NIST AI 100-1

إطار عمل إدارة مخاطر الذكاء الاصطناعي (AI RMF 1.0)

NIST | NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST AI 100-1

إطار عمل إدارة مخاطر الذكاء الاصطناعي (AI RMF 1.0)

يتوفر هذا المنشور مجانًا عبر الرابط:
https://doi.org/10.6028/NIST.AI.100-1

يناير 2023

وزارة التجارة الأمريكية - جينا ماري رايموندو،
وزيرة التجارة الأمريكية

المعهد الوطني للمعايير والتكنولوجيا

لوري إي لوكاسيو، مدير المعهد الوطني للمعايير والتكنولوجيا، ونائب وزير التجارة للمعايير والتكنولوجيا

جرى تحديد بعض الجهات أو المعدات أو المواد التجارية في هذه الوثيقة بهدف وصف إجراء أو مفهوم تجريبي على نحو مناسب. ولا يُقصد بهذا التحديد أن ينطوي ضمنًا على توصية أو تأييد من قِبل المعهد الوطني للمعايير والتكنولوجيا، كما لا يُقصد به بأنه يقتضي ضمنًا أن تلك الجهات أو المواد أو المعدات هي بالضرورة أفضل ما هو متاح لتحقيق هذا الغرض.

يتوفر هذا المنشور مجانًا عبر الرابط: https://doi.org/10.6028/NIST.AI.100-1

تحديث الجدول الزمني والإصدارات

يهدف إطار عمل إدارة مخاطر الذكاء الاصطناعي (AI RMF) إلى أن يكون بمثابة وثيقة قابلة للتعديل.

ويتولى المعهد الوطني للمعايير والتكنولوجيا عملية مراجعة محتوى إطار العمل وتقييم مدى فائدته على نحو منتظم، وذلك من أجل تحديد ما إذا كان إجراء التحديث مناسبًا؛ ومن المتوقع إجراء مراجعة بمدخلات رسمية صادرة عن مجتمع الذكاء الاصطناعي في موعد أقصاه عام 2028. إذ يستخدم إطار العمل نظام إصدار مكون من رقمين لتتبع التغيرات الرئيسية والثانوية وتحديدها، حيث يمثل الرقم الأول إنشاء إطار عمل إدارة مخاطر الذكاء الاصطناعي (AI RMF) والوثائق ذات الصلة المصاحبة له (على سبيل المثال، 1.0)، على أن يتغير هذا الرقم فقط مع إجراء المراجعات الرئيسية. ويجري تتبع المراجعات الطفيفة باستخدام "الرقم." بعد رقم إنشاء إطار العمل (على سبيل المثال، 1.1). وسيتم تتبع جميع التغييرات التي أُجريت باستخدام جدول التحكم في الإصدار الذي يحدد ما حدث، يتضمن ذلك رقم الإصدار وتاريخ التغيير ووصف التغيير. في حين يخطط المعهد الوطني للمعايير والتكنولوجيا لتحديث دليل إطار عمل إدارة مخاطر الذكاء الاصطناعي بانتظام. يمكن إرسال التعليقات بشأن دليل إطار العمل (AI RMF) عبر البريد الإلكتروني إلى AIframework@nist.gov في أي وقت، وسيجري مراجعتها وإدماجها على أساس نصف سنوي.

جدول المحتويات

قائمة الجداول

قائمة الأشكال التوضيحية

الملخص التنفيذي

تتميز تقنيات الذكاء الاصطناعي (AI) بإمكانيات ضخمة قادرة على تغيير حياة المجتمع والبشر، يتضمن ذلك مجالات التجارة والصحة والنقل والأمن السيبراني والبيئة وحتى الحياة على كوكبنا. وقد تؤدي تقنيات الذكاء الاصطناعي إلى دفع عجلة تحقيق النمو الاقتصادي الشامل ودعم التطورات العلمية الرامية إلى تحسين ظروف عالمنا. ومع ذلك، فقد تشكل تقنيات الذكاء الاصطناعي أيضًا مخاطر من شأنها التأثير سلبًا في حياة الأفراد والجماعات والمنظمات والمجتمعات المحلية والمجتمع والبيئة وكوكب الأرض. مثلها مثل المخاطر المتعلقة بأنواع التقنيات الأخرى، فقد تظهر مخاطر الذكاء الاصطناعي بالعديد من الطرق، والتي يمكن وصفها بأنها طويلة أو قصيرة الأمد، ذات احتمالية عالية أو منخفضة، نظامية أو محلية، وذات تأثير مرتفع أو منخفض.

يشير إطار عمل إدارة مخاطر الذكاء الاصطناعي (AI RMF) إلى نظام الذكاء الاصطناعي باعتباره نظامًا هندسيًا أو نظامًا قائمًا على الآلة يمكنه، لمجموعة معينة من الأهداف، تحقيق مخرجات مثل التنبؤات أو التوصيات أو القرارات التي تؤثر في البيئات الواقعية أو البيئات الافتراضية. فقد صُممت أنظمة الذكاء الاصطناعي للعمل ضمن مستويات مختلفة من الاستقلالية (مقتبس من: التوصية الصادرة عن منظمة التعاون الاقتصادي والتنمية بشأن الذكاء الاصطناعي: لعام 2019؛ ISO / IEC 22989: 2022).

في حين أن هناك الكثير من المعايير وأفضل الممارسات الرامية إلى مساعدة المؤسسات على التخفيف من مخاطر البرامج التقليدية أو الأنظمة القائمة على المعلومات، فإن المخاطر التي تشكلها أنظمة الذكاء الاصطناعي تتسم بتفردها من عدة أوجه (انظر الملحق ب). قد تكون أنظمة الذكاء الاصطناعي، على سبيل المثال، مؤهلة للتعامل مع البيانات التي يمكن أن تتغير بمرور الوقت، وأحيانًا بصورة ملموسة وعلى نحو غير متوقع، مما يؤثر في وظائف النظام وموثوقيته بطرق يصعب فهمها. وغالبًا ما تتسم أنظمة الذكاء الاصطناعي والسياقات التي تُنشر ضمنها بالتعقيد، مما يجعل من العسير اكتشاف الأعطال والاستجابة لها عند حدوثها. كما تتسم أنظمة الذكاء الاصطناعي بالطابعين الاجتماعي والتقني في جوهرها، مما يعني أنها تتأثر بالديناميكيات المجتمعية والسلوك البشري. فقد تنشأ مخاطر الذكاء الاصطناعي – وكذلك مزاياه – من تفاعل الجوانب التقنية التي تقترن بالعوامل المجتمعية المتعلقة بآلية استخدام النظام، وتفاعلاته مع أنظمة الذكاء الاصطناعي الأخرى، ومَن الجهة المنوطة بتشغيله، والسياق الاجتماعي الذي يُنشر فيه.

تساهم هذه المخاطر في جعل الذكاء الاصطناعي تقنية تحفّ بها تحديات فريدة فيما يتعلق بنشرها واستخدامها في المنظمات وداخل المجتمع. وفي حال الافتقار إلى الضوابط المناسبة، يمكن لأنظمة الذكاء الاصطناعي تضخيم أو إدامة أو مفاقمة النتائج غير المنصفة أو غير المرغوب فيها للأفراد والمجتمعات. وفي حال الاستعانة بالضوابط المناسبة، يمكن لأنظمة الذكاء الاصطناعي التخفيف من النتائج غير المنصفة وإدارتها.

وتمثل إدارة مخاطر الذكاء الاصطناعي عنصرًا أساسيًا في التطوير والاستخدام المسؤول لأنظمة الذكاء الاصطناعي. وقد تساعد ممارسات الذكاء الاصطناعي المسؤولة في مواءمة القرارات المتعلقة بتصميم نظام الذكاء الاصطناعي وتطويره واستخداماته مع الأهداف والقيم المنشودة. في حين تركز المفاهيم الأساسية في الذكاء الاصطناعي المسؤول على مركزية الإنسان والمسؤولية الاجتماعية والاستدامة. وقد تؤدي إدارة مخاطر الذكاء الاصطناعي إلى دفع الاستخدامات والممارسات المسؤولة من خلال حث المنظمات وفرقها الداخلية المعنية بتصميم الذكاء الاصطناعي وتطويره ونشره على التفكير الناقد بصورة أكبر بشأن السياق والتأثيرات السلبية والإيجابية المحتملة أو غير المتوقعة. إذ يساهم فهم مخاطر أنظمة الذكاء الاصطناعي وإدارتها على تعزيز مدى الجدارة بالثقة، وبالتالي ترسيخ ثقة الرأي العام.

يمكن أن تشير المسؤولية الاجتماعية إلى مسؤولية المنظمة "فيما يتعلق بتأثيرات قراراتها وأنشطتها في المجتمع والبيئة عن طريق اتباع سلوك شفاف وأخلاقي" (ISO 26000:2010). ويشير مفهوم الاستدامة إلى "حالة النظام العالمي، يتضمن ذلك الجوانب البيئية والاجتماعية والاقتصادية، حيث يتم تلبية احتياجات الجيل الحالي دون المساس بقدرة الأجيال المستقبلية على تلبية احتياجاتهم الخاصة" (ISO/IEC TR 24368:2022). في حين يهدف الذكاء الاصطناعي المسؤول إلى إنتاج تقنية منصفة وخاضعة للمساءلة أيضًا. ويعني التوقع تنفيذ الممارسات التنظيمية بما يتماشى مع "المسؤولية المهنية"، التي تحددها منظمة ISO على أنها نهج "يهدف إلى ضمان أن المهنيين الذين يصممون أو يطورون أو ينشرون أنظمة وتطبيقات الذكاء الاصطناعي أو المنتجات أو الأنظمة القائمة على تقنية الذكاء الاصطناعي، يدركون وضعها الفريد للتأثير في البشر والمجتمع ومستقبل الذكاء الاصطناعي" (ISO / IEC TR 24368: 2022).

وبناءً على توجيهات "قانون مبادرة الذكاء الاصطناعي الوطنية" لعام 2020 (P.L. 116-283), يكمن الهدف من وضع إطار عمل إدارة مخاطر الذكاء الاصطناعي (AI RMF) في تقديم مورد إلى المنظمات المعنية بتصميم أنظمة الذكاء الاصطناعي أو تطويرها أو نشرها أو استخدامها للمساعدة في إدارة العديد من مخاطر الذكاء الاصطناعي وتعزيز تطوير واستخدام أنظمة الذكاء الاصطناعي الجديرة بالثقة والمسؤولة. ويهدف إطار العمل إلى أن يكون طوعيًا، ويحافظ على الحقوق، وغير خاص بقطاع محدد، ومتوافق مع حالة الاستخدام، ويوفر المرونة اللازمة للمنظمات بمختلف الأحجام والقطاعات التابعة لها وفي جميع أنحاء المجتمع، وذلك من أجل تنفيذ النهج في إطار العمل هذا.

صُمم إطار العمل (AI RMF) بغرض تزويد المنظمات والأفراد – يشار إليهم في هذه الوثيقة باسم الجهات الفاعلة في مجال الذكاء الاصطناعي – بأساليب رامية إلى زيادة موثوقية أنظمة الذكاء الاصطناعي، وللمساعدة في تعزيز التصميم والتطوير ونشر واستخدام أنظمة الذكاء الاصطناعي بشكل مسؤول مع مرور الوقت. وتحدد منظمة التعاون الاقتصادي والتنمية (OECD) الجهات الفاعلة في مجال الذكاء الاصطناعي على أنها "الجهات التي تؤدي دورًا نشطًا في دورة حياة نظام الذكاء الاصطناعي، بما في ذلك المنظمات والأفراد الذين ينشرون أنظمة الذكاء الاصطناعي أو العاملين في هذا المجال" [منظمة التعاون الاقتصادي والتنمية لعام (2019)، الذكاء الاصطناعي في المجتمع—المكتبة الإلكترونية لمنظمة التعاون الاقتصادي والتنمية] (انظر الملحق أ).

يهدف إطار العمل (AI RMF) إلى أن يكون إطارًا عمليًا، للتكيف مع مشهد الذكاء الاصطناعي مع استمرار تطور تقنيات الذكاء الاصطناعي، ودخوله حيز التنفيذ من قِبل المنظمات بدرجات وقدرات متفاوتة حتى يتمكن المجتمع من الاستفادة من تقنيات الذكاء الاصطناعي بجانب الحماية أيضًا من أضراره المحتملة.

ومن المقرر تحديث إطار العمل والموارد الداعمة وتوسيع نطاقها وتحسينها استنادًا إلى التكنولوجيا المتطورة، ومشهد المعايير في جميع أنحاء العالم، وتجربة مجتمع الذكاء الاصطناعي وردود الفعل المتعلقة به. ويواصل المعهد الوطني للمعايير والتكنولوجيا مواءمة إطار العمل (AI RMF) والإرشادات ذات الصلة بالمعايير والمبادئ التوجيهية والممارسات الدولية السارية. وبمجرد وضع إطار العمل حيز الاستخدام، سيجري استخلاص دروس مستفادة إضافية للدلالة على التحديثات المستقبلية والموارد الإضافية.

ينقسم إطار العمل إلى جزأين، يناقش الجزء الأول كيف يمكن للمؤسسات تأطير المخاطر المتعلقة بالذكاء الاصطناعي ويصف فئات الجمهور المستهدفة. بعد ذلك، يجري تحليل مخاطر الذكاء الاصطناعي والجدارة بالثقة، وتحديد سمات أنظمة الذكاء الاصطناعي الجديرة بالثقة؛ التي تتضمن أن تكون تلك الأنظمة صالحة وموثوقة وآمنة ومأمونة ومرنة ومسؤولة وشفافة وقابلة للتفسير والتوضيح وتتسم بالخصوصية المحسنة والإنصاف مع إدارة تحيزاتها الضارة.

بينما يتناول الجزء الثاني "جوهر" إطار العمل، والذي يصف أربع وظائف محددة لمساعدة المؤسسات في معالجة مخاطر أنظمة الذكاء الاصطناعي على الصعيد العملي. وتُصنف هذه الوظائف الأربعة، التي تتضمن وظيفة الحوكمة ووظيفة التخطيط ووظيفة القياس ووظيفة الإدارة، إلى فئات رئيسية وفئات فرعية. بينما تطبق وظيفة الحوكمة على جميع مراحل عمليات إدارة مخاطر الذكاء الاصطناعي في المؤسسات وإجراءاتها المتبعة، في حين يمكن تطبيق وظائف التخطيط والقياس والإدارة في سياقات خاصة بنظام الذكاء الاصطناعي وخلال مراحل محددة من دورة حياة الذكاء الاصطناعي.

ترد الموارد الإضافية المتعلقة بإطار العمل في دليل إطار العمل (AI RMF)، والذي يُتاح على الموقع الإلكتروني التابع لدليل إطار عمل إدارة مخاطر الذكاء الاصطناعي الصادر عن المعهد الوطني للمعايير والتكنولوجيا (NIST AI RMF):
https://www.nist.gov/itl/ai-risk-management-framework.

إنَّ اضطلاع المعهد الوطني للمعايير والتكنولوجيا بتطوير إطار العمل (AI RMF) بالتعاون مع القطاعين العام والخاص هو أمر موجه ومتسق مع جهوده الأوسع نطاقًا في مجال الذكاء الاصطناعي المنصوص عليها في قانون المبادرة الوطنية للذكاء الاصطناعي لعام 2020، وتوصيات لجنة الأمن القومي للذكاء الاصطناعي، وخطة المشاركة الصادرة عن الحكومة الفيدرالية في تطوير المعايير الفنية والأدوات ذات الصلة.ويُعدّ الانخراط مع مجتمع الذكاء الاصطناعي في أثناء تطوير إطار العمل هذا – عن طريق الاستجابة لطلب رسمي يتعلق بتلقي المعلومات، وإقامة ثلاث ورش عمل حظيت بمشاركة واسعة، وإبداء تعليقات عامة حول ورقة مفاهيمية ومسودتين بشأن هذا الإطار، وإجراء مناقشات في منتديات عامة متعددة، وعقد العديد من اجتماعات التي تضم مجموعة صغيرة من الأفراد – بمثابة تطوير مستنير لإطار العمل (AI RMF 1.0)، بالإضافة إلى ما أجراه المعهد الوطني للمعايير والتكنولوجيا (NIST) بالتعاون مع جهات أخرى من بحوث بشأن مجال الذكاء الاصطناعي وتطويره وتقييمه. سيجري تطبيق البحوث ذات الأولوية والتوجيهات الإضافية التي من شأنها تعزيز إطار العمل هذا في "خارطة إطار عمل إدارة مخاطر الذكاء الاصطناعي" ذات الصلة، والتي يمكن أن يساهم فيها المعهد الوطني للمعايير والتقنية والمجتمع الأوسع نطاقًا.

الجزء الأول: المعلومات التأسيسية

1. تأطير المخاطر

توفر إدارة مخاطر الذكاء الاصطناعي مسارًا يهدف إلى الحد من الآثار السلبية المحتملة لأنظمة الذكاء الاصطناعي، مثل التهديدات التي تتعرض لها الحريات والحقوق المدنية، مع توفير الفرص أيضًا لتحقيق أقصى استفادة من التأثيرات الإيجابية. ويمكن أن تؤدي معالجة مخاطر الذكاء الاصطناعي والآثار السلبية المحتملة وتوثيقها وإدارتها بفعالية إلى تطوير أنظمة ذكاء اصطناعي أكثر موثوقية.

1.1 فهم المخاطر والتأثيرات والأضرار ومعالجتها

في سياق إطار العمل (AI RMF)، تشير المخاطر إلى المقياس المركب لاحتمال وقوع حدث ما وحجم أو درجة عواقب الحدث المقابل. وقد تتسم تأثيرات أنظمة الذكاء الاصطناعي أو عواقبه بالإيجابية أو السلبية أو كليهما، كما قد تؤدي إلى ظهور فرص أو تهديدات (مقتبس من: ISO 31000:2018). عند دراسة التأثير السلبي لحدث محتمل، فإن المخاطرة هي دالة تمثل 1) التأثير السلبي أو حجم الضرر الذي قد ينشأ في حالة حدوث الظرف أو الحدث، 2) احتمالية الحدوث (مقتبس من: OMB Circular A-130:2016). يمكن أن يتعرض الأفراد والجماعات والمجتمعات المحلية والمنظمات والمجتمع والبيئة وكوكب الأرض إلى هذا التأثير السلبي أو الضرر.

"تشير إدارة المخاطر إلى الأنشطة المنسقة لتوجيه ومراقبة المنظمة فيما يتعلق بالمخاطر" (المصدر: ISO 31000: 2018).

في حين تعالج عمليات إدارة المخاطر بوجه عام الآثار السلبية، يقدم إطار العمل هذا نُهُجًا تهدف إلى الحد من الآثار السلبية المتوقعة لأنظمة الذكاء الاصطناعي بجانب تحديد الفرص لتحقيق أقصى استفادة من التأثيرات الإيجابية. ويمكن أن تؤدي الإدارة الفعالة لمخاطر الأضرار المحتملة إلى تطوير أنظمة ذكاء اصطناعي أكثر موثوقية، مع إطلاق العنان للمزايا المحتملة التي تعود بالنفع على الشعوب (الأفراد والمجتمعات المحلية والمجتمع)، والمنظمات، والأنظمة/النظم البيئية. ويمكن لإدارة المخاطر أن تمكّن مطوري ومستخدمي الذكاء الاصطناعي من فهم التأثيرات ومراعاة القيود والشكوك المتأصلة في نماذجهم وأنظمتهم، التي بدورها يمكن أن تحسن الأداء العام للنظام والجدارة بالثقة واحتمال الاستعانة بتقنيات الذكاء الاصطناعي بطرق تعود بالفائدة.

يهدف إطار العمل (AI RMF) إلى معالجة المخاطر الجديدة كلما ظهرت. وتتسم هذه المرونة بأهمية خاصة، حيث لا يمكن التنبؤ بالتأثيرات بسهولة في الوقت الذي يتم فيه تطوير التطبيقات. وفي حين أن بعض مخاطر الذكاء الاصطناعي ومزاياه التي تعود بالنفع معروفة بشكل جيد، فقد يصعب تقييم الآثار السلبية ودرجة الأضرار. ويعرض الشكل التوضيحي (1) أمثلة على الأضرار المحتملة التي قد ترتبط بأنظمة الذكاء الاصطناعي.

وينبغي أن تأخذ الجهود المتعلقة بإدارة مخاطر الذكاء الاصطناعي في الاعتبار أن البشر قد يفترضون أن أنظمة الذكاء الاصطناعي تنجح – بل تحقق نجاحًا باهرًا– في التعامل مع جميع السياقات. فعلى سبيل المثال، سواء كانت وجهة النظر هذه صائبة أم خاطئة، غالبًا ما يُنظر إلى أنظمة الذكاء الاصطناعي باعتبارها أكثر موضوعية من البشر أو أنها تقدم قدرات أكبر من البرامج العامة.

