمجموعة أو أكثر مقارنة بالمجموعات الأخرى. وبالإضافة إلى ذلك، فإنه يقيِّم مدى تنوع البيانات الممثلة لكل مجموعة. هذا، ويتضمن الجدول التالي أمثلة على أدوات تقييم العدالة:

م	اسم الأداة	الرابط
1	Google Model Card Toolkit	https://github.com/tensorflow/model-card-toolkit
2	IBM AI Fairness 360	https://aif360.mybluemix.net/

3	Microsoft Fairlearn	https://fairlearn.org/
4	Google What-if Tool	https://pair-code.github.io/what-if-tool/
5	Aequitas Bias and Fairness Audit Toolkit	http://aequitas.dssg.io/
6	Veritas Fairness Assessment Tool	https://github.com/mas-veritas2/veritastool

- **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 صندوق أسود- 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 من 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 في أنظمة 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:** تجب 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 تلك الطرق على حماية البيانات من خلال تطبيق طرق تحويل البيانات، وخاصة بالنسبة للبيانات المصنفة على أنها حساسة، ويتم تقديم طرق حماية البيانات التالية كمثال، وينبغي أن تكون بعد تصنيف البيانات.

  - **إلغاء تحديد البيانات** (Data De-Identification) هو التخلص من البيانات المحددة للهوية الشخصية (Personally PIIIdentifiable Data) من أي مستند أو وسائط أخرى، بما في ذلك المعلومات الصحية المحمية للأفراد (PHIProtected Health Information).

  - **إخفاء هوية البيانات وعدم الكشف عنها** (Data Anonymization) 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** (Data Masking) هو التخلص من المعلومات أو إخفائها، وتستبدلها ببيانات بديلة واقعية أو حتى معلومات وهمية زائفة، والهدف منها هو 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** (Data Pseudonymization) هي طريقة لإخفاء البيانات التي تضمن عدم إمكانية إسناد البيانات الشخصية لشخص معين، دون استخدام معلومات إضافية خاضعة للتدابير الأمنية، وهي جزء لا يتجزأ من اللائحة العامة لحماية البيانات في النظام الأوروبي العام لحماية البيانات (GDPR)، والتي تحتوي على العديد من الحيثيات التي تحدد كيفية استخدام البيانات المستعارة والتوقيت المناسب لذلك.

  - **تشفير البيانات** (Data Encryption) 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** (Data Tokenization) 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 الرموز

(Tokens) عبارة عن أرقام عشوائية تماماً أو يتم إنشاؤها بواسطة وظائف أحادية أو متعددة الاتجاهات (hashes).

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

نوع الأداة	الأداة	التخطيط والتصميم	إعداد بيانات المدخلات	البناء والتحقق	التطبيق والمتابعة
الأدوات غير التقنية	إدارة المخاطر	✓	✓	✓	✓
الأدوات غير التقنية	تقرير عدالة الذكاء الاصطناعي	✓			
الأدوات غير التقنية	تقييم الأثر الأخلاقي	✓			✓
الأدوات غير التقنية	معايير الخصوصية والأمان	✓	✓	✓	✓
الأدوات التقنية	موثوقية بنية الذكاء الاصطناعي	✓			
الأدوات التقنية	تقييم الخوارزميات			✓	✓
الأدوات التقنية	تقييم العدالة			✓	
الأدوات التقنية	تقرير تفسير طريقة عمل الذكاء الاصطناعي			✓	✓
الأدوات التقنية	تدقيق الخوارزميات			✓	✓
الأدوات التقنية	التقييم الذاتي للسلامة	✓	✓	✓	✓
الأدوات التقنية	طرق حماية البيانات	✓	✓	✓	✓

