AI ACTION SUMMIT

التقرير الدولي حول سلامة الذكاء الاصطناعي

تقرير علمي حول سلامة الذكاء الاصطناعي المتقدم

يناير/كانون الثاني 2025

المساهمون

الرئيس

البروفيسور. يوشوا بينجيو، جامعة مونتريال / ميلا - معهد الذكاء الاصطناعي في كيبيك

لجنة الخبراء الاستشارية

رُشِحَت هذه اللجنة الدولية من قِبل حكومات البلدان الثلاثين المُدرجة أدناه، بالإضافة إلى الأمم المتحدة، والاتحاد الأوروبي، ومنظمة التعاون الاقتصادي والتنمية.

أستراليا: برونوين فوكس، جامعة نيو ساوث ويلز

البرازيل: أندريه كارلوس بونس دي ليون فيريرا دي كارفالو، معهد الرياضيات وعلوم الكمبيوتر، جامعة ساو باولو

كندا: منى نمر، المستشار العلمي الرئيسي لكندا

شيلي: راكيل بيزوا ريفيرا، الجامعة التقنية فيديريكو سانتا ماريا

الصين: يي زينج، الأكاديمية الصينية للعلوم

الاتحاد الأوروبي: جوها هيكيلا، المكتب الأوروبي لمنظمة العفو الدولية

فرنسا: غيوم أفرين، التنسيقية الوطنية للذكاء الاصطناعي

ألمانيا: أنطونيو كروجر، المركز الألماني لأبحاث الذكاء الاصطناعي

الهند: بالارامان رافيندران، كلية وادواني لعلوم البيانات والذكاء الاصطناعي، المعهد الهندي للتكنولوجيا في مدراس

إندونيسيا: همام رضا، البحث التعاوني والابتكار الصناعي في مجال الذكاء الاصطناعي (كوريكا)

أيرلندا: كيران سيويغي، أبحاث أيرلندا

إسرائيل: زيف كاتسير، هيئة الابتكار الإسرائيلية

إيطاليا: أندريا مونتي، الخبير القانوني لدى وكيل وزارة الدولة للتحول الرقمي، رئاسة مجلس الوزراء الإيطالي

اليابان: هيرواكي كيتانو، شركة مجموعة سوني

كينيا: نوسو موامانزي، وزارة تكنولوجيا المعلومات والاتصالات والاقتصاد الرقمي

المملكة العربية السعودية: فهد البلوي، الهيئة السعودية للبيانات والذكاء الاصطناعي

المكسيك: خوسيه رامون لوبيز بورتيلو، LobsterTel

هولندا: هارون الشيخ، المجلس العلمي الهولندي للسياسة الحكومية

نيوزيلندا: جيل جولي، وزارة الأعمال والابتكار والتوظيف

نيجيريا: أولوبونمي أجالا، وزارة الاتصالات والابتكار والاقتصاد الرقمي

منظمة التعاون الاقتصادي والتنمية: جيري شيهان، مدير مديرية العلوم والتكنولوجيا والابتكار

الفلبين: دومينيك فنسنت ليجوت، CirroLytix

جمهورية كوريا: كيونغ مو لي، قسم الهندسة الكهربائية والحاسوبية، جامعة سيول الوطنية

رواندا: كريستال روجيج، مركز الثورة لصناعية الرابعة (C4IR)

سنغافورة: دينيس وونغ، مجموعة ابتكار وحماية البيانات، هيئة تطوير وسائل الإعلام والاتصالات

إسبانيا: نوريا أوليفر، إليس أليكانتي

سويسرا: كريستيان بوش، الإدارة الاتحادية للشؤون الاقتصادية والتعليم والبحث

تركيا: أحمد هاليت خطيب، وزارة الصناعة والتكنولوجيا التركية

أوكرانيا: أوليكسي مولشانوفسكي، لجنة الخبراء المعنية بتطوير الذكاء الاصطناعي في أوكرانيا

الإمارات العربية المتحدة: مروان السركال، وزارة شؤون مجلس الوزراء، مكتب رئيس الوزراء

المملكة المتحدة: كريس جونسون، كبير المستشارين العلميين في وزارة العلوم والابتكار والتكنولوجيا

الأمم المتحدة: أماندیب سينغ جيل، وكيل الأمين العام لشؤون التكنولوجيا الرقمية والناشئة ومبعوث الأمين العام لشؤون التكنولوجيا