إلحاق الضرر بالبشر	إلحاق الضرر بمنظمة ما	إلحاق الضرر بنظام بيئي
• الفرد: إلحاق الضرر بالحريات المدنية المكفولة للفرد أو حقوقه أو سلامته الجسدية أو النفسية أو الفرص الاقتصادية المتوفرة أمامه.	• إلحاق الضرر بالعمليات التجارية في منظمة ما.	• إلحاق الضرر بالعناصر والموارد المترابطة والمتداخلة.
• فئة سكانية/مجتمع محلي: إلحاق الضرر بفئة سكانية مثل التعرض للتمييز ضد فئة فرعية سكانية.	• إلحاق الضرر بمنظمة ما مثل التعرض إلى الانتهاكات الأمنية أو الخسارة المالية.	• إلحاق الضرر بالنظام المالي العالمي أو سلسلة التوريد أو الأنظمة المترابطة.
• مجتمعي: إلحاق الضرر بالمشاركة الديمقراطية أو سُبُل الحصول على التعليم.	• إلحاق الإضرار بسمعة منظمة ما	• إلحاق الإضرار بالموارد الطبيعية والبيئة وكوكب الأرض.

الشكل التوضيحي (1) أمثلة على الأضرار المحتملة المتعلقة بأنظمة الذكاء الاصطناعي. تساهم أنظمة الذكاء الاصطناعي الجديرة بالثقة والاستخدام المسؤول في التخفيف من المخاطر السلبية، وتحقيق المنافع التي تعود على الأفراد والمنظمات والنظم البيئية.

1.2 تحديات إدارة مخاطر الذكاء الاصطناعي

يرد فيما يلي وصف العديد من التحديات، التي يتعين مراعاتها عند إدارة المخاطر سعيًا وراء تحقيق موثوقية الذكاء الاصطناعي.

1.2.1 قياس المخاطر

يصعب قياس مخاطر الذكاء الاصطناعي أو أوجه القصور غير المعروفة بشكل جيد أو غير المفهومة بشكل كافٍ كميًا أو نوعيًا. في حين أنه لا يعني عدم القدرة على قياس مخاطر الذكاء الاصطناعي بشكل مناسب أن نظام الذكاء الاصطناعي يشكل بالضرورة مخاطر عالية أو منخفضة المستوى. وتتضمن بعض تحديات قياس المخاطر ما يلي:

المخاطر المتعلقة ببرامج وأجهزة وبيانات الطرف الثالث: قد تساهم البيانات أو الأنظمة الخاصة بالطرف الثالث في تسريع وتيرة البحث والتطوير وتسهيل الانتقال التكنولوجي، كما أنها قد تؤدي إلى تعقيد قياس المخاطر. ويمكن أن تنشأ المخاطر من بيانات أو برامج أو أجهزة الخاصة بالطرف الثالث ذاته وكيفية استخدامها. ويجدر بالذكر أن مقاييس المخاطر أو منهجياته المتبعة التي تستخدمها المنظمة المسؤولة عن تطور نظام الذكاء الاصطناعي قد لا تتوافق مع مقاييس المخاطر أو منهجياته المتبعة التي تستخدمها المنظمة المسؤولة عن نشر نظام الذكاء الاصطناعي أو العاملة في هذا المجال. بالإضافة إلى ذلك، قد لا تتسم المنظمة المسؤولة عن تطور نظام الذكاء الاصطناعي بالشفافية فيما يتعلق بمقاييس المخاطر أو المنهجيات التي تتبعها. ويمكن أن يكون قياس المخاطر وإدارتها أمرًا معقدًا بسبب كيفية استخدام العملاء لبيانات أو أنظمة الطرف الثالث أو دمجها في منتجات أو خدمات الذكاء الاصطناعي، لا سيما في غياب وجود هياكل حوكمة داخلية كافية والافتقار إلى الضمانات التقنية. وبغض النظر عن ذلك، يتعين على جميع الأطراف المعنية والجهات الفاعلة في مجال الذكاء الاصطناعي إدارة المخاطر في أنظمة الذكاء الاصطناعي التي يطورونها أو ينشرونها أو يستخدمونها باعتبارها مكونات مستقلة أو متكاملة.

تتبع المخاطر الناشئة: سيجري تعزيز جهود إدارة المخاطر في المنظمات عن طريق تحديد المخاطر الناشئة وتتبعها والنظر في تقنيات قياسها.

وقد تساعد أساليب تقييم تأثير نظام الذكاء الاصطناعي الجهات الفاعلة في مجال الذكاء الاصطناعي على فهم التأثيرات أو الأضرار المحتملة ضمن سياقات بعينها.

توفير مقاييس موثوقة: يأتي في مقدمة التحديات التي تواجه قياس مخاطر الذكاء الاصطناعي، عدم توافق الآراء في الوقت الراهن بشأن طرق قياس قوية وقابلة للتحقق فيما يتعلق بالمخاطر والموثوقية، يليها قابلية التطبيق على مختلف حالات استخدام الذكاء الاصطناعي. وتتضمن الصعوبات المحتملة عند السعي إلى قياس المخاطر أو الأضرار السلبية حقيقة مفادها أن تطوير المقاييس غالبًا ما يكون جهدًا

مؤسسيًا، وقد يعكس عن غير قصد عوامل لا علاقة لها بالتأثير الأساسي. فضلًا عن ذلك، يمكن تبسيط أساليب القياس أو تحولها، أو غياب الفروق الدقيقة أو الاعتماد عليها بطرق غير متوقعة، أو الإخفاق في تفسير الاختلافات ضمن الفئات السكانية والسياقات المتأثرة.

وتعمل أساليب قياس التأثيرات في السكان بشكل أفضل إذا أدركت مدى أهمية السياقات، وأن الأضرار قد يختلف تأثيرها في فئة سكانية أو فئة سكانية فرعية متنوعة، وأن المجتمعات المحلية أو الفئات السكانية الفرعية الأخرى التي قد تتضرر ليست دائمًا تضم مستخدمين مباشرين لنظام الذكاء الاصطناعي.

يرد فيما يلي المخاطر في مراحل مختلفة من دورة حياة الذكاء الاصطناعي: قد يؤدي قياس المخاطر في مرحلة مبكرة من دورة حياة الذكاء الاصطناعي إلى تحقيق نتائج مختلفة عن قياس المخاطر في مرحلة لاحقة، قد تكون بعض المخاطر كامنة في وقت معين أو تزداد مع تكيف أنظمة الذكاء الاصطناعي وتطورها. وفضلًا عن ذلك، يمكن أن يكون لدى الجهات الفاعلة المختلفة في مجال الذكاء الاصطناعي خلال دورة حياة الذكاء الاصطناعي وجهات نظر مختلفة عن هذه المخاطر. فعلى سبيل المثال، يمكن لمطور الذكاء الاصطناعي الذي يوفر برنامج قائم على تقنية الذكاء الاصطناعي، مثل النماذج المدربة السابقة، أن يطرح منظورًا مختلفًا عن المخاطر مقارنةً بإحدى الجهات الفاعلة في مجال الذكاء الاصطناعي المسؤولة عن نشر هذا النموذج في حالة استخدام معينة. وقد لا تدرك الجهات المسؤولة عن النشر هذه أن استخداماتها الخاصة قد تنطوي على مخاطر تختلف عن تلك التي يتصورها المطور الأصلي للنموذج. وتتقاسم جميع الجهات الفاعلة في مجال الذكاء الاصطناعي المسؤوليات المتعلقة بكلٍ من تصميم وتطوير ونشر نظام ذكاء اصطناعي جدير بالثقة وملائم لتحقيق الغرض المنشود.

المخاطر الكامنة في السياقات الواقعية: في حين أن قياس مخاطر الذكاء الاصطناعي في المختبر أو البيئة الخاضعة للرقابة قد يسفر عنها استبصارات مهمة قبل نشر نظام الذكاء الاصطناعي، وقد تختلف هذه القياسات عن المخاطر الناشئة في السياقات التشغيلية الواقعية.

أوجه الغموض: قد تساهم أنظمة الذكاء الاصطناعي التي يكتنفها الغموض في تعقيد قياس المخاطر. وقد يكون الغموض نتيجة للطبيعة الغامضة التي تتسم بها أنظمة الذكاء الاصطناعي (قابلية التفسير أو التوضيح المحدودة)، أو الافتقار إلى الشفافية أو التوثيق في تطوير نظام الذكاء الاصطناعي أو نشره، أو عدم اليقين المتأصل في أنظمة الذكاء الاصطناعي.

المرجعية البشرية: تتطلب إدارة مخاطر أنظمة الذكاء الاصطناعي الرامية إلى تعزيز النشاط البشري أو استبداله، على سبيل المثال اتخاذ القرار، شكلًا من أشكال المقاييس المرجعية للمقارنة. ويصعب تنظيم ذلك، نظرًا لأن أنظمة الذكاء الاصطناعي تنفذ مهام مختلفة – وتؤديها على نحو مختلف – مقارنةً بالبشر.

1.2.2 تحمل المخاطر

بينما يمكن الاستعانة بإطار العمل (AI RMF) في تحديد أولويات المخاطر، إلا أنه لا ينص على تحمل المخاطر. ويشير تحمل المخاطر إلى استعداد المنظمة أو الجهات الفاعلة في مجال الذكاء الاصطناعي (يرجى الاطلاع على الملحق أ) لتحمل المخاطر من أجل تحقيق أهدافها المنشودة. ويمكن أن يتأثر تحمل المخاطر بالمتطلبات القانونية أو التنظيمية (مقتبس من: ISO GUIDE 73). ويعتمد تحمل المخاطر ومستوى المخاطر المقبول للمنظمات أو المجتمع بدرجة عالية على السياق، كما يتعلق بالتطبيق والاستخدام. وقد يتأثر تحمل المخاطر بالسياسات والمعايير التي وضعتها الجهات المالكة لأنظمة الذكاء الاصطناعي أو المنظمات أو الصناعات أو المجتمعات المحلية أو واضعي السياسات. فمن المرجح أن يتغير تحمل المخاطر بمرور الوقت مع تطور أنظمة الذكاء الاصطناعي وسياساتها وقواعدها. وقد يكون لدى المنظمات المختلفة تحمّل مخاطر متنوع وفقًا لأولوياتها التنظيمية الخاصة والاعتبارات المتعلقة بالموارد.

وسيستمر تطوير المعرفة والأساليب الناشئة لتوفير معطيات أفضل بشأن إلحاق الضرر/المفاضلة بين التكلفة والفائدة المتحققة، وكذلك مناقشتها من جانب الشركات والحكومات والأوساط الأكاديمية والمجتمع المدني. وذلك إلى الحد الذي تظل فيه التحديات التي تواجه تحديد القدرة على تحمل مخاطر الذكاء الاصطناعي دون التوصل إلى حل حاسم، قد تكون هناك سياقات لا يكون فيها إطار عمل إدارة المخاطر قابلًا للتطبيق بسهولة للحد من مخاطر الذكاء الاصطناعي السلبية.

يهدف إطار العمل إلى أن يكون مرنًا وأن يعزز ممارسات المخاطر الحالية التي يجب أن تتوافق مع القوانين واللوائح والمعايير السارية. وينبغي أن تتبع المنظمات اللوائح والمبادئ التوجيهية الحالية الخاصة بمعايير المخاطر وتحمل المخاطر والاستجابة التي وضعتها لتلبية المتطلبات التنظيمية أو المهنية، أو متطلبات المجال أو الانضباط أو القطاع. قد تكون بعض القطاعات أو الصناعات قد وضعت تعريفات عن الضرر أو أنشأت متطلبات التوثيق والإبلاغ والإفصاح. وفي القطاعات، قد تعتمد إدارة المخاطر على الإرشادات الحالية المتعلقة بتطبيقات محددة وسياقات حالة الاستخدام. في حال غياب الإرشادات المحددة، يتعين على المنظمات تحديد درجة تحمل المخاطر المعقولة. بمجرد تحديد تحمل المخاطر، يمكن استخدام إطار العمل (AI RMF) في إدارة المخاطر وتوثيق عمليات إدارة المخاطر.

1.2.3 ترتيب المخاطر حسب الأولوية

يمكن أن تؤدي محاولة القضاء على المخاطر السلبية تمامًا إلى ظهور نتائج عكسية من الناحية العملية، نظرًا لأنه لا يمكن القضاء على جميع الحوادث وحالات الإخفاق. وقد تؤدي التوقعات غير الواقعية بشأن المخاطر إلى قيام المنظمات بتخصيص الموارد بطريقة تجعل تصنيف المخاطر حسب الأولوية غير فعال أو غير عملي، أو تتسبب في إهدار الموارد النادرة. في حين تساهم ثقافة إدارة المخاطر في مساعدة المنظمات على إدراك عدم تماثل جميع مخاطر الذكاء الاصطناعي، مع إمكانية تخصيص الموارد بشكل هادف. وتحدد جهود إدارة المخاطر القابلة للتنفيذ إرشادات واضحة من أجل تقييم الجدارة بالثقة لكل نظام ذكاء اصطناعي تتولى منظمة ما مسؤولية تطويره أو نشره. وعليه، يتعين تحديد أولويات السياسات ومواردها وفقًا لمستوى المخاطر المقرر والتأثير المحتمل لنظام الذكاء الاصطناعي. إنَّ النطاق الذي يمكن أن يجري فيه تخصيص نظام الذكاء الاصطناعي أو تكييفه وفقًا لسياق الاستخدام المحدد بواسطة جهة نشر نظام الذكاء الاصطناعي قد يكون عاملًا مساهمًا.

عند تطبيق إطار العمل (AI RMF)، فإن المخاطر التي تحددها المنظمة بوصفه المخاطر الأعلى بالنسبة لأنظمة الذكاء الاصطناعي ضمن سياق استخدام معين، تتطلب تحديد الأولويات الأكثر إلحاحًا وعملية إدارة المخاطر الأكثر شمولًا. وفي الحالات التي يعرض فيها نظام الذكاء الاصطناعي مستويات مخاطر سلبية غير مقبولة – مثل عندما تكون الآثار السلبية البالغة وشيكة، أو تحدث أضرار جسيمة بالفعل، أو تسبب مخاطر كارثية – يتعين التوقف عن تطوير هذا النظام ونشره بطريقة آمنة حتى يجري إدارة المخاطر بشكل كافٍ. إذا وُجد أن حالات تطوير نظام الذكاء الاصطناعي ونشره واستخدامه منخفضة المخاطر ضمن سياق معين، فقد يشير ذلك إلى احتمال انخفاض الأولوية.

قد يختلف ترتيب المخاطر حسب الأولوية بين أنظمة الذكاء الاصطناعي المصممة أو المنشورة من أجل التفاعل المباشر مع البشر مقارنةً بأنظمة الذكاء الاصطناعي غير المصممة لذلك الغرض. وقد يُطلب ترتيب أولويات أولية أعلى في السياقات التي يجري فيها تدريب نظام الذكاء الاصطناعي على مجموعات بيانات كبيرة تتألف من بيانات حساسة أو محمية، مثل معلومات التعريف الشخصية، أو حيث يكون لمخرجات أنظمة الذكاء الاصطناعي تأثير مباشر في حياة البشر أو غير مباشر. قد تتطلب أنظمة الذكاء الاصطناعي المصممة من أجل التفاعل فقط مع الأنظمة الحاسوبية والمدربة على مجموعات البيانات غير الحساسة (مثل البيانات المستمدة من البيئة المادية) تحديد أولويات أولية أقل. ومع ذلك، فإن تقييم المخاطر وتحديد أولوياتها استنادًا إلى السياق يظل أمرًا حيويًا، نظرًا لأن أنظمة الذكاء الاصطناعي غير المواجهة للتعامل مع البشر يمكن أن يكون لها تدعيات على السلامة أو آثار اجتماعية مترتبة.

المخاطر المتبقية – تُعرّف بأنها المخاطر المتبقية بعد معالجة المخاطر (المصدر: ISO GUIDE 73) – يؤثر تأثيرًا مباشرًا في المستخدمين النهائيين أو المتضررين من الأفراد والمجتمعات المحلية. سيتطلب توثيق المخاطر المتبقية من مزود النظام إبلاء الاعتبار التام إلى مخاطر نشر منتج الذكاء الاصطناعي، بجانب إبلاغ المستخدمين النهائيين بالآثار السلبية المحتملة للتفاعل مع هذا النظام.

1.2.4 التكامل التنظيمي وإدارة المخاطر

ينبغي عدم النظر في مخاطر الذكاء الاصطناعي بمعزل عن غيرها، حيث تختلف المسؤوليات والوعي باختلاف الجهات الفاعلة في مجال الذكاء الاصطناعي اعتمادًا على أدوارها المنوطة في دورة الحياة الذكاء الاصطناعي. فعلى سبيل المثال، لا يتوفر غالبًا لدي المنظمات المسؤولة عن تطور نظام الذكاء الاصطناعي المعلومات المتعلقة بآلية استخدام هذا النظام. لذا، يتعين دمج إدارة مخاطر الذكاء الاصطناعي وإدماجها في استراتيجيات إدارة مخاطر المؤسسة وعملياتها الأوسع نطاقًا. وستؤدي معالجة مخاطر الذكاء الاصطناعي إلى جانب المخاطر الحرجة الأخرى، مثل الأمن السيبراني والخصوصية، إلى تحقيق نتائج أكثر تكاملًا وكفاءات تنظيمية.

ويمكن استخدام إطار العمل (AI RMF) جنبًا إلى جنب مع التوجيهات والأُطر ذات الصلة بإدارة مخاطر أنظمة الذكاء الاصطناعي أو المخاطر المؤسسية الأوسع نطاقًا. تُعدّ بعض المخاطر المتعلقة بأنظمة الذكاء الاصطناعي منتشرة في أنواع أخرى من تطوير ونشر البرامج القائمة على تقنية الذكاء الاصطناعي. وتشتمل أمثلة المخاطر المتداخلة على ما يلي: مخاوف الخصوصية المتعلقة باستخدام البيانات الأساسية لتدريب أنظمة الذكاء الاصطناعي، والآثار المترتبة على الطاقة والبيئة ذات الصلة بمتطلبات الحوسبة كثيفة الموارد، والمخاوف الأمنية المتعلقة بسرية النظام وسلامته وتوافره وبيانات التدريب ومخرجاته، والأمن العام للبرامج والأجهزة الأساسية لأنظمة الذكاء الاصطناعي.

يتعين على المنظمات إنشاء آليات المساءلة والأدوار والمسؤوليات والثقافة وهياكل الحوافز المناسبة والحفاظ عليها، وذلك من أجل إدارة المخاطر على نحو فعال. لن يؤدي استخدام إطار العمل (AI RMF) بمفرده إلى إحداث هذه التغييرات أو توفير الحوافز المناسبة. إذ تتحقق إدارة المخاطر الفعالة عن طريق الالتزام التنظيمي على المستويات العليا، وقد تتطلب تغييرًا ثقافيًا داخل المنظمة أو قطاع الصناعة. فضلًا عن ذلك، قد تواجه المنظمات الصغيرة والمتوسطة التي تدير مخاطر الذكاء الاصطناعي، أو تنفذ إطار عمل إدارة مخاطر الذكاء الاصطناعي تحديات مختلفة عن تلك التي تواجهها المؤسسات الكبيرة، وذلك اعتمادًا على قدراتها ومواردها.

2. فئات الجمهور

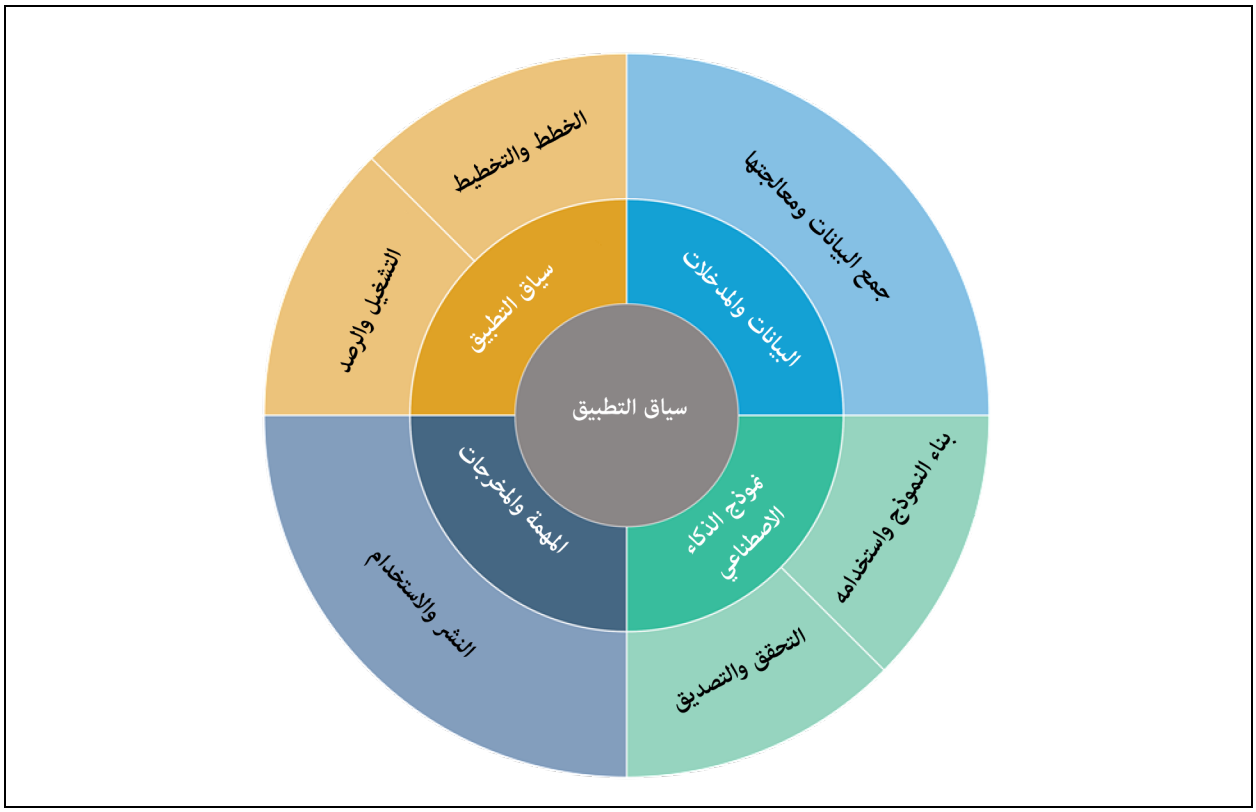
يتطلب تحديد وإدارة مخاطر الذكاء الاصطناعي والتأثيرات المحتملة – الإيجابية والسلبية على حد سواء– مجموعة واسعة النطاق من وجهات النظر والجهات الفاعلة خلال دورة حياة الذكاء الاصطناعي. ومن الناحية المثلى، ستمثل الجهات الفعالة في مجال الذكاء الاصطناعي مجموعة متنوعة من الخبرات والتجارب والخلفيات، كما سيشكلون فرقًا متنوعة على الصعيدين الديموغرافي والتخصصي. إذ يهدف إطار عمل إدارة مخاطر الذكاء الاصطناعي إلى استخدامه من جانب الجهات الفاعلة في مجال الذكاء الاصطناعي خلال دورة حياة الذكاء الاصطناعي وأبعاده.