الملحق ج: القائمة المرجعية لأخلاقيات الذكاء الاصطناعي

المرحلة الأولى لدورة عمل نظام الذكاء الاصطناعي: التخطيط والتصميم

الرقم	متطلبات القائمة المرجعية	ملزمة للجهات الخارجية	المبادئ
التخطيط والتصميم 1	هل تم تصميم مستوى إشرافي مناسب لنظام الذكاء الاصطناعي وحالة الاستخدام؟	نعم	المساءلة والمسؤولية
التخطيط والتصميم 2	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 3	هل تم تحديد عمليات الإشراف البشري باستخدام مؤشرات الأداء الرئيسية المناسبة وتحديد المسؤولية للأطراف ذات الصلة؟	لا	المساءلة والمسؤولية
التخطيط والتصميم 4	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 5	هل تمت مراعاة متطلبات حماية المسؤولية والجهة صاحبة البيانات وأخذها بعين الاعتبار؟	نعم	المساءلة والمسؤولية

التخطيط والتصميم 6	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 7	هل تم توفير التدريب والتعليم اللازم للمساعدة في تطوير ممارسات المساءلة؟	لا	المساءلة والمسؤولية
التخطيط والتصميم 8	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 9	هل تم التأكد من أن هيكل حوكمة أخلاقيات الذكاء الاصطناعي يتضمن آليات تدقيق داخلية أو خارجية؟	لا	المساءلة والمسؤولية
التخطيط والتصميم 10	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
التخطيط والتصميم 11	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
التخطيط والتصميم 12	هل تم تحديد مؤشرات الأداء الرئيسية لتقييم النزاهة؟	لا	النزاهة
التخطيط والتصميم 13	هل تم النظر في وضع آلية تشمل مشاركة مختلف الأطراف المعنية في تطوير واستخدام نظام الذكاء الاصطناعي؟	لا	النزاهة
التخطيط والتصميم 14	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 15	هل تم اتخاذ تدابير لضمان ألا يؤدي نظام الذكاء الاصطناعي إلى خداع الناس أو الإضرار بحرية اختيارهم بدون مبرر؟	نعم	القيم الإنسانية
التخطيط والتصميم 16	هل تمت مواءمة نظام الذكاء الاصطناعي لديكم مع المعايير أو السياسات ذات الصلة (مثل معيار الآيزو ومعيار IEEE وقانون خصوصية البيانات) أو البروتوكولات المعتمدة على نطاق واسع لإدارة وحوكمة البيانات اليومية؟	نعم	الخصوصية والأمان
التخطيط والتصميم 17	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 والأمان
التخطيط والتصميم 18	هل تم التأكد من أن التحكم في الوصول إلى البيانات يلبي متطلبات الأمان والخصوصية والالتزام؟ هل تم تصميم آلية تسجيل لأغراض التدقيق والتصحيح؟	نعم	الخصوصية والأمان
التخطيط والتصميم 19	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 20	هل تم تقييم مدى احتمالية أن يتسبب نظام الذكاء الاصطناعي في إلحاق الضرر أو الأذى لكل 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 21	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 22	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 23	هل تم تقييم ما يتماشى نموذج عمل النظام مع رؤية الجهة ورسالتها وكذلك مدونة قواعد السلوك؟	نعم	الشفافية والقابلية للتفسير
التخطيط والتصميم 24	هل تم تقييم تصميم نظام ذكاء اصطناعي قابل للتفسير بحيث تكون البيانات والخوارزميات والمخرجات والقرارات شفافة وقابلة للتفسير لجميع الأطراف المعنية ذات الصلة؟	نعم	الشفافية والقابلية للتفسير
التخطيط والتصميم 25	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 26	هل كان هناك افتراض بأن هناك مفاضلة تجارية؟	نعم	النزاهة
التخطيط والتصميم 27	هل تم وضع آلية لقياس أو تقييم أثر الخصوصية؟	نعم	الخصوصية والأمان
التخطيط والتصميم 28	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