وضعت منظمة التعاون والتنمية في الميدان الاقتصادي إطارًا لتصنيف الأنشطة المتعلقة بدورة حياة الذكاء الاصطناعي استنادًا إلى خمسة أبعاد اجتماعية وتقنية رئيسية، ولكل منها خصائص ذات صلة بسياسة الذكاء الاصطناعي وحوكمته، مثل إدارة المخاطر [OECD (2022) OECD Framework for the Classification of AI systems — OECD Digital Economy Papers]. يعرض الشكل التوضيحي (2) هذه الأبعاد الخمسة التي أجري عليها المعهد الوطني للمعايير والتكنولوجيا بعض التعديلات الطفيفة تحقيقًا لأغراض إطار العمل هذا. إذ يركز تعديل المعهد الوطني للمعايير والتكنولوجيا على أهمية ممارسات عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) طوال دورة حياة الذكاء الاصطناعي، ويعمم السياق التشغيلي لنظام الذكاء الاصطناعي.

وتمثل أبعاد الذكاء الاصطناعي المعروضة في الشكل التوضيحي (2) سياق التطبيق والبيانات والمدخلات ونموذج الذكاء الاصطناعي والمهمة المنوطة والمخرجات. إن الجهات الفاعلة في مجال الذكاء الاصطناعي المنخرطة في هذه الأبعاد والتي تتولى مسؤولية إدارة تصميم أنظمة الذكاء الاصطناعي وتطويرها ونشرها وتقييمها واستخدامها وتوجيه جهود إدارة مخاطر الذكاء الاصطناعي هي بمثابة فئة الجمهور الرئيسية لإطار عمل إدارة مخاطر الذكاء الاصطناعي (AI RMF).

يسرد الشكل التوضيحي (3) ممثلي الجهات الفاعلة في مجال الذكاء الاصطناعي عبر أبعاد دورة حياة الذكاء الاصطناعي، ثم وصفها بالتفصيل في (الملحق أ) ضمن إطار العمل (AI RMF)، حيث تتعاون جميع الجهات الفاعلة في مجال الذكاء الاصطناعي معًا لإدارة المخاطر وتحقيق أهداف الذكاء الاصطناعي الجديرة بالثقة والمسؤولة. تُدمج الجهات الفاعلة في مجال الذكاء الاصطناعي التي تتمتع بخبرة خاصة بممارسات عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) في دورة حياة الذكاء الاصطناعي، ومن المرجح أن تستفيد بشكل خاص من إطار العمل هذا. يمكن للمهام المنوطة بعمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) التي تُجري بانتظام، أن تطرح رؤى تتعلق بالمعايير أو القواعد الفنية والاجتماعية والقانونية والأخلاقية، وقد تساهم في توقع الآثار وتقييم المخاطر الناشئة وتتبعها. باعتبارها عملية منتظمة ضمن دورة حياة الذكاء الاصطناعي، تسمح ممارسات عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) بمعالجة منتصف المسار وإدارة المخاطر اللاحقة.

يمثل البُعد المتعلق بالبشر وكوكب الأرض في مركز الشكل التوضيحي (2) حقوق الإنسان والرفاهية الأوسع نطاقًا لأفراد المجتمع والكوكب. تتألف الجهات الفاعلة في مجال الذكاء الاصطناعي في هذا البُعد من فئة جمهور منفصلة لإطار العمل (AI RMF) تتولى مسؤولية إبلاغ فئة الجمهور الرئيسية. قد تتضمن هذه الجهات الفاعلة في مجال الذكاء الاصطناعي المؤسسات التجارية، ومنظمات وضع المعايير، والباحثين، والمجموعات الحقوقية، والجماعات المدافعة عن البيئة، ومنظمات المجتمع المدني، والمستخدمين النهائيين، والمتضررين من الأفراد والمجتمعات المحلية المحتملة.



الشكل التوضيحي (2): دورة الحياة نظام الذكاء الاصطناعي وأبعاده الرئيسية. معدَّل من منظمة التعاون الاقتصادي والتنمية (2022) OECD Framework for the Classification of AI systems — OECD Digital Economy Papers. توضح الدائرتان الداخليتان في الشكل (2) الأبعاد الرئيسية لأنظمة الذكاء الاصطناعي، بينما تُظهر الدائرة الخارجية مراحل دورة حياة الذكاء الاصطناعي. ومن الناحية المثلى، تبدأ جهود إدارة المخاطر بوظيفتي التخطيط والتصميم في سياق التطبيق، ويجري تنفيذها طوال دورة حياة نظام الذكاء الاصطناعي. انظر الشكل التوضيحي (3) لمعرفة ممثلي الجهات الفاعلة في مجال الذكاء الاصطناعي.

تعمل هذه الجهات الفاعلة على:

- المساعدة في توفير السياق وفهم الآثار المحتملة والفعلية،
- أن تكون مصدرًا لوضع المعايير والتوجيهات الرسمية أو شبه الرسمية فيما يتعلق بإدارة مخاطر الذكاء الاصطناعي،
- تعيين حدود لعمليات الذكاء الاصطناعي (تتضمن الحدود التقنية والمجتمعية والقانونية والأخلاقية).
- تشجيع إجراء مناقشة بشأن المفاضلات اللازمة لتحقيق التوازن بين القيم والأولويات المجتمعية المتعلقة بالحريات والحقوق المدنية والإنصاف والبيئة والكوكب والاقتصاد.

تعتمد إدارة المخاطر الناجحة على التحلي بالمسؤولية الجماعية بين الجهات الفاعلة في مجال الذكاء الاصطناعي كما هو موضح في الشكل التوضيحي (3). تتطلب وظائف إطار العمل (AI RMF)، الموضحة في القسم الخامس، وجهات نظر وتخصصات ومهن وخبرات متنوعة. حيث تساهم الفرق المتنوعة في المزيد من المشاركات أكثر انفتاحًا للأفكار والافتراضات بشأن الأغراض التكنولوجية ووظائفها – مما يجعل هذه الجوانب الضمنية أكثر وضوحًا. إذ يقدم هذا المنظور الجماعي الأوسع نطاقًا فرصًا لإلقاء الضوء على المشكلات وتحديد المخاطر القائمة والمحتملة.

الأبعاد الرئيسية	سياق التطبيق	البيانات والمدخلات	نموذج الذكاء الاصطناعي	نموذج الذكاء الاصطناعي	المهمة والمخرجات	سياق التطبيق	البشر وكوكب الأرض
مرحلة دورة الحياة	الخطط والتخطيط	جمع البيانات ومعالجتها	بناء النموذج واستخدامه	التحقق والتصديق	النشر والاستخدام	التشغيل والرصد	الاستخدام أو التأثير
TEVV	TEVV التدقيق وتقييم الأثر	TEVV التصديق على الصعيدين الداخلي والخارجي	TEVV اختبار النموذج	TEVV اختبار النموذج	TEVV التكامل واختبار الامتثال والتصديق	TEVV التدقيق وتقييم الأثر	TEVV التدقيق وتقييم الأثر
الأنشطة	توضيح وتوثيق مفهوم النظام وأهدافه والافتراضات الأساسية والسياق في ضوء المتطلبات القانونية والتنظيمية والاعتبارات الأخلاقية	جمع البيانات والتحقق منها وصقل وتوثيق البيانات الوصفية وسمات مجموعة البيانات، في ضوء الأهداف والاعتبارات القانونية والأخلاقية.	إنشاء الخوارزميات أو تحديدها، مثل نماذج التدريب.	التحقق من مخرجات النموذج والتصديق عليها وضبط قياسها وتفسيرها.	تجريب التوافق مع الأنظمة التقليدية وفحصها، والتحقق من الامتثال التنظيمي، وإدارة التغييرات التنظيمية، وتقييم تجربة المستخدم.	تشغيل نظام الذكاء الاصطناعي وتقييم توصياته وتأثيراته باستمرار (المقصودة وغير المقصودة) في ضوء الأهداف والمتطلبات القانونية والتنظيمية والاعتبارات الأخلاقية.	استخدام تكنولوجيا النظام، ورصد التأثيرات وتقييمها، والسعي إلى تخفيف التأثيرات ومناصرة الحقوق.
ممثلو الجهات الفاعلة	مشغلو النظام، والمستخدمون النهائيون، وخبراء المجال، ومصممو أنظمة الذكاء الاصطناعي، ومقيمو التأثيرات، وخبراء ممارسات عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV)، ومديرو المنتج، وخبراء الامتثال، والمدققون، وخبراء الحوكمة، والعاملون في الإدارة التنظيمية، والمديرون التنفيذيون على مستوى القيادة العليا، والمتضررون من الأفراد/المجتمعات المحلية، والمقيّمين.	علماء البيانات، ومهندسو البيانات، ومزودو البيانات، وخبراء المجال، والمحللون في القطاع الاجتماعي الثقافي، وخبراء العامل البشري، وخبراء ممارسات عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV).	مصممو النماذج ومهندسوها، وعلماء البيانات، والمطورون، وخبراء المجال بالتشاور مع القطاع الاجتماعي الثقافي، ومحللون على دراية بسياق التطبيق، وخبراء ممارسات عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV).		منظمو الأنظمة، والمطورون، ومهندسو النظام، ومهندسو البرمجة، وخبراء المجال، وموردي الطرف الثالث، والمديرون التنفيذيون على مستوى القيادة العليا بالتشاور مع خبراء العامل البشري، والمحللين في القطاع الاجتماعي الثقافي، وخبراء الحوكمة، وخبراء ممارسات عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV).	منظمو الأنظمة، والمطورون، ومهندسو النظام، ومهندسو البرمجة، وخبراء المجال، وموردي الطرف الثالث، والمديرون التنفيذيون على مستوى القيادة العليا بالتشاور مع خبراء العامل البشري، والمحللين في القطاع الاجتماعي الثقافي، وخبراء الحوكمة، وخبراء ممارسات عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV).	منظمو الأنظمة، والمطورون، ومهندسو النظام، ومهندسو البرمجة، وخبراء المجال، وموردي الطرف الثالث، والمديرون التنفيذيون على مستوى القيادة العليا بالتشاور مع خبراء العامل البشري، والمحللين في القطاع الاجتماعي الثقافي، وخبراء الحوكمة، وخبراء ممارسات عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV).

الشكل التوضيحي (3): الجهات الفاعلة في مجال الذكاء الاصطناعي عبر مراحل دورة حياة الذكاء الاصطناعي. انظر الملحق (أ) للاطلاع على الأوصاف التفصيلية للمهام المنوطة بالجهات الفاعلة في مجال الذكاء الاصطناعي، يتضمن ذلك تفاصيل بشأن المهام المتعلقة بعمليات الاختبار والتقييم والتحقق والتصديق. يُرجى ملاحظة أن الجهات الفاعلة في مجال الذكاء الاصطناعي المذكورة في بُعد نموذج الذكاء الاصطناعي (الوارد في الشكل التوضيحي 2) تُعرض على نحو منفصل بوصفها أفضل الممارسات، مع فصل الجهات المعنية بوضع النماذج واستخدامها عن تلك الجهات المعنية بالتحقق من النماذج والتصديق عليها.

3 مخاطر الذكاء الاصطناعي والجدارة بالثقة

لتصبح أنظمة الذكاء الاصطناعي جديرة بالثقة، فإنها بحاجة إلى الاستجابة لتلبية العديد من المعايير ذات القيمة للأطراف المعنية. وقد تساهم الأساليب التي تعزز موثوقية الذكاء الاصطناعي في الحد من المخاطر السلبية المتعلقة بالذكاء الاصطناعي. ويوضح إطار العمل هذا **السمات** التالية للذكاء الاصطناعي الجدير بالثقة، ويقدم إرشادات لمعالجتها. تشمل سمات أنظمة الذكاء الاصطناعي الجديرة بالثقة: التي تتضمن أن تكون الأنظمة **صالحة وموثوقة وآمنة ومأمونة ومرنة ومسؤولة وشفافة وقابلة للتفسير والتوضيح وتتسم الخصوصية المحسنة والإنصاف مع إدارة تحيزاتها الضارة**. ويتطلب إنشاء ذكاء اصطناعي جدير بالثقة موازنة كل من هذه السمات استنادًا إلى سياق استخدام نظام الذكاء الاصطناعي. في حين أن جميع السمات هي سمات نظام اجتماعي تقني، فإن المساءلة والشفافية ترتبطان ارتباطًا وثيقًا بالعمليات والأنشطة الداخلية لنظام الذكاء الاصطناعي ومحيطه الخارجي. يمكن أن يؤدي إهمال هذه السمات إلى زيادة احتمالية وحجم النتائج السلبية.

الشكل التوضيحي (4): سمات أنظمة الذكاء الاصطناعي الجديرة بالثقة. تُعدّ الصلاحية والموثوقية من الشروط الضرورية للجدارة بالثقة، كما تظهران باعتبارهما أساسًا لغيرهما من سمات الجدارة بالثقة. وتُعرض المساءلة والشفافية في مربع عمودي، نظرًا لأنهما يتعلقان بجميع السمات الأخرى.

ترتبط سمات الجدارة بالثقة (المبينة في الشكل التوضيحي 4) ارتباطًا وثيقًا بالسلوك الاجتماعي التنظيمي، ومجموعات البيانات المستخدمة بواسطة أنظمة الذكاء الاصطناعي، واختيار نماذج وخوارزميات الذكاء الاصطناعي والقرارات التي يتخذها من يبنونها، والتفاعلات مع البشر الذين يقدمون رؤى والمشرفين على مثل هذه الأنظمة. يتعين استخدام الحكم البشري عند اتخاذ قرار بشأن المقاييس المحددة المتعلقة بسمات موثوقية الذكاء الاصطناعي والقيم الدقيقة لتلك المقاييس.

إنَّ معالجة سمات موثوقية الذكاء الاصطناعي بشكل فردي لن يضمن موثوقية نظام الذكاء الاصطناعي، فعادةً ما ينطوي على المفاضلات، ونادرًا ما تُطبق جميع السمات في كل سياق، وذلك نظرًا لأن بعض السياقات ستكون أكثر أهمية أو أقل تبعًا لاختلاف المواقف. وفي نهاية المطاف، تُعدّ الجدارة بالثقة أو الموثوقية مفهومًا اجتماعيًا يمتد عبر طيف واسع النطاق ولا يكون إلا بقوة أضعف سماته.

عند إدارة مخاطر الذكاء الاصطناعي، يمكن أن تواجه المنظمات قرارات صعبة فيما يتعلق بموازنة هذه السمات. فعلى سبيل المثال، في بعض السيناريوهات، قد تظهر مفاضلات بين تحسين قابلية التفسير وتحقيق الخصوصية. وفي حالات أخرى، قد تواجه المنظمات مفاضلة بين الدقة التنبؤية وقابلية التفسير. أو في ظل ظروف معينة مثل تباين البيانات، يمكن أن تؤدي تقنيات تحسين الخصوصية إلى فقدان الدقة، مما يؤثر في القرارات المتعلقة بالانصاف والقيم الأخرى في مجالات معينة. إذ يتطلب التعامل مع المفاضلات مراعاة سياق اتخاذ القرارات. فقد تسلط هذه التحليلات الضوء على وجود ومدى المفاضلات بين مختلف المقاييس، ولكنها لا تجيب بالضرورة على الأسئلة المتعلقة بكيفية التعامل مع المفاضلة، حيث تعتمد على القيم الموجودة في السياق ذي الصلة، ويتعين معالجتها بطريقة شفافة ومبررة بشكل مناسب.

ثمة أساليب متعددة لتعزيز الوعي السياقي في دورة حياة الذكاء الاصطناعي. فعلى سبيل المثال، يمكن للخبراء المتخصصين المساعدة في تقييم نتائج عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV)، وكذلك العمل مع فرق المنتج والنشر من أجل مواءمة معايير (TEVV) مع المتطلبات وظروف النشر. عند توفير الموارد بشكل مناسب، يمكن أن تؤدي زيادة اتساع المدخلات وتنوعها من الأطراف المعنية والجهات الفاعلة ذات الصلة بمجال الذكاء الاصطناعي، خلال دورة حياة الذكاء الاصطناعي، إلى تعزيز فرص تقديم التقييمات

الحساسة للسياق وتحديد مزايا نظام الذكاء الاصطناعي وتأثيراته الإيجابية. وقد تزيد هذه الممارسات من احتمالية إدارة المخاطر الناشئة في السياقات الاجتماعية على نحو مناسب.

يعتمد فهم سمات الجدارة بالثقة ومعالجتها على الدور الخاص المنوط بالجهات الفاعلة في مجال الذكاء الاصطناعي خلال دورة حياة الذكاء الاصطناعي. بالنسبة لأي نظام ذكاء اصطناعي محدد، قد يكون لمصمم أو مطور الذكاء الاصطناعي تصور مختلف فيما يتعلق بالسمات عن تصور الجهة المسؤولة عن النشر.

تؤثر سمات الجدارة بالثقة الموضحة في هذه الوثيقةعلى بعضها البعض. هناك أنظمة شديدة الأمان ولكنها غير منصفة، وأنظمة تتسم بالدقة ولكنها لا تتمتع بالشفافية وغير قابلة للتفسير، وفي الجانب الأخرى ثمة أنظمة لا تتسم بالدقة ولكنها آمنة ومحسّنة للخصوصية وتتمتع بالشفافية، إلا أن كل تلك الأنظمة غير مرغوب فيها. في حين يتطلب النهج الشامل لإدارة المخاطر موازنة المفاضلات بين سمات الجدارة بالثقة. تقع على عاتق جميع الجهات الفاعلة في مجال الذكاء الاصطناعي مسؤولية مشتركة لتحديد ما إذا كانت تقنية الذكاء الاصطناعي أداة مناسبة أو ضرورية لسياق أو تحقيق غرض معين، وكيفية استخدامها على نحو مسؤول. إذ يتعين أن يستند قرار تكليف أو نشر نظام الذكاء الاصطناعي إلى إجراء تقييم سياقي بشأن سمات الجدارة بالثقة والمخاطر والتأثيرات والتكاليف والمزايا النسبية، وأن تكون مجموعة واسعة من الأطراف المعنية على دراية بها.

3.1 الصلاحية والموثوقية

يُقصد بالصلاحية "التأكيد، من خلال توفير دليل موضوعي، على أن متطلبات الاستخدام المحدد أو التطبيق المقصود قد تمت تلبيتها" (المصدر: ISO 9000:2015). ويؤدي نشر أنظمة الذكاء الاصطناعي غير الدقيقة أو غير الموثوق بها أو المعممة بشكل ضعيف إلى البيانات والسياقات، بخلاف تدريبها، إلى زيادة المخاطر السلبية للذكاء الاصطناعي وتقليل الجدارة بالثقة.

وتُعرف الجدارة بالثقة في المقاس ذاته على أنها "قدرة عنصر ما على الأداء كما هو مطلوب، دون التعرض للإخفاق، خلال فترة زمنية محددة وفي ظل ظروف معينة" (المصدر: ISO/IEC TS 5723:2022). الموثوقية هي هدف الإتقان الشامل لتشغيل نظام الذكاء الاصطناعي في ظل ظروف الاستخدام المتوقع ولفترة زمنية معينة، يتضمن ذلك العمر الافتراضي للنظام بالكامل.

تساهم الدقة والاعتمادية في صلاحية أنظمة الذكاء الاصطناعي وموثوقيتها، ويمكن أن تتعارض مع بعضها البعض في أنظمة الذكاء الاصطناعي.

تُعرف الدقة وفقًا للتعريف الوارد في (ISO / IEC TS 5723: 2022) على أنها "تقارب نتائج الملاحظات أو الحسابات أو التقديرات مع القيم الحقيقية أو القيم المقبولة على أنها قيم صحيحة". ويتعين أن تأخذ تدابير الدقة في الاعتبار المقاييس التي تتمحور حول الحساب (مثل المعدلات الإيجابية الخاطئة والسلبية الخاطئة)، والفرق التي تجمع بين العامل البشري والذكاء الاصطناعي، وإثبات الصلاحية الخارجية (قابلة للتعميم خارج شروط التدريب). ويلزم إقران قياسات الدقة بمجموعات اختبار محددة بوضوح وواقعية – والتي تمثل ظروف الاستخدام المتوقع – وتفاصيل حول منهجية الاختبار، ويتعين أن يتم تضمين ذلك في الوثائق ذات الصلة. وقد تشمل قياسات الدقة تفصيل النتائج لمختلف قطاعات البيانات.

وتُعرّف الاعتمادية أو القابلية للتعميم بأنها "قدرة النظام على الحفاظ على مستوى أدائه في ظل تعرضه إلى مجموعة متنوعة من الظروف" (المصدر: ISO/IEC TS 5723:2022). وتعتبر الاعتمادية هدفًا فيما يتعلق بوظائف النظام المناسبة ضمن مجموعة واسعة من الظروف والأوضاع، يتضمن ذلك استخدامات أنظمة الذكاء الاصطناعي التي لم تكن متوقعة في بداية الأمر. لا تتطلب الاعتمادية فقط أن يعمل النظام تمامًا كما يفعل في ظل الاستخدامات المتوقعة، ولكن أيضًا يجب أن يعمل بطرق تقلل من الأضرار المحتملة على الأفراد، إذا كان يعمل في بيئة غير متوقعة.

غالبًا ما يتم تقييم الصلاحية والموثوقية لأنظمة الذكاء الاصطناعي المنشورة عن طريق عمليتي الاختبار أو الرصد باستمرار لضمان عمل النظام على النحو المنشود. ويساهم قياس الصلاحية والدقة والاعتمادية والموثوقية في تحقيق الجدارة بالثقة، كما يتعين أن يأخذ في الاعتبار أن أنواعًا معينة من حالات الإخفاق قد تسبب ضررًا أكبر. وينبغي أن تعطي جهود إدارة مخاطر الذكاء الاصطناعي الأولوية لتقليل الآثار السلبية المحتملة، وقد يتطلب الأمر تضمين تدخل بشري في الحالات التي يتعذر على نظام الذكاء الاصطناعي اكتشاف الأخطاء أو تصحيحها.

3.2 الأمان

يجب ألا تعمل أنظمة الذكاء الاصطناعي "في ظل ظروف محددة تعرض فيها حياة الإنسان أو صحته أو ممتلكاته أو البيئة إلى الخطر" (المصدر: ISO/IEC TS 5723:2022). جرى تحسين التشغيل الآمن لأنظمة الذكاء الاصطناعي من خلال:

- ممارسات التصميم والتطوير والنشر المسؤولة،
- معلومات واضحة للجهات المسؤولة عن النشر بشأن الاستخدام المسؤول للنظام،
- اتخاذ القرار المسؤول من جانب الجهات المسؤولة عن النشر والمستخدمين النهائيين،
- تفسيرات وتوثيق للمخاطر استنادًا إلى الأدلة التجريبية للحوادث.

قد تتطلب الأنواع المختلفة من مخاطر السلامة اتباع أساليب مخصصة لإدارة مخاطر الذكاء الاصطناعي وفقًا للسياق وشدة المخاطر المحتملة المقدمة. تتطلب مخاطر السلامة التي تشكل خطرًا محتملًا للإصابة الخطيرة أو الوفاة تحديد الأولويات الأكثر إلحاحًا وعملية إدارة المخاطر الأكثر شمولًا.

وقد يؤدي استخدام اعتبارات السلامة خلال دورة حياة نظام الذكاء الاصطناعي، والبدء في التخطيط والتصميم في أقرب وقت ممكن إلى منع الإخفاقات أو الظروف التي تجعل النظام خطيرًا. وغالبًا ما تتعلق الأساليب العملية الأخرى لسلامة الذكاء الاصطناعي بالمحاكاة الصارمة والاختبار في المجال، والمراقبة في الوقت الفعلي، والقدرة على إيقاف أو تعديل أو التدخل البشري في الأنظمة التي تنحرف عن الوظائف المنشودة أو المتوقعة.

يتعين أن تأخذ أساليب إدارة مخاطر السلامة فيما يتعلق بالذكاء الاصطناعي إشارات من الجهود والإرشادات الخاصة بالسلامة في مجالات مثل النقل والرعاية الصحية، وأن تتماشى مع الإرشادات أو المعايير الحالية الخاصة بالقطاع المعني أو التطبيق.

3.3 آمنة ومرنة

نصف أنظمة الذكاء الاصطناعي والأنظمة البيئية التي يتم نشرها فيها، بأنها مرنة في حال قدرتها على تحمل الأحداث السلبية غير المتوقعة أو التغييرات غير المتوقعة في بيئتها أو استخدامها - أو إذا كان بإمكانها الحفاظ على وظائفها وهيكلها في مواجهة الظروف الداخلية والتغيير الخارجي والتحلل بأمان عند الضرورة (مقتبس من: ISO/IEC TS 5723:2022). تتعلق المخاوف الأمنية الشائعة بأمثلة عدائية، مثل تسميم البيانات، وتسرب النماذج أو بيانات التدريب أو الملكية الفكرية الأخرى من خلال نقاط النهائية لنظام الذكاء الاصطناعي. ويمكن القول بأن أنظمة الذكاء الاصطناعي آمنة في حال قدرتها على الحفاظ على السرية والنزاهة والتوافر من خلال اتباع آليات الحماية تمنع الوصول والاستخدام غير المصرح بهما. إنَّ المبادئ التوجيهية الواردة في إطار عمل الأمن السيبراني NIST، وإطار عمل إدارة المخاطر، هي من بين المبادئ التي يمكن تطبيقها في هذا السياق.

يُعد الأمن والمرونة سمات مترابطة ولكنها متباينة. في حين أن المرونة هي القدرة على العودة إلى الوظيفة المعتادة بعد وقوع حدث ضار غير متوقع، فإن الأمن يضمن مفهوم المرونة ولكنه يشمل أيضًا بروتوكولات لتجنب الهجمات أو الحماية منها أو الاستجابة لها أو التعافي منها. وترتبط المرونة بالقوة وتتجاوز مصدر البيانات لتشمل الاستخدام غير المتوقع أو العدائي (انتهاك أو إساءة استخدام) النموذج أو البيانات.

3.4 المساءلة والشفافية

يعتمد الذكاء الاصطناعي الجدير بالثقة على المساءلة، في حين أن المساءلة تفترض الشفافية. إذ تعكس الشفافية مدى توفر المعلومات بشأن نظام الذكاء الاصطناعي ومخرجاته فيما يتعلق بالأفراد الذين يتفاعلون مع مثل هذا النظام – بغض النظر عما إذا كانوا يدركون أنهم يفعلون ذلك. وتوفر الشفافية الهادفة الوصول إلى مستويات مناسبة من المعلومات وفقًا لمرحلة دورة حياة الذكاء الاصطناعي، والتي تكون مصممة خصيصًا لتتوافق مع دور أو معرفة الجهات الفاعلة في مجال الذكاء الاصطناعي أو الأفراد الذين يتفاعلون مع نظام الذكاء الاصطناعي أو يستخدمونه. ومن خلال تعزيز مستويات أعلى من الفهم، تزيد الشفافية مستوى الثقة في نظام الذكاء الاصطناعي.

يمتد نطاق هذه الخاصية بدءًا من قرارات التصميم حتى بيانات التدريب وصولًا إلى تدريب النموذج، وهيكل النموذج، وحالات الاستخدام المقصودة، وكيف ومتى تم اتخاذ قرارات النشر أو ما بعد النشر، أو المستخدم النهائي وبواسطة مَن. وغالبًا ما تكون الشفافية ضرورية

للإنصاف القابل للتنفيذ فيما يتعلق بمخرجات نظام الذكاء الاصطناعي غير الصائبة أو التي تؤدي بطريقة أخرى إلى وقوع تأثيرات سلبية. ويتعين أن تأخذ الشفافية في الاعتبار التفاعل بين الإنسان والذكاء الاصطناعي: على سبيل المثال، كيف يتم إخطار عامل بشري أو مستخدم عند اكتشاف نتيجة عكسية محتملة أو فعلية ناجمة عن نظام الذكاء الاصطناعي.

لا يُعتبر النظام الذي يتسم بالشفافية بالضرورة نظامًا دقيقًا أو محسّنًا للخصوصية أو آمنًا أو منصفًا. ومع ذلك، فإنه يصعب تحديد ما إذا كان النظام الغامض يمتلك مثل هذه السمات، والقيام بذلك بمرور الوقت مع تطور الأنظمة المعقدة.

وينبغي مراعاة دور الجهات الفاعلة في مجال الذكاء الاصطناعي عند السعي إلى المساءلة عن نتائج أنظمة الذكاء الاصطناعي. حيث تختلف العلاقة بين المخاطر والمساءلة المرتبطة بالذكاء الاصطناعي والأنظمة التكنولوجية على نطاق أوسع عبر السياقات الثقافية والقانونية والقطاعية والمجتمعية. وعندما تكون العواقب وخيمة، كما هو الحال عندما تكون الحياة والحرية على المحك، يتعين على مطوري الذكاء الاصطناعي والجهات القائمة على نشره النظر في تعديل ممارسات الشفافية والمساءلة بشكل متناسب واستباقي. ويمكن أن يساعد الحفاظ على الممارسات التنظيمية وهياكل الحوكمة للحد من الضرر، مثل إدارة المخاطر، في أن يؤدي إلى إنشاء أنظمة أكثر مساءلة.

ويتعين أن تأخذ تدابير تعزيز الشفافية والمساءلة في الاعتبار تأثير هذه الجهود في الجهة المسؤولة عن التنفيذ، بما في ذلك مستوى الموارد اللازمة وكذلك الحاجة إلى حماية المعلومات الخاضعة لحقوق الملكية.

قد يساهم الحفاظ على مصدر بيانات التدريب، بجانب دعم إسناد قرارات نظام الذكاء الاصطناعي إلى المجموعات الفرعية من بيانات التدريب في تحقيق الشفافية والمساءلة على حد سواء. وقد تخضع بيانات التدريب أيضًا لحقوق الطبع والنشر، كما يجب أن تمتثل قوانين حقوق الملكية الفكرية المعمول بها.

وتزامنًا مع استمرار تطور أدوات الشفافية المتعلقة بأنظمة الذكاء الاصطناعي والوثائق ذات الصلة، يتم تشجيع مطوري أنظمة الذكاء الاصطناعي على اختبار أنواع مختلفة من أدوات الشفافية بالتعاون مع الجهات المسؤولة عن نشر الذكاء الاصطناعي، وذلك لضمان استخدام أنظمة الذكاء الاصطناعي على النحو المنشود.

3.5 القابلية للتفسير والتوضيح

تشير القابلية للتفسير إلى تمثيل الآليات الكامنة وراء تشغيل أنظمة الذكاء الاصطناعي، بينما تشير القابلية للتوضيح إلى معنى مخرجات أنظمة الذكاء الاصطناعي في سياق الأغراض الوظيفية المصممة لها. وتساعد القابلية للتفسير والتوضيح معًا الجهات المسؤولة عن تشغيل نظام الذكاء الاصطناعي والإشراف عليه، وكذلك مستخدمي نظام الذكاء الاصطناعي، وذلك من أجل اكتساب رؤى أعمق بشأن وظائف النظام وموثوقيته، بما في ذلك مخرجاته. ويكمن الافتراض الأساسي في أن تصورات المخاطر السلبية تنبع من عدم القدرة على فهم أو تأطير مخرجات النظام على نحو مناسب. في حين تقدم أنظمة الذكاء الاصطناعي القابلة للتفسير والتوضيح معلومات من شأنها مساعدة المستخدمين النهائيين على فهم الأغراض والتأثير المحتمل لنظام الذكاء الاصطناعي.

يمكن إدارة المخاطر الناجمة عن نقص القابلية للتفسير عن طريق وصف كيفية عمل أنظمة الذكاء الاصطناعي من خلال الأوصاف المصممة للاختلافات الفردية مثل دور المستخدم ومستوى معرفته ومهارته. ويمكن تصحيح أخطاء الأنظمة القابلة للتفسير ورصدها بسهولة أكبر، كما أنها تصلح لتوثيق وتدقيق وحوكمة أكثر شمولًا.

غالبًا ما يمكن معالجة المخاطر المتعلقة بقابلية التفسير من خلال وصف السبب الذي يجعل نظام الذكاء الاصطناعي يقدم تنبؤًا أو توصية معينة. (انظر "المبادئ الأربعة للذكاء الاصطناعي القابلة للتفسير" و"الأسس النفسية للتفسير والتوضيح في الذكاء الاصطناعي" في الربط من هنا)

تُعد الشفافية وقابلية التفسير والتوضيح سمات مميزة تدعم بعضها البعض، ويمكن للشفافية الإجابة على سؤال "ماذا حدث" في النظام؟ كما يمكن أن تجيب قابلية التفسير على سؤال "كيف" تم اتخاذ القرار في النظام؟ بينما يمكن أن تجيب قابلية التوضيح على سؤال "لماذا" اتخذ النظام قرارًا، وما معناه، أو سياقه بالنسبة إلى المستخدم؟

3.6 الخصوصية المحسنة

تشير الخصوصية بشكل عام إلى القواعد والممارسات التي تساهم في حماية استقلالية الإنسان وهويته والحفاظ على كرامته. تتناول هذه القواعد والممارسات عادةً التحرر من التطفل أو تقييد المراقبة أو وكالة الأفراد للموافقة على الكشف عن جوانب من هوياتهم الشخصية أو التحكم فيها (مثل الجسد والبيانات والسمعة). (انظر للاطلاع The NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management.)

يتعين أن توجه قيم الخصوصية مثل عدم الكشف عن الهوية والسرية والتحكم بشكل عام في خيارات تصميم نظام الذكاء الاصطناعي وتطويره ونشره. قد تؤثر المخاطر المتعلقة بالخصوصية في الأمان والتحيز والشفافية، وتأتي مع المفاضلات هذه السمات الأخرى.على غرار السلامة والأمن، قد تعزز الخصائص التقنية المحددة لنظام الذكاء الاصطناعي الخصوصية أو تعمل على تقليلها. كما يمكن أن تشكل أنظمة الذكاء الاصطناعي مخاطر جديدة على الخصوصية من خلال السماح بالاستدلال بتحديد هوية الأفراد أو المعلومات الخاصة بالأفراد سابقًا.

وقد تساهم تقنيات الخصوصية المحسنة ("PETs") المتعلقة بالذكاء الاصطناعي، بالإضافة إلى اتباع أساليب تقليل البيانات مثل إلغاء تحديد الهوية والتجميع لمخرجات نماذج معينة، في دعم التصميم لأنظمة الذكاء الاصطناعي المحسّنة للخصوصية. وفي ظل ظروف معينة مثل تباين البيانات، يمكن أن تؤدي تقنيات الخصوصية المحسنة إلى فقدان الدقة، مما يؤثر في اتخاذ القرارات المتعلقة بالإنصاف والقيم الأخرى في مجالات معينة.

3.7 الإنصاف- مع إدارة التحيزات الضارة

يتضمن مفهوم الإنصاف في الذكاء الاصطناعي المخاوف بشأن المساواة والإنصاف من خلال معالجة قضايا مثل التحيزات الضارة والتعرض إلى التمييز. ويمكن أن تكون معايير الإنصاف معقدة ويصعب تحديدها، نظرًا لأن تصورات الإنصاف تختلف بين الثقافات وقد تتغير اعتمادًا على التطبيق. وسيجري تعزيز جهود إدارة المخاطر في المنظمات من خلال التعرف على هذه الاختلافات ومراعاتها ولا تُعدّ الأنظمة التي يتم فيها الحد من التحيزات الضارة منصفة بالضرورة. فعلى سبيل المثال، لا تزال الأنظمة التي تكون فيها التوقعات متوازنة إلى حد ما عبر المجموعات الديموغرافية غير متوفرة لذوي الهمم أو متأثرة بالفجوة الرقمية أو قد تؤدي إلى تفاقم التباينات القائمة أو التحيزات المنهجية.

يُعد مفهوم التحيز أوسع نطاقًا من التوازن الديموغرافي وتمثيل البيانات. فقد حدد المعهد القومي للمعايير والتكنولوجيا (NIST) ثلاث فئات رئيسية من انحياز الذكاء الاصطناعي اللازم مراعاتها وإدارتها: المنهجية، والحسابية والإحصائية، والإنسانية المعرفية. وقد يحدث ذلك في حال عدم وجود التحيز أو الانحياز أو قصد التمييز. يمكن أن يوجد التحيز النظامي في مجموعات بيانات الذكاء الاصطناعي، والمعايير والممارسات والعمليات التنظيمية خلال دورة حياة الذكاء الاصطناعي، والمجتمع الأوسع نطاقًا الذي يستخدم أنظمة الذكاء الاصطناعي. وقد توجد التحيزات الحسابية والإحصائية في مجموعات بيانات الذكاء الاصطناعي والعمليات الحسابية، وغالبًا ما تنجم عن أخطاء منهجية ناتجة عن وجود عينات غير تمثيلية. وترتبط التحيزات المعرفية البشرية بآلية إدراك الفرد أو المجموعة لمعلومات نظام الذكاء الاصطناعي لاتخاذ قرار أو استكمال المعلومات الناقصة، أو معرفة كيفية تفكير البشر في أغراض نظام الذكاء الاصطناعي ووظائفه. بينما التحيزات المعرفية البشرية منتشرة في كل مكان في عمليات اتخاذ القرار خلال دورة حياة الذكاء الاصطناعي واستخدام النظام، بما في ذلك تصميم الذكاء الاصطناعي وتنفيذه وتشغيله وصيانته.

يظهر التحيز في صور عديدة، ويمكن أن يصبح متأصلًا في الأنظمة الآلية التي تساعد في اتخاذ القرارات بشأن حياتنا. وفي حين أن التحيز ليس دائمًا ظاهرة سلبية، فمن المحتمل أن تزيد أنظمة الذكاء الاصطناعي من سرعة وحجم التحيزات، وتؤدي إلى تضخيم واستمرار الأضرار التي تلحق بالأفراد والجماعات والمجتمعات المحلية والمنظمات والمجتمع بأكمله. إذ يرتبط التحيز ارتباطًا وثيقًا بمفاهيم الشفافية والإنصاف في المجتمع. (لمعرفة مزيد من المعلومات بشأن التحيز، بما في ذلك الفئات الثلاثة، يُرجى الاطلاع على المنشور الخاص بالمعهد الوطني للمعايير والتكنولوجيا رقم (1270)، Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.)

4. فعالية إطار عمل إدارة مخاطر الذكاء الاصطناعي

ستشكل تقييمات مدى فعالية إطار عمل إدارة مخاطر الذكاء الاصطناعي، بما في ذلك طرق قياس التحسينات النهائية في موثوقية أنظمة الذكاء الاصطناعي، جزءًا من أنشطة المعهد الوطني للمعايير والتكنولوجيا في المستقبل، بالتنسيق مع مجتمع الذكاء الاصطناعي.

كما أن المؤسسات وغيرهم من مستخدمي إطار العمل مدعون إلى إجراء تقييم دوري لتقييم ما إذا كان إطار عمل إدارة مخاطر الذكاء الاصطناعي قد حسّن من قدرتهم على إدارة مخاطر الذكاء الاصطناعي، بما في ذلك، على سبيل المثال لا الحصر، سياساتهم وعملياتهم وممارساتهم وخطط التنفيذ والمؤشرات والقياسات والنتائج المتوقعة. يعتزم المعهد الوطني للمعايير والتكنولوجيا العمل بالتعاون مع الآخرين لتطوير المقاييس والمنهجيات والأهداف لتقييم مدى فعالية إطار عمل إدارة مخاطر الذكاء الاصطناعي، ومشاركة النتائج والمعلومات الداعمة على نطاق واسع. من المنتظر أن يستفيد مستخدمو إطار العمل من:

- تعزيز العمليات من أجل حوكمة مخاطر الذكاء الاصطناعي وتخطيطها وقياسها وإدارتها، وتوثيق النتائج بوضوح.
- تحسين مستوى الوعي بالعلاقات والمفاضلات بين سمات الموثوقية والنُهج الاجتماعية-التقنية ومخاطر الذكاء الاصطناعي.
- العمليات الواضحة لاتباع نظام المضي/عدم المضي عند اتخاذ قرارات التكليف والنشر.
- وضع السياسات والعمليات والممارسات والإجراءات لتحسين الجهود المبذولة بشأن المساءلة التنظيمية ذات الصلة بمخاطر أنظمة الذكاء الاصطناعي.
- تعزيز الثقافة التنظيمية التي تمنح الأولوية لتحديد وإدارة مخاطر أنظمة الذكاء الاصطناعي والتأثيرات المحتملة في الأفراد والمجتمعات والمؤسسات والمجتمع.
- تبادل المعلومات على نحو أفضل داخل المؤسسات وفيما بينها حول المخاطر وعمليات صنع القرار والمسؤوليات والمآزق الشائعة، وممارسات عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV)، ونُهُج التحسين المستمر.
- تعزيز المعارف السياقية لإذكاء الوعي بالمخاطر النهائية.
- تعزيز المشاركة مع الأطراف المعنية والجهات الفاعلة المعنية في مجال الذكاء الاصطناعي.
- زيادة القدرات على تحقيق عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) لأنظمة الذكاء الاصطناعي والمخاطر ذات الصلة بها.