الرقم	متطلبات القائمة المرجعية	ملزمة للجهات الخارجية	المبادئ
تهيئة البيانات 1	هل توجد آلية ثابتة تحدد المشكلات المتعلقة بخصوصية البيانات أو حمايتها في عملية جمع البيانات ومعالجتها؟	نعم	الخصوصية والأمان
تهيئة البيانات 2	هل تمت مراجعة البيانات من حيث النطاق والتصنيف؟	لا	الخصوصية والأمان
تهيئة البيانات 3	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 والأمان
تهيئة البيانات 4	هل توجد آلية محددة للتحكم في استخدام البيانات الشخصية (مثل الموافقة الصحيحة وإمكانية الإلغاء، عند الاقتضاء)؟	نعم	الخصوصية والأمان
تهيئة البيانات 5	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 والأمان
تهيئة البيانات 6	هل تم تقييم جودة ومصدر البيانات التي تم الحصول عليها من خلال عمليات محددة؟	لا	الخصوصية والأمان
تهيئة البيانات 7	هل كان هناك تقييم حول مدى إمكانية إجراء التحليل بعد التدريب واختبار البيانات؟	لا	الشفافية والقابلية للتفسير

تهيئة البيانات 8	هل تمت دراسة أو مراجعة التنوع وشمول مجموعة البيانات الحالية؟	لا	النزاهة
تهيئة البيانات 9	هل هناك آلية محددة لقياس ما إذا تم تقييم سلامة وجودة ودقة جمع البيانات ومصادرها وتحديثها؟	لا	المساءلة والمسؤولية
تهيئة البيانات 10	هل تم تطوير عملية تحليل لمزايا الخصائص الحساسة؟	نعم	النزاهة
تهيئة البيانات 11	هل قام الفريق بتقييم تصنيف البيانات ومعالجتها والوصول إليها لضمان الحصول عليها بشكل صحيح؟	نعم	القيم الإنسانية
تهيئة البيانات 12	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 13	هل تم تصنيف البيانات استناداً إلى توصيات مكتب إدارة البيانات الوطنية؟ وفي حال استخدام معايير أخرى غير الواردة، يرجى ذكرها.	نعم	المزايا الاجتماعية والبيئية
تهيئة البيانات 14	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

الرقم	متطلبات القائمة المرجعية	ملزمة للجهات الخارجية	المبادئ
البناء وقياس الأداء 1	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 2	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 3	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 4	هل هناك آلية تواصل ثابتة لضمان موثوقية النظام من قبل المستخدمين النهائيين؟	نعم	الموثوقية والسلامة
البناء وقياس الأداء 5	هل هناك تعريفات واضحة ومفهومة لشرح سبب اتخاذ نظام الذكاء الاصطناعي قرار محدد؟	لا	الشفافية والقابلية للتفسير
البناء وقياس الأداء 6	هل تم بناء النموذج بطريقة بسيطة وقابلة للتفسير؟	لا	الشفافية والقابلية للتفسير
البناء وقياس الأداء 7	هل تم اختبار قابلية تفسير نموذج الذكاء الاصطناعي بنجاح بعد تدريب النموذج؟	لا	الشفافية والقابلية للتفسير
البناء وقياس الأداء 8	هل تم إجراء بحث يتعلق باستخدام الأدوات التقنية المتاحة لتحسين فهم البيانات والنموذج وأدائه؟	لا	النزاهة

البناء وقياس الأداء 9	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
البناء وقياس الأداء 10	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 11	هل وافقت الأطراف المعنية على الاختبارات الناجحة وجولات التحقق من قبول المستخدمين قبل إعداد نماذج الذكاء الاصطناعي؟	نعم	المساءلة والمسؤولية
البناء وقياس الأداء 12	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
البناء وقياس الأداء 13	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