الجزء الثاني: جوهر إطار العمل وملفات التعريف

5. جوهر إطار عمل إدارة مخاطر الذكاء الاصطناعي

يقدم جوهر إطار عمل إدارة مخاطر الذكاء الاصطناعي النتائج والإجراءات التي من شأنها تمكين الحوار والفهم والأنشطة اللازمة لإدارة مخاطر الذكاء الاصطناعي وتطوير أنظمة ذكاء اصطناعي جديرة بالثقة على نحو مسؤول. وكما هو موضح في الشكل التوضيحي (5)، يتكون جوهر إطار العمل من أربع وظائف: الحوكمة، والتخطيط، والقياس، والإدارة. تُقسم كل من هذه الوظائف عالية المستوى إلى فئات رئيسية وفئات فرعية. وتنقسم الفئات الرئيسية والفئات الفرعية إلى إجراءات ونتائج محددة. ولا تشكل الإجراءات قائمة مرجعية في حد ذاتها، كما أنها لا تُعد بالضرورة مجموعة خطوات مرتبة.

الشكل التوضيحي (5) تعمل الوظائف على تنظيم أنشطة إدارة مخاطر الذكاء الاصطناعي وفق أعلى مستوياتها، وذلك بغرض حوكمة مخاطر الذكاء الاصطناعي وتخطيطها وقياسها وإدارتها. وقد صُممت الحوكمة لتكون وظيفة شاملة الأبعاد بغرض إرشاد الوظائف الثلاث الأخرى ودمجها فيها.

يتعين أن تكون عملية إدارة المخاطر مستمرة ومنضبطة من حيث الوقت، مع ضرورة إجراؤها عبر أبعاد دورة حياة نظام الذكاء الاصطناعي. كما يجب تنفيذ وظائف جوهر إطار عمل إدارة مخاطر الذكاء الاصطناعي على نحو يعكس وجهات النظر المتنوعة ومتعددة التخصصات، بما في ذلك وجهات نظر الجهات الفاعلة في مجال الذكاء الاصطناعي خارج المؤسسة. إذ يسهم وجود فريق متنوع في زيادة المشاركة المنفتحة للأفكار والافتراضات حول أغراض التكنولوجيا ووظائفها التي يُجرى تصميمها أو تطويرها أو نشرها أو تقييمها، والتي من شأنها إيجاد فرص لإظهار المشكلات وتحديد المخاطر الحالية والناشئة.

يتوفر مورد مصاحب لإطار عمل إدارة مخاطر الذكاء الاصطناعي على شبكة الإنترنت؛ وهو دليل إطار عمل إدارة مخاطر الذكاء الاصطناعي الصادر عن المعهد الوطني للمعايير والتكنولوجيا، والذي بإمكانه مساعدة المؤسسات استكشاف إطار عمل إدارة مخاطر الذكاء الاصطناعي وتحقيق نتائجه من خلال إجراءات العمل التكتيكية المقترحة التي يمكن للمؤسسات تطبيقها في إطار سياقاتها الخاصة. وعلى غرار إطار عمل إدارة مخاطر الذكاء الاصطناعي، فإن دليل التشغيل هذا اختياري، ويمكن للمؤسسات الاستفادة من المقترحات وفقًا لاحتياجاتها واهتماماتها. يمكن لمستخدمي هذا الدليل وضع مبادئ توجيهية مصممة وفق الحاجة ومختارة من المواد المقترحة لاستخدامهم الخاص والإسهام باقتراحاتهم لمشاركتها مع المجتمع الأوسع نطاقًا. إلى جانب إطار عمل إدارة مخاطر الذكاء الاصطناعي، يعد الدليل جزءًا من مركز موارد المعهد الوطني للمعايير والتكنولوجيا حول الذكاء الاصطناعي الجدير بالثقة والمسؤولية.

قد يطبق مستخدمو إطار العمل هذه الوظائف وفق ما يناسب احتياجاتهم لإدارة مخاطر الذكاء الاصطناعي استنادًا إلى مواردهم وقدراتهم. قد تختار بعض المؤسسات الانتقاء من بين الفئات الرئيسية والفئات الفرعية؛ قد يختار البعض الآخر جميع الفئات الرئيسية والفئات الفرعية وقد تتوفر لديه القدرة على تطبيقها. وعلى افتراض أن هناك هيكل حوكمة موضع التنفيذ، فإنه يمكن أداء الوظائف بأي ترتيب عبر دورة حياة الذكاء الاصطناعي على النحو الذي يظن مستخدم إطار العمل من شأنه أن يضيف قيمة. بعد صياغة النتائج في وظيفة الحوكمة، سيبدأ معظم مستخدمي إطار عمل إدارة مخاطر الذكاء الاصطناعي أولًا بوظيفة التخطيط وثم الاستمرار في وظيفتي القياس أو الإدارة. ومع ذلك، في حال دمج المستخدمون الوظائف، فيتعين أن تكون العملية تكرارية، مع ضرورة إجراء الإحالة المرجعية بين الوظائف عند الحاجة. وعلى نحو مماثل، ثمة فئات رئيسية وفئات فرعية بها عناصر تنطبق على عديد من الوظائف، أو التي يقتضي المنطق تنفيذها قبل اتخاذ قرارات فئة فرعية بعينها.

5.1 الحوكمة

وظيفة الحوكمة:

- غرس ثقافة إدارة المخاطر وتطبيقها داخل المؤسسات التي تعمل على تصميم أنظمة الذكاء الاصطناعي أو تطويرها أو نشرها أو تقييمها أو اقتنائها.
- تحديد الخطوط العريضة للعمليات والوثائق والمخططات التنظيمية التي من شأنها توقع وتحديد وإدارة المخاطر التي يمكن أن تتشكل عن النظام، بما في ذلك على المستخدمين وغيرهم من الأفراد في المجتمع، وكذلك تحديد الإجراءات اللازمة لتحقيق تلك النتائج.
- تضمين العمليات بغرض تقييم التأثيرات المحتملة.
- توفير هيكل يمكن أن تتواءم من خلاله وظائف إدارة مخاطر الذكاء الاصطناعي مع المبادئ التنظيمية والسياسات والأولويات الإستراتيجية.
- ربط الجوانب التقنية لتصميم أنظمة الذكاء الاصطناعي وتطويرها بالقيم والمبادئ التنظيمية، وتمكين الممارسات والكفاءات التنظيمية للأفراد المشاركين في عمليات اقتناء هذه الأنظمة والتدريب عليها ونشرها ورصدها.
- التعامل مع دورة حياة المنتج الكاملة والعمليات المرتبطة بها، بما في ذلك المشكلات القانونية وغيرها من المشكلات المتعلقة باستخدام البرامج أو أنظمة الأجهزة والبيانات التابعة للجهات الخارجية.

الحوكمة هي وظيفة شاملة الأبعاد يُجرى دمجها في على مدار جميع مراحل عملية إدارة مخاطر الذكاء الاصطناعي، ومن شأنها تمكين الوظائف الأخرى للعملية. كما يتعين إدماج جوانب وظيفة الحوكمة، ولا سيما تلك المتعلقة بالامتثال أو التقييم، في كل وظيفة من الوظائف الأخرى. إذ يُعد إيلاء الاهتمام بالحوكمة مطلبًا مستمرًا وجوهريًا للإدارة الفعالة لمخاطر الذكاء الاصطناعي على مدار عمر نظام الذكاء الاصطناعي والتسلسل الهرمي للمؤسسة.

يمكن للحوكمة القوية أن توجيه الممارسات والمعايير الداخلية وتحفيزها وتعزيزها من أجل تيسير نشر ثقافة الوعي بالمخاطر التنظيمية. يمكن للسلطات الحاكمة تحديد السياسات الشاملة التي من شأنها توجيه مهمة المؤسسة وأهدافها وقيمها وثقافتها ودرجات تحمل المخاطر. كما تحدد القيادة العليا المسار المتوخى اتباعه لإدارة المخاطر داخل المؤسسة، إلى جانب تحديد الثقافة التنظيمية. إذ تعمل الإدارة على مواءمة الجوانب التقنية لإدارة مخاطر الذكاء الاصطناعي مع السياسات والعمليات. ويمكن للتوثيق أن يعمل على تعزيز مستوى الشفافية، وتحسين عمليات المراجعة البشرية، وتعزيز تطبيق المساءلة بين فِرق أنظمة الذكاء الاصطناعي.

بعد وضع الهياكل والأنظمة والعمليات والفرق الموضحة في وظيفة الحوكمة، يتعين على المؤسسات الاستفادة من الثقافة المكرسة لتحقيق الغرض التي ترتكز على فهم المخاطر وإدارتها. كما يتحتم على مستخدمي إطار العمل مواصلة تنفيذ وظيفة الحوكمة، نظرًا إلى تطور المعارف والثقافات واحتياجات الجهات الفاعلة أو توقعاتها فيما يتعلق بالذكاء الاصطناعي مع مرور الوقت.

يرد وصف الممارسات المتعلقة بحوكمة مخاطر الذكاء الاصطناعي في دليل إطار عمل إدارة مخاطر الذكاء الاصطناعي الصادر عن المعهد الوطني للمعايير والتكنولوجيا. يسرد (الجدول 1) الفئات الرئيسية والفئات الفرعية لوظيفة الحوكمة.

(الجدول 1): الفئات الرئيسية والفئات الفرعية لوظيفة الحوكمة.

الفئات الرئيسية	**الفئات الفرعية**
الحوكمة 1: السياسات والعمليات والإجراءات والممارسات المطبقة عبر المؤسسة ذات الصلة بتخطيط مخاطر الذكاء الاصطناعي وقياسها وإدارتها، التي تتسم بالشفافية ويُجرى تنفيذها بفعالية.	**الحوكمة 1.1:** المتطلبات القانونية والتنظيمية المتعلقة بالذكاء الاصطناعي مفهومة ومدارة وموثقة. **الحوكمة 1.2:** سمات الذكاء الاصطناعي الجدير بالثقة مدمجة في السياسات والعمليات والإجراءات والممارسات التنظيمية. **الحوكمة 1.3:** تطبيق العمليات والإجراءات والممارسات بهدف تحديد المستوى اللازم لأنشطة إدارة المخاطر إستنادًا إلى درجات تحمل المخاطر في المؤسسة. **الحوكمة 1.4:** إنشاء عملية إدارة المخاطر ونتائجها من خلال تطبيق سياسات وإجراءات تتسم بالشفافية، مع وضع ضوابط أخرى استنادًا إلى أولويات المخاطر التنظيمية. **الحوكمة 1.5:** التخطيط لعمليات الرصد المستمر والمراجعة الدورية لعملية إدارة المخاطر ونتائجها، وتحديد الأدوار والمسؤوليات التنظيمية بوضوح، بما في ذلك تحديد وتيرة عمليات المراجعة الدورية. **الحوكمة 1.6:** وضع آليات لحصر أنظمة الذكاء الاصطناعي، مع تزويدها بالموارد وفقًا لأولويات المخاطر التنظيمية. **الحوكمة 1.7:** تطبيق العمليات والإجراءات بهدف إيقاف تشغيل أنظمة الذكاء الاصطناعي والتخلي عنها تدريجيًا بطريقة آمنة لا تسهم في زيادة المخاطر أو تقليل مدى موثوقية المؤسسة.
الحوكمة 2: وضع هياكل المساءلة موضع التنفيذ بحيث يحظى الأفراد والفرق المناسبين بالتمكين والمسؤولية، مع تدريبهم على كيفية مخاطر الذكاء الاصطناعي وقياسها وإدارتها.	**الحوكمة 2.1:** توثيق الأدوار والمسؤوليات وقنوات الاتصال المتعلقة بتخطيط مخاطر الذكاء الاصطناعي وقياسها وإدارتها، مع توضيحها للأفراد والفرق في جميع أنحاء المؤسسة. **الحوكمة 2.2:** تلقي موظفي المؤسسة وشركائها التدريب على كيفية إدارة مخاطر الذكاء الاصطناعي لتمكينهم من أداء واجباتهم ومسؤولياتهم على نحو يتسق مع السياسات والإجراءات والاتفاقيات ذات الصلة.

	الحوكمة 2.3: تحمل القيادة التنفيذية في المؤسسة مسؤولية القرارات المتعلقة بالمخاطر المرتبطة بتطوير نظام الذكاء الاصطناعي ونشره.
الحوكمة 3: إيلاء الأولوية لعمليات التنويع في صفوف القوى العاملة وتحقيق المساواة والشمول وإمكانية الوصول عند تخطيط مخاطر الذكاء الاصطناعي وقياسها وإدارتها طوال دورة الحياة.	**الحوكمة 3.1:** اتخاذ القرارات المتعلق بتخطيط مخاطر الذكاء الاصطناعي وقياسها وإدارتها طوال دورة الحياة على يد فريق متنوع (على سبيل المثال، التنوع في التركيبة الديموغرافية والتخصصات والتجارب والخبرات الفنية والخلفيات). الحوكمة 3.2: وضع السياسات والإجراءات الرامية إلى تحديد الأدوار والمسؤوليات، والتمييز بينها وبين التكوينات المشتركة بين الذكاء الاصطناعي والبشر، والإشراف على أنظمة الذكاء الاصطناعي.
الحوكمة 4: التزام الفرق التنظيمية بتبني ثقافة مراعية لمخاطر الذكاء الاصطناعي وتعلن عنها.	**الحوكمة 4.1:** تطبيق السياسات والممارسات التنظيمية الهادفة إلى تعزيز التفكير الناقد وتبني عقلية "التفكير في السلامة أولًا" عند تصميم أنظمة الذكاء الاصطناعي وتطويرها ونشرها واستخدامها، وذلك بغرض الحد من التأثيرات السلبية المحتملة. **الحوكمة 4.2:** توثيق الفرق التنظيمية للمخاطر والتأثيرات المحتملة لتقنيات الذكاء الاصطناعي التي تصممها وتطورها وتنشرها وتقيّمها وتستخدمها، مع الإعلان عن التأثيرات على نطاق أوسع. **الحوكمة 4.3:** تطبيق الممارسات التنظيمية بهدف تمكين عمليات اختبار الذكاء الاصطناعي والتعرف على الحوادث ومشاركة المعلومات.
الحوكمة 5: تنفيذ العمليات من أجل ضمان تحقق المشاركة القوية مع الجهات الفاعلة المعنية في الذكاء الاصطناعي.	**الحوكمة 5.1:** تطبيق السياسات والممارسات التنظيمية بهدف جمع ودراسة وترتيب أولويات ودمج الملاحظات المتلقاة من الأطراف خارج الفريق التي طورت نظام الذكاء الاصطناعي أو نشرته، وذلك فيما يتعلق بالتأثيرات الفردية والمجتمعية المحتملة ذات الصلة بمخاطر الذكاء الاصطناعي. **الحوكمة 5.2:** إنشاء آليات لتمكين الفريق الذي طور أنظمة الذكاء الاصطناعي أو نشرها من الإدماج المنتظم للملاحظات التي نظرت فيها الجهات الفاعلة المعنية في مجال الذكاء الاصطناعي، وذلك عند تصميم النظام وتنفيذه.
الحوكمة 6: تطبيق السياسات والإجراءات الرامية إلى التصدي لمخاطر الذكاء الاصطناعي والفوائد المترتبة عن برامج وبيانات الأطراف الثالثة وغيرها من المسائل المتعلقة بسلاسل الإمداد.	**الحوكمة 6.1:** تطبيق السياسات والإجراءات الرامية إلى التصدي لمخاطر الذكاء الاصطناعي المرتبطة بجهات الأطراف الثالثة؛ ومن بينها مخاطر انتهاك حقوق الملكية الفكرية أو غيرها من الحقوق الخاصة بالأطراف الثالثة. **الحوكمة 6.2:** تنفيذ عمليات التأهب لحالات الطوارئ بهدف التعامل مع حالات الإخفاق أو الحوادث فيما يتعلق ببيانات الأطراف الثالثة أو أنظمة الذكاء الاصطناعي التي تعتبر عالية الخطورة.

5.2 التخطيط

تحدد وظيفة التخطيط السياق لتأطير المخاطر المتعلقة بنظام الذكاء الاصطناعي. وتتكون دورة حياة الذكاء الاصطناعي من العديد من الأنشطة المترابطة التي تشمل مجموعة متنوعة من الجهات الفاعلة (اطلع على الشكل التوضيحي 3). أما من الناحية العملية، في أغلب الأحيان لا تتوفر لدى الجهات الفاعلة في مجال الذكاء الاصطناعي المسؤولة عن أحد جوانب العملية الرؤية أو السيطرة الكاملة على الجوانب الأخرى والسياقات المرتبطة بها. إذ يمكن أن تؤدي أوجه الترابط بين هذه الأنشطة وبين الجهات الفاعلة المعنية في مجال الذكاء الاصطناعي إلى مواجهة صعوبة في توقع تأثيرات أنظمة الذكاء الاصطناعي على نحو موثوق. فعلى سبيل المثال، يمكن أن يؤدي اتخاذ

قرارات مبكرة بشأن تحديد أغراض نظام الذكاء الاصطناعي وأهدافها إلى تغير سلوكه وقدراته، كما يمكن أن تسهم ديناميات إعداد النشر (مثل المستخدمين النهائيين أو الأفراد المتضررين) في تحديد الشكل الذي ستكون عليه التأثيرات المترتبة على قرارات نظام الذكاء الاصطناعي. وبالتالي، يمكن أن تتعرض أفضل المقاصد ضمن أحد أبعاد دورة حياة الذكاء الاصطناعي إلى الإخلال من خلال التفاعلات مع القرارات والشروط الواردة في الأنشطة الأخرى اللاحقة.

قد يسفر هذه الطبيعة المعقدة والمستويات المتفاوتة من الرؤية عن ظهور أوجه عدم اليقين حيال ممارسات إدارة المخاطر. غير أن عمليات التوقع والتقييم والتصدي للمصادر المحتملة للمخاطر السلبية من شأنه الحد من أوجه عدم اليقين هذه، بالإضافة إلى تعزيز نزاهة عملية اتخاذ القرار.

علاوة على ذلك، يمكن للمعلومات التي تُجمع أثناء تنفيذ وظيفة التخطيط أن تساعد في الحيلولة دون ظهور المخاطر السلبية وإرشاد عملية اتخاذ القرار بشأن العمليات مثل إدارة النماذج، فضلًا عن اتخاذ قرار أولي بشأن مدى ملاءمة حل الذكاء الاصطناعي أو الحاجة إليه. تُشكل النتائج في وظيفة التخطيط الأساس الذي تستند إليه وظائف القياس والإدارة. وبدون المعرفة السياقية والوعي بالمخاطر ضمن السياقات المحددة، يصعب أداء مهام إدارة المخاطر. إذ تهدف وظيفة التخطيط إلى تعزيز قدرة المؤسسة على تحديد المخاطر والعوامل المساهمة الأوسع نطاقًا.

يُجرى تعزيز سبل تنفيذ هذه الوظيفة من خلال إدماج وجهات نظر الفريق الداخلي المتنوع وبمشاركة الأطراف خارج الفريق التي طورت نظام الذكاء الاصطناعي أو نشرته. وقد يختلف مستوى التعامل مع الجهات المتعاونة الخارجية والمستخدمين النهائيين والمجتمعات التي يُحتمل تأثرها وغيرهم من الأطراف استنادًا إلى مستوى المخاطر التي يحفل بها نظام ذكاء اصطناعي بعينه وتركيبة الفريق الداخلي والسياسات التنظيمية المطبقة. ويمكن لجمع وجهات النظر واسعة النطاق هذه أن يساعد المؤسسات في الحيلولة دون ظهور المخاطر السلبية بشكل استباقي، وتطوير أنظمة ذكاء اصطناعي أكثر جدارة بالثقة، وذلك من خلال:

- تحسين قدرات المؤسسات على فهم السياقات.
- التحقق من افتراضات المؤسسات حول سياق الاستخدام.
- تمكين القدرة على التعرف على الحالات التي لا تعمل فيها الأنظمة داخل سياقها المنشود أو خارجه.
- تحديد الاستخدامات الإيجابية والمفيدة لأنظمة الذكاء الاصطناعي الموجودة لدى المؤسسات. • تحسين مستوى فهم القيود المفروضة على عمليات الذكاء الاصطناعي وتعلم الآلة.
- تحديد القيود المفروضة على التطبيقات الواقعية التي قد ينجم عنها تأثيرات سلبية.
- تحديد التأثيرات السلبية المعروفة والمتوقعة المتعلقة بالاستخدام المزمع لأنظمة الذكاء الاصطناعي.
- توقع مخاطر استخدام أنظمة الذكاء الاصطناعي التي قد تتجاوز الاستخدام المزمع.

بعد الانتهاء من أداء وظيفة التخطيط، يجب أن يتحقق لدى مستخدمي إطار العمل معرفة سياقية كافية حول تأثيرات نظام الذكاء الاصطناعي وذلك للاسترشاد بها عند اتخاذ القرار الأولي بالمضي/عدم المضي بشأن احتمالية تصميم نظام ذكاء اصطناعي أو تطويره أو نشره. وفي حال التوصل إلى قرار بالمضي قدمًا، فيتعين على المؤسسات الاستفادة من وظائف القياس والإدارة إلى جانب تطبيق السياسات والإجراءات المحددة في وظيفة الحوكمة، وذلك للمساعدة في جهود إدارة مخاطر الذكاء الاصطناعي. ويتعين على مستخدمي إطار العمل مواصلة تطبيق وظيفة التخطيط على أنظمة الذكاء الاصطناعي، نظرًا إلى تطور السياق والقدرات والمخاطر والفوائد والتأثيرات المحتملة بمرور الوقت.

يرد وصف الممارسات المتعلقة بعملية تخطيط مخاطر الذكاء الاصطناعي في دليل إطار عمل إدارة مخاطر الذكاء الاصطناعي الصادر عن المعهد الوطني للمعايير والتكنولوجيا. يسرد (الجدول 2) الفئات الرئيسية والفئات الفرعية لوظيفة التخطيط.

(الجدول 2): الفئات الرئيسية والفئات الفرعية لوظيفة التخطيط.

الفئات الرئيسية	الفئات الفرعية
التخطيط 1: تأسيس السياق وفهمه	**التخطيط 1.1:** فهم وتوثيق الأغراض المنشودة، والاستخدامات المفيدة المحتملة، والقوانين محددة السياق، والمعايير والتوقعات، والبيئات المتوقع أن ينشر فيها نظام الذكاء الاصطناعي. وتشمل الاعتبارات: البيئة المحددة أو أنواع المستخدمين فضلًا عن توقعاتهم، والتأثيرات الإيجابية والسلبية المحتملة لاستخدامات النظام على الأفراد والمجتمعات والمؤسسات والمجتمع

	والكوكب، والافتراضات والقيود ذات الصلة بأغراض نظام الذكاء الاصطناعي واستخداماته ومخاطره على مدار عملية التطوير أو دورة حياة منتج الذكاء الاصطناعي، وعمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) ومقاييس النظام ذات الصلة. **التخطيط 1.2:** نجاح الجهات الفاعلة متعددة التخصصات في مجال الذكاء الاصطناعي والكفاءات والمهارات والقدرات اللازمة لإنشاء السياق في تجسيد التنوع الديموغرافي والخبرة الفنية الواسعة في المجال وتجربة المستخدم، مع توثيق مشاركة تلك الجهات. منح الأولوية لفرص التعاون متعدد التخصصات. **التخطيط 1.3:** فهم وتوثيق مهمة المؤسسة والأهداف ذات الصلة بتقنيات الذكاء الاصطناعي. **التخطيط 1.4:** تحديد قيمة الأعمال أو سياق استخدام الأعمال بشكل واضح، أو إعادة تقييمها في حالة تقييم أنظمة الذكاء الاصطناعي الحالية. **التخطيط 1.5:** تحديد درجات تحمل المخاطر التنظيمية وتوثيقها. **التخطيط 1.6:** استنباط متطلبات النظام (على سبيل المثال، "ضرورة احترام النظام لخصوصية مستخدميه") وفهمها من قِبل الجهات الفاعلة المعنية في مجال الذكاء الاصطناعي. مراعاة قرارات التصميم للتداعيات الاجتماعية-التقنية عند التصدي لمخاطر الذكاء الاصطناعي.
التخطيط 2: إجراء تصنيف لنظام الذكاء الاصطناعي.	**التخطيط 2.1:** تحديد المهام والأساليب المحددة المستخدمة لتنفيذ المهام التي سيدعمها نظام الذكاء الاصطناعي (على سبيل المثال، المصنِفات، النماذج التوليدية، الموصيات). **التخطيط 2.2:** توثيق المعلومات حول حدود المعرفة لنظام الذكاء الاصطناعي وكيفية استخدام مخرجات النظام والإشراف عليها بشريًا. إتاحة التوثيق للمعلومات كافية اللازمة لمساعدة الجهات الفاعلة المعنية في مجال الذكاء الاصطناعي عند اتخاذ القرارات والإجراءات اللاحقة. **التخطيط 2.3:** تحديد وتوثيق اعتبارات السلامة العلمية وعمليات الاختبار والتقييم والتحقق والمصادقة (TEVV)، بما في ذلك تلك الاعتبارات المتعلقة بالتصميم التجريبي، وجمع البيانات والاختيار (على سبيل المثال، التوافر، والتمثيل، ومدى الملاءمة)، وجدارة النظام بالثقة، والتحقق من صلاحية النتائج.
التخطيط 3: فهم قدرات الذكاء الاصطناعي والاستخدام المستهدف والأهداف والفوائد والتكاليف المتوقعة مقارنة بالمعايير المرجعية المناسبة.	**التخطيط 3.1:** دراسة وتوثيق الفوائد المحتملة للصلاحية الوظيفية والأداء المنشودين لنظام الذكاء الاصطناعي. **التخطيط 3.2:** دراسة وتوثيق التكاليف المحتملة، بما في ذلك التكاليف غير النقدية، التي تنتج عن أخطاء الذكاء الاصطناعي المتوقعة أو المحققة أو الصلاحية الوظيفية للنظام وجدارته بالثقة، باعتبارها مرتبطة بدرجات تحمل المخاطر التنظيمية. **التخطيط 3.3:** تحديد نطاق التطبيق المستهدف وتوثيقه بناءً على قدرة النظام والسياق المحدد وتصنيف نظام الذكاء الاصطناعي. **التخطيط 3.4:** تحديد وتقييم وتوثيق عمليات المعنية بكفاءة المشغل والممارس وأداء نظام الذكاء الاصطناعي ومدى موثوقيته، وكذلك المعايير الفنية وشهادات المصادقة ذات الصلة. **التخطيط 3.5:** تحديد عمليات الإشراف البشري وتقييمها وتوثيقها وفقًا للسياسات التنظيمية الواردة في وظيفة الحوكمة
التخطيط 4: تخطيط المخاطر والفوائد لجميع مكونات نظام الذكاء	**التخطيط 4.1:** تطبيق ومتابعة وتوثيق نُهُج تخطيط تقنيات الذكاء الاصطناعي والمخاطر القانونية لمكوناتها؛ بما في ذلك استخدام بيانات أو برامج الأطراف الثالثة، وكذلك مخاطر انتهاك حقوق الملكية الفكرية أو غيرها من الحقوق الخاصة بالأطراف الثالثة.

الاصطناعي؛ بما في ذلك برامج وبيانات الأطراف الثالثة.	**التخطيط 4.2:** تحديد وتوثيق ضوابط المخاطر الداخلية لمكونات نظام الذكاء الاصطناعي، بما في ذلك تقنيات الذكاء الاصطناعي التابعة للأطراف الثالثة.
التخطيط 5: وصف التأثيرات على الأفراد والجماعات والمجتمعات والمؤسسات والمجتمع.	**التخطيط 5.1:** تحديد وتوثيق احتمالية وحجم كل تأثير محدد (كلا من التأثيرات المحتملة المفيدة والضارة) بناءً على حجم الاستخدام المتوقع، والاستخدامات السابقة لأنظمة الذكاء الاصطناعي في سياقات مماثلة، وتقارير الحوادث العامة، وردود الفعل الأطراف خارج الفريق التي طور نظام الذكاء الاصطناعي أو نشرته، أو غيرها من البيانات. **التخطيط 5.2:** جاهزية الممارسات والموظفين اللازمين لدعم المشاركة المنتظمة مع الجهات الفاعلة المعنية في مجال الذكاء الاصطناعي وتوثيق ذلك، مع إدماج الملاحظات حول التأثيرات الإيجابية والسلبية وغير المتوقعة.

5.3 القياس

تستعين وظيفة القياس بأدوات وتكنولوجيات كمّية أو نوعية أو مختلطة الأساليب بهدف تحليل مخاطر الذكاء الاصطناعي والتأثيرات ذات الصلة وتقييمها وقياسها ورصدها. كما تستعين وظيفة القياس بالمعرفة ذات الصلة بمخاطر الذكاء الاصطناعي المحددة في وظيفة التخطيط، وتسترشد بوظيفة الإدارة. لذا يتعين اختبار أنظمة الذكاء الاصطناعي قبل نشرها وبشكل منتظم في أثناء تشغيلها. وتشمل قياسات مخاطر الذكاء الاصطناعي توثيق جانبي الصلاحية الوظيفية للأنظمة ومدى موثوقيتها.

تتضمن عملية قياس مخاطر الذكاء الاصطناعي تتبع مقاييس السمات الجديرة بالثقة والتأثير الاجتماعي والتكوينات المشتركة بين الذكاء الاصطناعي والبشر. ويجب أن تتضمن العمليات المطورة أو المعتمدة في وظيفة القياس إجراء اختبارات صارمة للبرامج، واتباع منهجيات تقييم الأداء مزود بها مقاييس لأوجه عدم اليقين المرتبطة بها، وعقد مقارنات مع المعايير المرجعية للأداء، وإعداد تقارير رسمية، وتوثيق النتائج. يمكن لعمليات المراجعة المستقلة أن تحسن من فعالية عمليات الاختبار، ويمكن أن تخفف أيضًا من التحيزات الداخلية وتضارب المصالح المحتمل.

عندما تنشأ مفاضلات بين السمات الجديرة بالثقة، يمكن للقياسات أن توفر أساسًا يمكن تتبعه للاسترشاد به عند اتخاذ القرارات المتعلقة بالإدارة. وقد تتضمن الخيارات كلًا من إعادة المعايرة، أو تخفيف التأثير، أو إزالة النظام من التصميم، أو التطوير، أو الإنتاج، أو الاستخدام، إلى جانب مجموعة من ضوابط التعويض والاستكشاف والردع والتوجيه والاسترداد.

بعد الانتهاء من إجراء وظيفة القياس، يتم إجراء عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) الموضوعية أو القابلة للتكرار أو القابلة للتوسع؛ بما في ذلك المقاييس والأساليب والمنهجيات، ومتابعتها وتوثيقها. يجب أن تلتزم المقاييس ومنهجيات القياس بالمعايير العلمية والقانونية والأخلاقية وأن يُجرى تنفيذها في إطار عملية مفتوحة وشفافة. وقد يلزم تطوير أنواع القياسات الجديدة، سواء النوعية والكمّية. كما يجب مراعاة المدى الذي يمكن لكل نوع من أنواع القياسات أن يوفر من خلاله معلومات فريدة وذات مغزى فيما يتعلق بتقييم مخاطر الذكاء الاصطناعي. كما يتعين على مستخدمي إطار العمل تعزيز قدرتهم على إجراء تقييمات شاملة لمدى موثوقية النظام، وتحديد المخاطر الحالية والناشئة وتتبعها، والتحقق من مدى فعالية المقاييس. بالإضافة إلى ذلك، سيُجرى الاستعانة بنتائج القياسات في وظيفة الإدارة، وذلك بهدف المساعدة في رصد المخاطر وجهود الاستجابة. كما يتحتم على مستخدمي إطار العمل مواصلة تطبيق وظيفة القياس على أنظمة الذكاء الاصطناعي، نظرًا لتطور المعارف والمنهجيات والمخاطر والتأثيرات بمرور الوقت.

يرد وصف الممارسات المتعلقة بقياس مخاطر الذكاء الاصطناعي في دليل إطار عمل إدارة مخاطر الذكاء الاصطناعي الصادر عن المعهد الوطني للمعايير والتكنولوجيا. يسرد (الجدول 3) الفئات الرئيسية والفئات الفرعية لوظيفة القياس.

(الجدول 3): الفئات الرئيسية والفئات الفرعية لوظيفة القياس.

الفئات الرئيسية	الفئات الفرعية

القياس 1: تحديد الأساليب والمقاييس المناسبة وتطبيقها.	**القياس 1.1:** اختيار الأساليب والمقاييس لقياس مخاطر الذكاء الاصطناعي المنصوص عليها في وظيفة التخطيط لبدء التنفيذ ابتداءً بأهم مخاطر الذكاء الاصطناعي. التوثيق الصحيح للمخاطر أو سمات الجدارة بالثقة التي لن أو لا يمكن قياسها. **القياس 1.2:** تقييم مدى ملاءمة مقاييس الذكاء الاصطناعي وفعالية الضوابط الحالية وتحديثها بانتظام، على أن يتضمن ذلك تقارير الأخطاء والتأثيرات المحتملة على المجتمعات المتضررة. **القياس 1.3:** إشراك الخبراء الداخليين الذين لم يعملوا بصفتهم المطورين المباشرين للنظام و/أو المقيّمين المستقلين في إجراء عمليات التقييم والتحديث المنتظمة. استشارة الخبراء في المجال والمستخدمين والجهات الفاعلة في مجال الذكاء الاصطناعي خارج الفريق الذين طوروا نظام الذكاء الاصطناعي أو نشروه والمجتمعات المتضررة من أجل دعم التقييمات، عند الحاجة ووفقًا لدرجة تحمل المخاطر التنظيمية.
القياس 2: تقييم أنظمة الذكاء الاصطناعي من حيث السمات الجديرة بالثقة.	**القياس 2.1:** توثيق مجموعات الاختبار والمقاييس والتفاصيل المتعلقة بالأدوات المستخدمة خلال عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV). **القياس 2.2:** تلبية التقييمات التي تُجرى على العنصر البشري للمتطلبات واجبة التطبيق (ومن بينها حماية العنصر البشري)، وتمثيلها للفئات المعنية. **القياس 2.3:** قياس أداء نظام الذكاء الاصطناعي أو معايير ضمان جودته نوعيًا أو كمّيًا، وثبوت ذلك من خلال خضوعه لظروف مشابهة لإعداد (إعدادات) النشر. بالإضافة إلى توثيق الإجراءات. **القياس 2.4:** مراقبة الصلاحية الوظيفية وسلوك نظام الذكاء الاصطناعي ومكوناته، وفقًا للنحو المحدد في وظيفة التخطيط، في أثناء عملية الإنتاج. **القياس 2.5:** ثبوت أن نظام الذكاء الاصطناعي المقرر نشره صالح وموثوق. توثيق حدود القابلية للتعميم خارج إطار الظروف التي يُجرى فيها تطوير التكنولوجيا. **القياس 2.6:** تقييم نظام الذكاء الاصطناعي بانتظام لمعرفة مخاطر السلامة، وذلك على النحو المحدد في وظيفة التخطيط. ثبوت أن نظام الذكاء الاصطناعي المقرر نشره آمن، وأن مخاطره السلبية المتبقية لا تتجاوز درجة تحمل المخاطر، وأنه يمكن أن يتعطل بطريقة مأمونة، ولا سيما في حال تشغيله خارج نطاق حدود معرفته. تجسيد مقاييس الأمان لمدى موثوقية النظام وقوته، وعمليات الرصد الآنية، وأوقات الاستجابة لأعطال نظام الذكاء الاصطناعي. **القياس 2.7:** تقييم وتوثيق أمن نظام الذكاء الاصطناعي ومرونته، على النحو المحدد في وظيفة التخطيط. **القياس 2.8:** دراسة وتوثيق المخاطر المرتبطة بالشفافية والمساءلة، على النحو المحدد في وظيفة التخطيط. **القياس 2.9:** شرح نموذج الذكاء الاصطناعي والتحقق من صحته وتوثيقه، وتفسير مخرجات نظام الذكاء الاصطناعي في سياقه، وذلك على النحو المحدد في وظيفة التخطيط، بهدف الاسترشاد بها لتحقيق الاستخدام المسؤول والحوكمة. **القياس 2.10**: دراسة وتوثيق مخاطر الخصوصية لنظام الذكاء الاصطناعي، على النحو المحدد في وظيفة التخطيط.

	القياس 2.11: تقييم مدى الإنصاف والتحيز، على النحو المحدد في وظيفة التخطيط، وتوثيق النتائج. **القياس 2.12:** تقييم وتوثيق التأثير البيئي والاستدامة لأنشطة التدريب والإدارة لنموذج الذكاء الاصطناعي، وذلك على النحو المحدد في وظيفة التخطيط. **القياس 2.13:** تقييم وتوثيق مدى فعالية مقاييس وعمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) المستخدمة في وظيفة القياس.
القياس 3: تطبيق آليات تتبع مخاطر الذكاء الاصطناعي المحددة على مدار الوقت.	**القياس 3.1:** جاهزية الأساليب والموظفين وعمليات التوثيق بهدف إجراء عمليات التحديد والتتبع المنتظمة لمخاطر الذكاء الاصطناعي الحالية وغير المتوقعة والناشئة بناءً على عوامل مثل الأدائين المنشود والفعلي في السياقات المنشورة. **القياس 3.2:** مراعاة نُهُج تتبع المخاطر للبيئات التي يصعب فيها تقييم مخاطر الذكاء الاصطناعي باستخدام تقنيات القياس المتاحة حاليًا أو حيث لا تتوفر فيها المقاييس بعد. **القياس 3.3:** إنشاء عمليات تلقي الملاحظات من المستخدمين النهائيين والمجتمعات المتضررة للإبلاغ عن المشكلات والتماس نتائج النظام وإدماجها في مقاييس تقييم نظام الذكاء الاصطناعي.
القياس 4: جمع الملاحظات حول مدى فعالية عملية القياس وتقييمها.	**القياس 4.1:** ارتباط نُهُج القياس لتحديد مخاطر الذكاء الاصطناعي بسياق (سياقات) النشر واستلهامها من خلال المشاورات المجراة مع الخبراء في المجال والمستخدمين النهائيين الآخرين. بالإضافة إلى توثيق النُهُج. **القياس 4.2:** استلهام نتائج القياس المتعلقة بمدى موثوقية نظام الذكاء الاصطناعي في سياق (سياقات) النشر وعلى مدار دورة حياة الذكاء الاصطناعي من خلال الإسهامات والمدخلات المقدمة من الخبراء في المجال والجهات الفاعلة المعنية في مجال الذكاء الاصطناعي، وذلك بهدف التحقق مما إذا كان النظام يعمل بشكل متسق على النحو المنشود. بالإضافة إلى توثيق النتائج. **القياس 4.3:** تحديد وتوثيق تحسينات الأداء القابلة للقياس أو الإخفاقات استنادًا إلى المشاورات المجراة مع الجهات الفاعلة المعنية في مجال الذكاء الاصطناعي؛ بما في ذلك المجتمعات المتضررة، مع تحديد وتوثيق البيانات الميدانية حول المخاطر ذات الصلة بالسياق وسمات الجدارة بالثقة.

5.4 الإدارة

تستبع وظيفة الإدارة تخصيص موارد المخاطر لتخطيط المخاطر وقياسها على أساس منتظم وعلى النحو المحدد وفقًا لوظيفة الحوكمة. حيث تتضمن عمليات معالجة المخاطر وضع خطط للاستجابة للحوادث أو الأحداث والتعافي منها والإبلاغ عنها.

يُجرى الاستعانة بالمعلومات السياقية المستقاة من المشاورات المجراة الخبراء والمدخلات المستمدة من الجهات الفاعلة المعنية في مجال الذكاء الاصطناعي؛ التي جرى إنشاؤها في وظيفة الحوكمة وتنفيذها في وظيفة التخطيط، في هذه الوظيفة بهدف تقليل احتمالية تعطل النظام والتأثيرات السلبية. كما تعمل ممارسات التوثيق المنهجية المنشأة في وظيفة الحوكمة والمستخدمة في وظيفتي التخطيط والقياس على تعزيز جهود إدارة مخاطر الذكاء الاصطناعي وزيادة الشفافية والمساءلة. بالإضافة إلى ذلك، تُطبق عمليات تقييم المخاطر الناشئة، إلى جانب وضع آليات للتحسين المستمر.

بعد الانتهاء من إجراء وظيفة الإدارة، سيُجرى وضع خطط لتحديد أولويات المخاطر والرصد والتحسين بشكل منتظم. وسيحظى مستخدمو إطار العمل بقدرة معززة على إدارة مخاطر أنظمة الذكاء الاصطناعي المنتشرة وتخصيص موارد إدارة المخاطر بناءً على المخاطر التي جرى تقييمها وتحديدها وفقًا للأولوية. كما يتحتم على مستخدمي إطار العمل مواصلة تطبيق وظيفة الإدارة على أنظمة

الذكاء الاصطناعي المنشورة، نظرًا إلى تطور الأساليب والسياقات والمخاطر واحتياجات الجهات الفاعلة المعنية في مجال الذكاء الاصطناعي أو توقعاتها مع مرور الوقت.