الرقم	متطلبات القائمة المرجعية	ملزمة للجهات الخارجية	المبادئ
التطبيق والمتابعة 1	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 2	هل أجرى الفريق تقييماً لنقاط ضعف نظام الذكاء الاصطناعي لمواجهة الهجمات السيبرانية المحتملة أو الكشف عن البيانات الحساسة أو خرق السرية؟	نعم	الخصوصية والأمان
التطبيق والمتابعة 3	هل هناك آليات لقياس ما إذا كان النظام ينتج كمية غير مقبولة من التوقعات غير الدقيقة؟	لا	المساءلة والمسؤولية
التطبيق والمتابعة 4	هل توجد استراتيجية محددة لمتابعة وقياس ما إذا كان نظام الذكاء الاصطناعي يحقق الأهداف والأغراض والتطبيقات على نحو المطلوب؟	لا	الموثوقية والسلامة
التطبيق والمتابعة 5	هل يتمتع الأشخاص المخولين بالوصول إلى البيانات بالكفاءات اللازمة لفهم تفاصيل متطلبات حماية البيانات؟	لا	الخصوصية والأمان
التطبيق والمتابعة 6	هل هناك آليات مطبقة لتقييم مستوى تأثير نظام الذكاء الاصطناعي على عمليات صنع القرار للمستخدمين النهائيين؟	لا	الشفافية والقابلية للتفسير
التطبيق والمتابعة 7	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 إليها؟	لا	الشفافية والقابلية للتفسير
التطبيق والمتابعة 8	هل هناك عملية محددة لجمع ملاحظات المستخدمين النهائيين ودراستها واعتمادها على النظام؟	نعم	الشفافية والقابلية للتفسير
التطبيق والمتابعة 9	هل هناك عمليات قائمة وتحليل كمي لمتابعة التحيزات والإنصاف العام للنظام أثناء مراحل التنفيذ؟	نعم	النزاهة
التطبيق والمتابعة 10	في حال وجود تباين، هل تم وضع آلية لقياس أو تقييم الأثر المحتمل لمثل هذا التباين على الحقوق الأساسية؟	لا	النزاهة

التطبيق والمتابعة 11	هل هناك آليات محددة لضمان العدالة والإنصاف في أنظمة الذكاء الاصطناعي الخاصة بك؟	لا	النزاهة
التطبيق والمتابعة 12	هل يمكن للمستخدمين النهائيين للتقنيات المساعدة الوصول إلى المعلومات المتعلقة بنظام الذكاء الاصطناعي؟	لا	النزاهة
التطبيق والمتابعة 13	هل هناك آليات محددة لقياس الأثر الاجتماعي والبيئي لتنفيذ واستخدام نظام الذكاء الاصطناعي؟	نعم	المزايا الاجتماعية والبيئية
التطبيق والمتابعة 14	هل هناك آليات محددة لضمان تطبيق حقوق الإنسان الأساسية؟	نعم	المساءلة والمسؤولية
التطبيق والمتابعة 15	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 16	هل هناك آليات محددة لإثبات مدى التزامك بالمبادئ المنصوص عليها في هذا المستند؟	لا	المساءلة والمسؤولية
التطبيق والمتابعة 17	هل هناك آليات محددة تسمح بالتعويض في حالة حدوث أي ضرر أو تأثير سلبي؟	نعم	المساءلة والمسؤولية
التطبيق والمتابعة 18	هل هناك آليات محددة توفر معلومات للمستخدمين النهائيين أو الجهات الخارجية بخصوص فرص التعويض المتاحة؟	نعم	المساءلة والمسؤولية
التطبيق والمتابعة 19	هل هناك تقنيات مراقبة مستمرة لضمان الحفاظ على الخصوصية والأمان في نظام الذكاء الاصطناعي؟	نعم	الخصوصية والأمان
التطبيق والمتابعة 20	هل هناك تقييمات دورية لأنظمة الذكاء الاصطناعي المستخدمة لضمان تطبيق حقوق الإنسان الأساسية والقيم الثقافية للمملكة؟	نعم	القيم الإنسانية