يرد وصف الممارسات المتعلقة بإدارة مخاطر الذكاء الاصطناعي في دليل إطار عمل إدارة مخاطر الذكاء الاصطناعي الصادر عن المعهد الوطني للمعايير والتكنولوجيا. يسرد (الجدول 4) الفئات الرئيسية والفئات الفرعية لوظيفة الإدارة.

(الجدول 4): الفئات الرئيسية والفئات الفرعية لوظيفة الإدارة.

الفئات الرئيسية	الفئات الفرعية
الإدارة 1: تحديد أولويات مخاطر الذكاء الاصطناعي المستندة إلى التقييمات والمخرجات التحليلية الأخرى الناتجة عن وظائف التخطيط والقياس، والاستجابة لها وإدارتها.	**الإدارة 1.1:** تحديد ما إذا كان نظام الذكاء الاصطناعي يحقق الأغراض المنشودة والأهداف المعلنة وما إذا كان ينبغي المضي قدمًا في تطويره أو نشره. **الإدارة 1.2:** تحديد أولويات معالجة مخاطر الذكاء الاصطناعي الموثقة بناءً على التأثير والاحتمالية والموارد المتاحة أو الأساليب المستخدمة. **الإدارة 1.3:** تطوير وتخطيط وتوثيق سبل الاستجابة لمخاطر الذكاء الاصطناعي التي تعتبر ذات أولوية عالية، على النحو المحدد في وظيفة التخطيط. يمكن أن تشمل خيارات الاستجابة للمخاطر كلًا من وسائل تخفيفها أو تحويلها أو تجنبها أو قبولها. **الإدارة 1.4:** توثيق المخاطر المتبقية السلبية (المعرفة باعتبارها مجموع جميع المخاطر التي لا يمكن تخفيفها)، وذلك بالنسبة إلى لكل من المشترين النهائيين لأنظمة الذكاء الاصطناعي والمستخدمين النهائيين.
الإدارة 2: تخطيط إستراتيجيات تعظيم فوائد الذكاء الاصطناعي والحد من التأثيرات السلبية وإعدادها وتنفيذها وتوثيقها واستلهامها من المدخلات المقدمة من الجهات الفاعلة المعنية في مجال الذكاء الاصطناعي.	**الإدارة 2.1:** النظر في الموارد اللازمة لإدارة مخاطر الذكاء الاصطناعي، إلى جانب الأنظمة أو النُهُج أو الأساليب البديلة قابلة للتطبيق بخلاف الذكاء الاصطناعي، وذلك بهدف تقليص حجم التأثيرات المحتملة أو احتمالية وقوعها. **الإدارة 2.2:** وضع آليات للحفاظ على قيمة أنظمة الذكاء الاصطناعي المنتشرة وتطبيقها. **الإدارة 2.3:** اتباع إجراءات الاستجابة للمخاطر المجهولة في السابق عند تحديدها وإجراءات التعافي منها. **الإدارة 2.4:** وضع الآليات وتطبيقها، وتحديد المسؤوليات وفهمها، وذلك حتى تحل محل أنظمة الذكاء الاصطناعي التي تظهر أداءًا أو نتائج غير متوافقة مع الاستخدام المنشود أو تفكيكها أو تعطيلها.
الإدارة 3: إدارة مخاطر الذكاء الاصطناعي والفوائد المتوخاة من جهات الأطراف الثالثة.	**الإدارة 3.1:** رصد مخاطر الذكاء الاصطناعي وفوائده المتوخاة من موارد الأطراف الثالثة بشكل منتظم، وتطبيق ضوابط المخاطر وتوثيقها. **الإدارة 3.2:** رصد النماذج المُدرَبة مسبقًا والمستخدمة للتطوير في إطار عمليات الرصد والصيانة المنتظمة لنظام الذكاء الاصطناعي.
الإدارة 4: توثيق ومراقبة إجراءات معالجة المخاطر، بما في ذلك إجراءات الاستجابة والتعافي، وخطط التواصل بشأن مخاطر الذكاء الاصطناعي التي جرى تحديدها وقياسها، مع رصدها بشكل منتظم.	**الإدارة 4.1:** تنفيذ خطط رصد نظام الذكاء الاصطناعي في مرحلة ما بعد النشر، ويتضمن ذلك آليات استخلاص وتقييم المدخلات من المستخدمين والجهات الفاعلة المعنية الأخرى في مجال الذكاء الاصطناعي، وكذلك آليات الالتماس والتجاوز، وإيقاف التشغيل، والاستجابة للحوادث، والتعافي، وإدارة التغيير. **الإدارة 4.2:** دمج الأنشطة القابلة للقياس المعنية بالتحسينات المستمرة في تحديثات نظام الذكاء الاصطناعي التي تشمل المشاركة المنتظمة مع الأطراف ذات المصلحة، بما في ذلك الجهات الفاعلة المعنية في مجال الذكاء الاصطناعي.

	الإدارة 4.3: إبلاغ الحوادث والأخطاء إلى الجهات الفاعلة المعنية في مجال الذكاء الاصطناعي، بما في ذلك المجتمعات المتضررة. متابعة وتوثيق عمليات تتبع الحوادث والأخطاء والاستجابة لها والتعافي منها.

6. ملفات تعريف إطار عمل إدارة مخاطر الذكاء الاصطناعي

ملفات تعريف حالة استخدام إطار عمل إدارة مخاطر الذكاء الاصطناعي هي عمليات تنفيذ وظائف إطار عمل إدارة مخاطر الذكاء الاصطناعي وفئاتها الرئيسية وفئاتها الفرعية في إطار بيئة أو تطبيق محدد بناءً على المتطلبات ودرجة تحمل المخاطر وموارد مستخدم إطار العمل، ومنها على سبيل المثال، ملف تعريف التوظيف في إطار عمل إدارة مخاطر الذكاء الاصطناعي أو ملف الإسكان العادل في إطار عمل إدارة مخاطر الذكاء الاصطناعي. ومن شأن ملفات التعريف أن توضح وتقدم رؤى حول كيفية إدارة المخاطر في المراحل المختلفة على مدار دورة حياة منتج الذكاء الاصطناعي أو فيما يتعلق بقطاع أو تقنية محددة أو تطبيقات الاستخدام النهائي. كما تساعد ملفات تعريف إطار عمل إدارة مخاطر الذكاء الاصطناعي المؤسسات في تحديد أفضل طريقة لإدارة مخاطر الذكاء الاصطناعي على متواءم جيدًا مع أهدافها، كما أنها تراعي المتطلبات القانونية/التنظيمية وأفضل الممارسات، وتعكس أولويات إدارة المخاطر.

ملفات التعريف الزمنية لإطار عمل إدارة مخاطر الذكاء الاصطناعي هي أوصاف للحالة الحالية أو الحالة المستهدفة المنشودة لأنشطة إدارة مخاطر الذكاء الاصطناعي المحددة ضمن قطاع أو صناعة أو مؤسسة أو سياق تطبيق محدد. كما يشير ملف تعريف إطار عمل إدارة مخاطر الذكاء الاصطناعي الحالي إلى كيفية الإدارة الحالية للذكاء الاصطناعي والمخاطر ذات الصلة من حيث النتائج الحالية. ويشير ملف التعريف المستهدف إلى النتائج المطلوبة لتحقيق أهداف إدارة مخاطر الذكاء الاصطناعي المنشودة أو المستهدفة.

عند عقد مقارنة بين ملفات التعريف الحالية والمستهدفة، من المرجح أن يكشف ذلك عن وجود ثغرات يتعين معالجتها، وذلك للتمكن من تحقيق أهداف إدارة مخاطر الذكاء الاصطناعي. ويمكن وضع خطط عمل لمعالجة هذه الفجوات من أجل تحقيق النتائج في فئة رئيسية أو فئة فرعية محددة. ويتم تحديد أولويات التخفيف من الفجوات وفقًا لاحتياجات المستخدم وعمليات إدارة المخاطر. إذ يسمح هذا النهج القائم على المخاطر أيضًا لمستخدمي إطار العمل بمقارنة نُهُجهم المتبعة مع الأساليب الأخرى واستقصاء الموارد اللازمة (على سبيل المثال التوظيف والتمويل)، وذلك لتحقيق أهداف إدارة مخاطر الذكاء الاصطناعي بشكل فعال من حيث التكلفة والأولوية العليا.

تغطي ملفات التعريف متعددة قطاعات في إطار عمل إدارة مخاطر الذكاء الاصطناعي مخاطر النماذج أو التطبيقات التي يمكن استخدامها على صعيد حالات الاستخدام أو القطاعات. ويمكن أن تغطي ملفات التعريف متعددة القطاعات أيضًا سبل حوكمة المخاطر المتعلقة بالأنشطة أو أساليب سير العمل المشتركة عبر القطاعات مع تخطيطها وقياسها وإدارتها، ويشمل ذلك استخدام نماذج اللغة الكبيرة أو الخدمات المستندة إلى السحابة أو الاستحواذ.

لا يفرض إطار العمل هذا وصفًا نماذج ملفات التعريف؛ ما يسمح بالمرونة في التنفيذ.

الملحق (أ):
أوصاف مهام الجهة الفاعلة في مجال الذكاء الاصطناعي وفق الشكلين التوضحيين (2) و(3)

تُنفذ مهام تصميم الذكاء الاصطناعي في أثناء مراحل تحديد سياق التطبيق وجمع البيانات والمدخلات من دورة حياة الذكاء الاصطناعي المبينة في الشكل التوضيحي (2). تتولى الجهات الفاعلة المسؤولة عن تصميم الذكاء الاصطناعي إنشاء مفهوم أنظمة الذكاء الاصطناعي وأهدافها، كما أنها تتولى مسؤولية مهام التخطيط والتصميم وجمع بيانات نظام الذكاء الاصطناعي ومعالجتها، بحيث يكون نظام الذكاء الاصطناعي قانونيًا وملائمًا للغرض. وتشمل المهام صياغة وتوثيق مفهوم النظام وأهدافه والافتراضات الأساسية والسياق والمتطلبات، وجمع البيانات وتنقيتها، وتوثيق البيانات الوصفية وسمات مجموعة البيانات. وتشمل الجهات الفاعلة في مجال الذكاء الاصطناعي في هذه الفئة كلًا من علماء البيانات، والخبراء في المجال، ومحللي الظواهر الاجتماعية والثقافية، وخبراء في مجال التنوع والإنصاف والشمولية وإمكانية الوصول، وأعضاء المجتمعات المتضررة، وخبراء العوامل البشرية (مثل تصميم واجهة المستخدم (UI / UX))، وخبراء الحوكمة، ومهندسي البيانات، ومزودي البيانات، وممولي النظام، ومديري المنتجات، وجهات الأطراف الثالثة، والمقيّمين، وخبراء الحوكمة القانونية والمتعلقة بالخصوصية.

تُنفذ مهام **تطوير الذكاء الاصطناعي** في أثناء مرحلة بناء نموذج الذكاء الاصطناعي من دورة الحياة المبينة في الشكل التوضيحي (2). توفر الجهات الفاعلة المسؤولة عن تطوير الذكاء الاصطناعي البنية التحتية الأولية لأنظمة الذكاء الاصطناعي، كما تتولى مسؤولية مهام بناء النماذج وتفسيرها، والتي تتضمن إنشاء النماذج أو الخوارزميات و/أو اختيارها و/أو معايرتها و/أو التدريب عليها و/أو اختبارها. وتشمل الجهات الفاعلة في مجال الذكاء الاصطناعي في هذه الفئة كلًا من خبراء التعلم الآلي، وعلماء البيانات، والمطورين، وجهات الأطراف الثالثة، وخبراء الحوكمة القانونية والمتعلقة بالخصوصية، والخبراء في العوامل الاجتماعية والثقافية والسياقية المرتبطة ببيئة النشر.

تُنفذ مهام **نشر الذكاء الاصطناعي** في أثناء مرحلة المهام والمخرجات من دورة الحياة المبينة في الشكل التوضيحي (2). تتولى الجهات الفاعلة المسؤولة عن نشر الذكاء الاصطناعي مسؤولية اتخاذ القرارات السياقية المتعلقة بكيفية استخدام نظام الذكاء الاصطناعي لضمان نشر النظام حتى مرحلة الإنتاج. وتشمل المهام ذات الصلة كلًا من الإطلاق التجريبي للنظام، والتحقق من مدى التوافق مع الأنظمة القديمة، وضمان الامتثال للوائح التنظيمية، وإدارة التغيير التنظيمي، وتقييم تجربة المستخدم. وتشمل الجهات الفاعلة في مجال الذكاء الاصطناعي في هذه الفئة كلًا من المختصين في تكامل الأنظمة، ومطوري البرامج، والمستخدمين النهائيين، والمشغلين والممارسين، والمقيّمين، والخبراء في المجال ذوي الخبرة في مجالات العوامل البشرية والتحليل الاجتماعي والثقافي والحوكمة.

تُنفذ مهام **التشغيل والرصد** في سياق التطبيق/مرحلة التشغيل والرصد من دورة الحياة المبينة في الشكل التوضيحي (2). يتولى تنفيذ هذه المهام الجهات الفاعلة في مجال الذكاء الاصطناعي المسؤولة عن تشغيل نظام الذكاء الاصطناعي والعاملة مع الجهات الأخرى، وذلك بهدف تقييم مخرجات النظام وتأثيراته بشكل منتظم. تشمل الجهات الفاعلة في مجال الذكاء الاصطناعي في هذه الفئة كلًا من مشغلي النظام، والخبراء في المجال، ومصممي الذكاء الاصطناعي، والمستخدمين الذين يفسرون أو يدمجون مخرجات أنظمة الذكاء الاصطناعي، ومطوري المنتجات، والمقيّمين والمراجعين، وخبراء الامتثال، والإدارة التنظيمية، وأعضاء المجتمع موضع البحث.

تُنفذ مهام **عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV)** طوال دورة حياة الذكاء الاصطناعي. وتتولى تنفيذها الجهات الفاعلة في مجال الذكاء الاصطناعي التي تعمل على دراسة نظام الذكاء الاصطناعي أو مكوناته، أو تسكتشف المشكلات وتعالجها. من الناحية المثلى، تختلف الجهات الفاعلة في مجال الذكاء الاصطناعي التي تقوم بمهام التحقق والمصادقة عن تلك الجهات المسؤولة عن إجراءات الاختبار والتقييم. يمكن دمج المهام في مرحلة مبكر مثل مرحلة التصميم، حيث يمكن تخطيط الاختبارات وفقًا لمتطلبات التصميم.

- قد تركز مهام عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) المعنية بالتصميم والتخطيط والبيانات على التحقق الداخلي والخارجي من الافتراضات المتعلقة بتصميم النظام، وجمع البيانات، والقياسات ذات الصلة بالسياق المنشود للنشر أو التطبيق.
- تتضمن مهام عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) المعنية بالتطوير (أي بناء النموذج) التحقق من النموذج وتقييمه.
- تتضمن مهام عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) المعنية بالنشر التحقق من النظام والتكامل في الإنتاج، مع إجراء الاختبارات وإعادة المعايرة للأنظمة وتكامل العمليات، وتجربة المستخدم، والامتثال للمواصفات القانونية والتنظيمية والأخلاقية الحالية.

- تتضمن مهام عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) المعنية بالعمليات إجراء عمليات رصد مستمرة للتحديثات الدورية، والاختبارات، وإعادة المعايرة على يد الخبراء المتخصصين (SME)، وتتبع الحوادث أو الأخطاء المبلغ عنها وإدارتها، واكتشاف الخصائص الناشئة والتأثيرات ذات الصلة، وعمليات الإصلاح والاستجابة.

يمكن إيجاد المهام والأنشطة المتعلقة **بالعوامل البشرية** في جميع أبعاد دورة حياة الذكاء الاصطناعي، حيث تشمل ممارسات ومنهجيات التصميم التي ترتكز على الإنسان، وتعزيز المشاركة النشطة للمستخدمين النهائيين والأطراف المعنية الأخرى والجهات الفاعلة المعنية في مجال الذكاء الاصطناعي، وإدماج المعايير والقيم محددة السياق في تصميم النظام، وتقييم تجارب المستخدم النهائي وتكييفها، والتكامل الواسع للبشر والديناميات البشرية في جميع مراحل دورة حياة الذكاء الاصطناعي. يسهم المتخصصون في مجال العوامل البشرية في تقديم مهارات ووجهات نظر متعددة التخصصات لفهم سياق الاستخدام، والإفادة في التنوع الديموغرافي متعدد التخصصات، والمشاركة في العمليات الاستشارية، وتصميم تجربة المستخدم وتقييمها، وإجراء عمليات التقييم والاختبار المرتكزة على الإنسان، والإفادة في تقييمات التأثير.

تتضمن مهام **الخبراء في المجال** المدخلات المقدمة من الممارسين متعددي التخصصات أو العلماء الذين يقدمون المعرفة أو الخبرة في وحول القطاع الصناعي أو القطاع الاقتصادي أو السياق أو مجال التطبيق الذي يُستخدم فيه نظام الذكاء الاصطناعي. يمكن للجهات الفاعلة في مجال الذكاء الاصطناعي من الخبراء في المجال تقديم إرشادات جوهرية عن تصميم نظام الذكاء الاصطناعي وتطويره، وتفسير المخرجات لدعم العمل الذي تؤديه عمليات الاختبار والتقييم والتحقق والمصادقة (TEVV) وفرق تقييم تأثير الذكاء الاصطناعي.

تتضمن مهام **تقييم تأثير الذكاء الاصطناعي** تقدير وتقييم متطلبات مساءلة نظام الذكاء الاصطناعي، ومكافحة التحيز الضار، ودراسة تأثيرات أنظمة الذكاء الاصطناعي، وسلامة المنتجات، والمسؤولية، والأمن، من جملة أمور أخرى. توفر الجهات الفاعلة في مجال الذكاء الاصطناعي، مثل مقدّري ومقيّمي التأثير، الخبرة الفنية والبشرية والاجتماعية والثقافية والقانونية.

تُنفذ مهام **الشراء** الجهات الفاعلة في مجال الذكاء الاصطناعي مع السلطة المالية أو القانونية أو الإدارية للسياسات، وذلك بغرض شراء نماذج الذكاء الاصطناعي أو منتجاته أو خدماته من مطور أو بائع أو متعاقد تابع لأطراف ثالثة.

تتولى مهام **الحوكمة والرقابة** الجهات الفاعلة في مجال الذكاء الاصطناعي التي تتمتع بالسلطة الإدارية والائتمانية والقانونية والمسؤولة عن المؤسسة التي يُجرى فيها تصميم نظام الذكاء الاصطناعي و/أو تطويره و/أو نشره. تتولى الجهات الفاعلة الرئيسية في مجال الذكاء الاصطناعي مسؤولية حوكمة الذكاء الاصطناعي؛ التي تشمل الإدارة التنظيمية والقيادة العليا ومجلس الإدارة. وتُعد الجهات الفاعلة هذه هي الأطراف المعنية بتأثير المؤسسة واستدامتها بشكل عام.

الجهات الفاعلة الأخرى في مجال الذكاء الاصطناعي

تشمل **جهات الأطراف الثالثة** كلًا من مقدمي الخدمات، والمطورين، والبائعين، ومقيمي البيانات و/أو الخوارزميات و/أو النماذج و/أو الأنظمة والخدمات ذات الصلة لمؤسسة أخرى أو عملاء المؤسسة أو المتعاملين معها. تتولى جهات الأطراف الثالثة مسؤولية مهام تصميم الذكاء الاصطناعي وتطويره، كليًا أو جزئيًا. وبحكم التعريف، تُعد هذه الجهات أطرافًا خارج فريق التصميم أو التطوير أو النشر في المؤسسة التي تشتري تقنياتها أو خدماتها. وقد تكون التقنيات التي يُجرى شراؤها من أطراف ثالثة معقدة أو غامضة، وقد لا تتوافق معدلات تحمل المخاطر مع مؤسسة النشر أو التشغيل.

المستخدمون النهائيون لنظام الذكاء الاصطناعي هم الأفراد أو المجموعات التي تستخدم النظام لأغراض محددة. ويتفاعل هؤلاء الأفراد أو المجموعات مع نظام الذكاء الاصطناعي في إطار سياق محدد. قد يكون هناك تنوع بين المستخدمين النهائيين من حيث الكفاءة؛ ويشمل ذلك خبراء الذكاء الاصطناعي حتى المستخدمين النهائيين للتكنولوجيا لأول مرة.

يشمل **الأفراد/المجتمعات المتضررين** جميع الأفراد أو المجموعات أو المجتمعات أو المؤسسات المتأثرة بشكل مباشر أو غير مباشر بأنظمة الذكاء الاصطناعي أو القرارات المستندة إلى مخرجات أنظمة الذكاء الاصطناعي. ولا يتفاعل هؤلاء الأفراد بالضرورة مع النظام أو التطبيق الذي جرى نشره.

قد تقدم **الجهات الفاعلة الأخرى في مجال الذكاء الاصطناعي** معايير أو إرشادات رسمية أو شبه رسمية لتحديد مخاطر الذكاء الاصطناعي وإدارتها. وقد تشكل هذه الجهات كلًا من الاتحادات التجارية، ومنظمات وضع المعايير، ومجموعات المناصرة، والباحثين، والجماعات البيئية، ومنظمات المجتمع المدني.

من المرجح أن يتعرض **الجمهور العام** بشكل مباشر للتأثيرات الإيجابية والسلبية لتقنيات الذكاء الاصطناعي. قد توفر هذه الفئة الدافع للإجراءات التي تتخذها الجهات الفاعلة في مجال الذكاء الاصطناعي. وقد تشمل هذه المجموعة الأفراد والمجتمعات والمستهلكين المرتبطين بالسياق الذي يُجرى فيه تطوير نظام الذكاء الاصطناعي أو نشره.

الملحق (ب):
كيف تختلف مخاطر الذكاء الاصطناعي عن مخاطر البرامج التقليدية

كما هو الحال مع البرامج التقليدية، يمكن أن تكون المخاطر المترتبة على التقنيات القائمة على الذكاء الاصطناعي أكبر قدرة الشركة، وقد تمتد إلى مؤسسات أخرى وتحدث تأثيرات مجتمعية. ويمكن أن ينشأ عن أنظمة الذكاء الاصطناعي أيضًا مجموعة من المخاطر التي لا يمكن معالجتها بشكل شامل من خلال أُطر العمل ونُهُج المخاطر الحالية. وقد يكون بعض ميزات نظام الذكاء الاصطناعي التي تنطوي على مخاطر فوائد أيضًا. فعلى سبيل المثال، يمكن للنماذج المُدَّربة مسبقًا ونقل التعلم تطوير البحوث وزيادة الدقة والمرونة عند مقارنتها بالنماذج والنُهُج الأخرى. ومن شأن تحديد العوامل السياقية في وظيفة التخطيط مساعدة الجهات الفاعلة في مجال الذكاء الاصطناعي على تحديد مستوى المخاطر وجهود الإدارة المحتملة.

مقارنةً بالبرامج التقليدية، تتضمن المخاطر الخاصة بالذكاء الاصطناعي الجديدة أو المطردة ما يلي:

- قد لا توفر البيانات المستخدمة في بناء نظام ذكاء اصطناعي تمثيلًا صحيحًا أو مناسبًا للسياق أو الاستخدام المنشود لنظام الذكاء الاصطناعي، وقد لا تكون الحقائق المرجعية موجودة أو غير متاحة. بالإضافة إلى ذلك، يمكن أن يؤثر التحيز الضار ومشكلات جودة البيانات الأخرى في موثوقية نظام الذكاء الاصطناعي، الأمر الذي قد يسفر عن تأثيرات سلبية.
- يرتبط الاعتماد على نظام الذكاء الاصطناعي والاعتماد على البيانات لأغراض مهام التدريب، إلى جانب زيادة الحجم والتعقيد عادةً بهذا النوع من البيانات.
- قد تؤدي التغييرات المتعمدة أو غير المتعمدة في أثناء مرحلة التدريب إلى تغيير أداء نظام الذكاء الاصطناعي بشكل جوهري.
- قد تصبح مجموعات البيانات المستخدمة للتدريب على أنظمة الذكاء الاصطناعي منفصلة عن سياقها الأصلي والمنشود، أو قد تصبح قديمة أو متقادمة بالنسبة إلى سياق النشر.
- نطاق نظام الذكاء الاصطناعي ومدى تعقيده (تحتوي العديد من الأنظمة على مليارات بل تريليونات من نقاط اتخاذ القرار) الموجودة داخل تطبيقات البرامج الأكثر تقليدية.
- يمكن أن يؤدي استخدام النماذج المُدَّربة مسبقًا التي يمكنها تعزيز مستوى البحوث وتحسين الأداء إلى زيادة مستويات أوجه عدم اليقين الإحصائية والتسبب في مشكلات تتعلق بإدارة التحيز والصحة العلمية وقابلية التكرار.
- مواجهة درجة أعلى من الصعوبة عند التنبؤ بأنماط تعطل الخصائص الناشئة للنماذج واسعة النطاق المُدَّربة مسبقًا.
- مخاطر الخصوصية بسبب القدرة التجميعية للبيانات المحسنة في أنظمة الذكاء الاصطناعي.
- قد تتطلب أنظمة الذكاء الاصطناعي إجراء المزيد من عمليات الصيانة المتكررة والمحفزات لإجراء الصيانة التصحيحية بسبب البيانات أو النموذج أو انحراف المفهوم.
- الغموض والمخاوف المتزايدة بشأن قابلية التكرار.
- معايير اختبار البرامج غير المطورة وعدم القدرة على توثيق الممارسات القائمة على الذكاء الاصطناعي وفقًا للمعيار المتوقع من البرامج المصممة تقليديًا لجميع الحالات باستثناء أبسطها.
- مواجهة صعوبة في إجراء عمليات الاختبار المنتظمة للبرامج القائمة على الذكاء الاصطناعي، أو تحديد ما يجب اختباره، لأن أنظمة الذكاء الاصطناعي لا تخضع الضوابط ذاتها مثل عملية تطوير الأكواد التقليدية.
- التكاليف الحسابية لتطوير أنظمة الذكاء الاصطناعي وتأثيرها في البيئة والكوكب.
- عدم القدرة على التنبؤ أو اكتشاف الآثار الجانبية للأنظمة القائمة على الذكاء الاصطناعي علة نحو يتجاوز المقاييس الإحصائية.

يمكن تطبيق اعتبارات نُهج إدارة مخاطر الخصوصية والأمن السيبراني عند تصميم أنظمة الذكاء الاصطناعي وتطويرها ونشرها وتقييمها واستخدامها. إذ تعتبر مخاطر الخصوصية والأمن السيبراني أيضًا جزءًا من اعتبارات إدارة مخاطر المؤسسة الأوسع نطاقًا التي قد تتضمن مخاطر الذكاء الاصطناعي. وفي إطار الجهود المبذولة للتعامل مع سمات الجدارة بالثقة في الذكاء الاصطناعي مثل "الأمن والمرونة" و "تحسين الخصوصية"، قد تفكر المؤسسات في الاستفادة من المعايير والإرشادات المتاحة التي تقدم إرشادات واسعة النطاق للمؤسسات بهدف الحد من مخاطر الأمان والخصوصية، ومنها على سبيل المثال: إطار عمل الأمن السيبراني الخاص بالمعهد الوطني للمعايير والتكنولوجيا، وإطار عمل الخصوصية الخاص بالمعهد الوطني للمعايير والتكنولوجيا، وإطار عمل إدارة المخاطر الخاص بالمعهد الوطني للمعايير والتكنولوجيا، وإطار عمل التطوير الآمن للبرامج. تتشارك جميع أطر العمل هذه في بعض الميزات مع إطار عمل إدارة مخاطر الذكاء الاصطناعي. وعلى غرار معظم نُهج إدارة المخاطر، فهي تستند إلى النتائج وليست إلزامية، وغالبًا ما تتمحور حول مجموعة من وظائف جوهر إطار العمل وفئاتها الرئيسية وفئاتها الفرعية. في حين أن هناك اختلافات جوهرية بين أطر العمل هذه

بناءً على المجال الذي يُجرى التصدي له، ونظرًا إلى أن إدارة مخاطر الذكاء الاصطناعي تتطلب التصدي إلى العديد من أنواع المخاطر الأخرى، فإن أطر العمل مثل تلك المذكورة أعلاه يمكن الاسترشاد بها في اعتبارات الأمان والخصوصية في وظائف التخطيط والقياس والإدارة الخاصة بإطار عمل إدارة مخاطر الذكاء الاصطناعي.

في الوقت نفسه، لا تتطرق الإرشادات المتاحة قبل نشر إطار إطار عمل إدارة مخاطر الذكاء الاصطناعي هذا بشكل شامل إلى العديد من مخاطر أنظمة الذكاء الاصطناعي. فعلى سبيل المثال، أطر العمل والإرشادات الحالية لا يمكنها:

- إدارة مشكلة التحيز الضار في أنظمة الذكاء الاصطناعي على النحو الصحيح. • مواجهة المخاطر الباعثة على التحدي المتعلقة بالذكاء الاصطناعي التوليدي.
- التصدي بشكل شامل للمخاوف الأمنية المتعلقة بالتهرب، أو استخراج النموذج، أو هجمات استنتاج العضوية، أو الإتاحة، أو غير ذلك من هجمات نماذج التعلم الآلي.
- تفسير ثغرات سطح الهجوم المعقد في أنظمة الذكاء الاصطناعي أو الانتهاكات الأمنية الأخرى التي تمكّنها أنظمة الذكاء الاصطناعي.
- النظر في المخاطر المرتبطة بتقنيات الذكاء الاصطناعي لجهات الأطراف الثالثة، ونقل التعلم، والاستخدام بدون تصريح حيث يمكن تدريب أنظمة الذكاء الاصطناعي على اتخاذ القرار خارج ضوابط أمان المؤسسة أو التدريب في مجال واحد ثم "تعديل ضبطها" في مجال آخر.

تخضع تقنيات وأنظمة الذكاء الاصطناعي والبرامج التقليدية للابتكار سريعة الوتيرة. لذا يتعين رصد التقدم التكنولوجي ونشره للاستفادة من تلك التطورات والعمل نحو تهيئة مستقبل للذكاء الاصطناعي جدير بالثقة ومسؤول.

الملحق (ج):
إدارة مخاطر الذكاء الاصطناعي والتفاعل بين الإنسان والذكاء الاصطناعي

يمكن المؤسسات المعنية بتصميم أنظمة الذكاء الاصطناعي أو تطويرها أو نشرها لاستخدامها في البيئات التشغيلية العمل على تعزيز إدارة مخاطر الذكاء الاصطناعي من خلال فهم القيود الحالية للتفاعل بين الإنسان والذكاء الاصطناعي. ويوفر إطار عمل إدارة مخاطر الذكاء الاصطناعي فرصًا لتحديد الأدوار والمسؤوليات البشرية المختلفة والتمييز بينها بوضوح عند استخدام أنظمة الذكاء الاصطناعي أو التفاعل معها أو إدارتها.

ثمة العديد من الأساليب القائمة على البيانات التي تعتمد عليها أنظمة الذكاء الاصطناعي في محاولةً منها لتحويل ممارسات الرصد واتخاذ القرار الفردية والاجتماعية إلى كميات قابلة للقياس أو تمثيلها. وقد يتحقق تمثيل الظواهر البشرية المعقدة بنماذج رياضية على حساب إزالة السياق الضروري. قد يؤدي فقدان السياق بدوره إلى صعوبة فهم التأثيرات الفردية والمجتمعية التي تعتبر أساسية في جهود إدارة مخاطر الذكاء الاصطناعي.

تشمل المسائل التي تستحق المزيد من الدراسة والبحث ما يلي:

1. **ضرورة تحديد وتمييز الأدوار والمسؤوليات البشرية في عملية صنع القرار والإشراف على أنظمة الذكاء الاصطناعي.** يمكن أن تمتد التكوينات المشتركة بين الذكاء الاصطناعي والبشر لتشمل تلك المستقلة (الذاتية) بالكامل إلى يدوية بالكامل. ويمكن لأنظمة الذكاء الاصطناعي اتخاذ القرارات بشكل مستقل، أو إحالة عملية اتخاذ القرار إلى خبير بشري، أو استخدامها من جاني صانع قرار بشري باعتبارها رأيًا إضافيًا. قد لا تتطلب بعض أنظمة الذكاء الاصطناعي إشرافًا بشريًا، مثل النماذج المستخدمة لتحسين ضغط الفيديو. وقد تتطلب أنظمة أخرى إشرافًا بشريًا على وجه التحديد.
2. **تعكس القرارات التي تدخل في تصميم أنظمة الذكاء الاصطناعي وتطويرها ونشرها وتقييمها واستخدامها التحيزات المعرفية النظامية والبشرية.** ويمكن أن تسهم الجهات الفاعلة في الذكاء الاصطناعي في إدخال تحيزاتها المعرفية الفردية والجماعية في العملية أيضًا. وقد تنبع التحيزات من مهام صنع القرار للمستخدم النهائي، ويتم تقديمها على مدار دورة حياة الذكاء الاصطناعي من خلال الافتراضات البشرية والتوقعات والقرارات في أثناء مهام التصميم والنمذجة. قد تتفاقم هذه التحيزات، التي قد لا تكون بالضرورة ضارة دائمًا، نتيجة غموض نظام الذكاء الاصطناعي وما ينتج عن ذلك من نقص في الشفافية. وقد تؤثر التحيزات النظامية في المستوى التنظيمي فيما يتعلق بكيفية هيكلة الفرق وتحديد المتحكم في عمليات صنع القرار على مدار دورة حياة الذكاء الاصطناعي. ويمكن أن تؤثر هذه التحيزات أيضًا في القرارات النهائية من جانب المستخدمين النهائيين وصناع القرار وواضعي السياسات، الأمر الذي قد يؤدي بدوره إلى تأثيرات سلبية.
3. **تختلف نتائج التفاعل بين الإنسان والذكاء الاصطناعي.** ففي ظل ظروف معينة ؛على سبيل المثال في مهام الحكم على أساس الإدراك الحسي يمكن لجزء الذكاء الاصطناعي من التفاعل بين الإنسان والذكاء الاصطناعي أن يضخم التحيزات البشرية، الأمر الذي من شأنه أن يؤدي إلى قرارات أكثر تحيزًا مقارنة بالقرارات المتخذة من جانب الذكاء الاصطناعي أو الإنسان كل على حدة. عندما مراعاة هذه الاختلافات عند تنظيم الفرق المشتركة بين الذكاء الاصطناعي والبشر، فإنها يمكن أن تؤدي إلى تحقيق التكامل وتحسين الأداء العام.
4. **تُعد عملية تقديم معلومات نظام الذكاء الاصطناعي للبشر أمرًا معقدًا.** إذ يتمتع البشر بالقدرة على الإدراك ويستمدون المعنى من مخرجات وتفسيرات نظام الذكاء الاصطناعي بطرق مختلفة، ما يعكس التفضيلات والسمات والمهارات الفردية المختلفة.

توفر وظيفة **الحوكمة** للمؤسسات الفرصة لتوضيح وتحديد الأدوار والمسؤوليات للبشر في التكوينات المشتركة بين الذكاء الاصطناعي والفريق البشري وتلك الجهات المشرفة على أداء نظام الذكاء الاصطناعي. إذ تعمل وظيفة الحوكمة أيضًا على إنشاء آليات للمؤسسات لجعل عمليات صنع القرار لديها أكثر وضوحًا، وللمساعدة في مواجهة التحيزات النظامية.

تقترح وظيفة **التخطيط** فرصًا لتحديد وتوثيق العمليات لضمان كفاءة المشغل والممارس مع أداء نظام الذكاء الاصطناعي ومفاهيم الجدارة بالثقة، وتحديد المعايير وشهادات المصادقة الفنية ذات الصلة. من شأن تنفيذ الفئات الرئيسية والفئات الفرعية لوظيفة التخطيط مساعدة المؤسسات على تحسين كفاءتها الداخلية لتحليل السياق، وتحديد القيود الإجرائية والنظامية، واستكشاف ودراسة تأثيرات الأنظمة القائمة على الذكاء الاصطناعي في العالم الحقيقي، وتقييم عمليات صنع القرار طوال دورة حياة الذكاء الاصطناعي.

تصف وظيفتا **الحوكمة والتخطيط** أهمية تعدد التخصصات والفرق المتنوعة ديموغرافيًا والاستفادة من الملاحظات المقدمة من الأفراد والمجتمعات التي يُحتمل تأثرها. ويمكن للجهات الفاعلة في مجال الذكاء الاصطناعي التي تم استدعاؤها في إطار قياس الذكاء الاصطناعي التي تؤدي مهام وأنشطة العوامل البشرية أن تساعد الفرق الفنية من خلال ترسيخ ممارسات التصميم والتطوير لنوايا المستخدم وممثلي مجتمع الذكاء الاصطناعي الأوسع نطاقًا والقيم المجتمعية. ويمكن لهذه الجهات الفاعلة المساعدة أيضًا في إدماج المعايير والقيم الخاصة بالسياق في تصميم النظام وتقييم تجارب المستخدم النهائي، بالتنسيق مع أنظمة الذكاء الاصطناعي.

سيُجرى تعزيز نُهُج إدارة مخاطر الذكاء الاصطناعي للتكوينات المشتركة بين الذكاء الاصطناعي والبشر من خلال البحث والتقييم المستمر. فعلى سبيل المثال، تتطلب درجة تمكين البشر وتحفيزهم لتحدي مخرجات نظام الذكاء الاصطناعي 'جراء المزيد من الدراسات. ويمكن للبيانات المتعلقة بمدى التكرار والأساس المنطقي التي ينقض فيها البشر مخرجات نظام الذكاء الاصطناعي في الأنظمة المنتشرة أن تكون ذات فائدة عند تجميع البيانات وتحليلها.

الملحق (د):
سمات إطار عمل إدارة مخاطر الذكاء الاصطناعي

قدم المعهد الوطني للمعايير والتكنولوجيا وصفًا للعديد من السمات الرئيسية لإطار عمل إدارة مخاطر الذكاء الاصطناعي عند بدء العمل على إطار العمل لأول مرة. وظلت هذه السمات على حالها، وجرى استخدامها في توجيه عملية تطوير إطار عمل إدارة مخاطر الذكاء الاصطناعي. ويرد ذكر هذه السمات أدناه باعتبارها مرجعًا.

يسعى إطار عمل إدارة مخاطر الذكاء الاصطناعي إلى:

1. أن يكون قائماً على المخاطر، ومتسم بالكفاءة من حيث الموارد، وداعمًا للابتكار، واختياريًا.

2. أن يكون مدفوعًا بتوافق الآراء، ومتطورًا ومُحدَّثًا بانتظام من خلال عملية مفتوحة وشفافة. إذ يجب أن تتاح لجميع أصحاب المصلحة الفرصة للمساهمة في تطوير إطار عمل إدارة مخاطر الذكاء الاصطناعي.

3. أن يكون مستخدمًا للغة واضحة وبسيطة يمكن لجمهور عريض فهمها، بمن في ذلك كبار الموظفين التنفيذيين والمسؤولين الحكوميين وقيادة المنظمات غير الحكومية والأفراد غير متخصصين في مجال الذكاء الاصطناعي الذي يحظون بمعرفة فنية متعمقة كافية للاستفادة بها من جانب الممارسين. ويتعين على إطار عمل إدارة مخاطر الذكاء الاصطناعي أن يتيح إمكانية الإبلاغ عن مخاطر الذكاء الاصطناعي عبر المؤسسة، وبين المؤسسات، والعملاء، والجمهور بشكل عام.

4. أن يوفر لغة وفهم مشتركين لإدارة مخاطر الذكاء الاصطناعي. إذ يتعين على إطار عمل إدارة مخاطر الذكاء الاصطناعي أن يقدم التصنيفات والمصطلحات والتعريفات والمقاييس والتوصيفات المتعلقة بمخاطر الذكاء الاصطناعي.

5. أن يكون سهل الاستخدام ويتناسب بشكل جيد مع الجوانب الأخرى لإدارة المخاطر. إذ يتعين أن يكون إطار العمل سهل الاستخدام وقابلًا للتكيف بسهولة باعتباره جزءًا من إستراتيجية وعمليات إدارة المخاطر الأوسع للمؤسسة. كما يجب أن يكون متسقًا أو متواءمًا مع الأساليب الأخرى لإدارة مخاطر الذكاء الاصطناعي.

6. أن يكون ذا فائدة لمجموعة واسعة من وجهات النظر والقطاعات ومجالات التكنولوجيا. إذا يتعين أن يكون إطار عمل إدارة مخاطر الذكاء الاصطناعي قابلًا للتطبيق عالميًا على أي تقنية ذكاء اصطناعي وفي حالات الاستخدام ذات السياق المخصص.

7. أن يكون مرتكزًا على النتائج وغير إلزاميًا. إذا يتعين أن يوفر إطار العمل قائمة للنتائج والنُهُج بدلًا من وصف متطلبات واحدة تناسب الجميع.

8. أن يحقق الاستفادة ويعزز الوعي بالمعايير القائمة، والإرشادات، وأفضل الممارسات، والمنهجيات، والأدوات المستخدمة لإدارة مخاطر الذكاء الاصطناعي، مع توضيح الحاجة إلى موارد إضافية ومحسنة.

9. أن يتمتع بالحيادية تجاه القوانين واللوائح التنظيمية. إذ يتعين على إطار العمل أن يدعم قدرات المؤسسات على العمل في ظل الأنظمة القانونية أو التنظيمية المحلية والدولية المعمول بها.

10. أن يكون وثيقة حية وقابلة للتعديل. إذ يتعين تحديث إطار عمل إدارة مخاطر الذكاء الاصطناعي بسهولة متى تغيرت التقنيات ومستوى الفهم والنُهُج ذات الصلة بموثوقية الذكاء الاصطناعي واستخدامات الذكاء الاصطناعي، حيث يدرك أصحاب المصلحة هذا الأمر عند تنفيذ إدارة مخاطر الذكاء الاصطناعي بشكل عام وإطار العمل هذا على وجه الخصوص.

هذا المنشور متاح مجانًا على الرابط التالي:

https://doi.org/10.6028/NIST.AI.100-1