الولايات المتحدة: سيف م. خان، الولايات المتحدة. وزارة التجارة

كاتب رئيسي

دانيال بريفيتيرا، مركز كيرا

مجموعة الكتابة

تاماي بيسير أوغلو، Epoch AI

ريشي بوماساني، جامعة ستانفورد

ستيفن كاسبر، معهد ماساتشوستس للتكنولوجيا

يجين تشوي، جامعة ستانفورد

فيليب فوكس، مركز كيرا

بن جارفينكل، جامعة أكسفورد

دانييل جولدفارب، ميلا - معهد الذكاء الاصطناعي في كيبيك

مانتاس مازيكا، جامعة إلينوي في أوربانا-شامبين

جوليان مايكل، جامعة نيويورك

جيسيكا نيومان، جامعة كاليفورنيا، بيركلي

كوان يي نج، Concordia AI

جيريش ساستري، مستقل

هدى حيدري، جامعة كارنيجي ميلون

تشيناسا ت. أوكولو، مؤسسة بروكينجز

ديبورا راجي، جامعة كاليفورنيا، بيركلي

أنسون هو، Epoch AI

ساياش كابور، جامعة برينستون

ليلى خالتبري، جامعة هونج كونج للعلوم والتكنولوجيا

شاين لونجبري، معهد ماساتشوستس للتكنولوجيا

سام مانينغ، مركز حوكمة الذكاء الاصطناعي

فاسيليوس مافروديس، معهد آلان تورينج

ثيودورا سكيداس، Humane Intelligence

توبين ساوث، معهد ماساتشوستس للتكنولوجيا

إيما ستروبيل، جامعة كارنيجي ميلون

فلوريان ترامير، المعهد الفيدرالي السويسري للتكنولوجيا في زيورخ

لوسيا فيلاسكو، جامعة ماستريخت

نيكول ويلر، جامعة برمنجهام

كبار المستشارين

دارون أسيموجلو، معهد ماساتشوستس للتكنولوجيا

أولوبايو أديكانمبي، ساهم كمستشار أول قبل توليه منصبه في إيكوالايز إيه آي (EqualyzAI)

ديفيد دالريمبل، وكالة الأبحاث المتقدمة والاختراع

توماس ج. ديتريش، جامعة ولاية أوريغون

ادوارد دبليو. فيلتن، جامعة برينستون

باسكال فونج، ساهمت كأحد كبار المستشارين قبل تولي منصبها في Meta

بيير أوليفييه جورينشاس، قسم الأبحاث، صندوق النقد الدولي

فريدريك هاينتز، جامعة لينشوبينغ

بيرنهارد شولكوبف، معهد إليس توبنغن

دون سونغ، جامعة كاليفورنيا، بيركلي

ألفارو سوتو، الجامعة البابوية الكاثوليكية في تشيلي

نيك جينينجز، جامعة لوفبورو

أندرياس كراوس، المعهد الفيدرالي السويسري للتكنولوجيا في زيورخ

سوزان ليفي، جامعة كلية دبلن

بيرسي ليانغ، جامعة ستانفورد

تيريزا لوديرمير، الجامعة الفيدرالية في بيرنامبوكو

فيدوشي ماردا، إيه آي كولابوراتيف (AI Collaborative)

إليزابيث سيجر (كاتبة عامة)، ديموس

هيلين مارغيتس، جامعة أكسفورد

جون ماكديرميد، جامعة يورك

جيفري هينتون، جامعة تورنتو

القيادة العلمية

سورين ميندرمان، ميلا - معهد الذكاء الاصطناعي في كيبيك

ألوندرا نيلسون، معهد الدراسات المتقدمة
كلارا نيبل، معهد مهندسي الكهرباء والإلكترونيات (IEEE)
أليس أوه، مدرسة KAIST للحوسبة
جوبال رامشورن، Responsible AI UK
ستيوارت راسل، جامعة كاليفورنيا، بيركلي
ماريبيتجي شاكي، جامعة ستانفورد

لي تيدريتش، جامعة ديوك
جايل فاروكو، المعهد الفرنسي للأبحاث في علوم الكمبيوتر والتشغيل الآلي (Inria)
أندرو ياو، معهد علوم المعلومات متعددة التخصصات، جامعة تسينغهوا
يا تشين تشانغ، جامعة تسينغهوا
جين مونجا، مؤسسة كارنيغي للسلام الدولي
أرفيند نارايانان، جامعة برينستون

الأمانة العامة

معهد سلامة الذكاء الاصطناعي (AI Safety Institute)

باران أكار

بن كليفورد

لامبريني داس

كلير دينيس

فريا هيمبلمان

هانا ميرشانت

ريان أوفري

بن سنودين

ميلا - معهد الذكاء الاصطناعي في كيبيك

جوناثان باري

بنيامين برودوم

الإفادات

المراجعون من المجتمع المدني والصناعة

المجتمع المدني: معهد آدا لوفليس، منتدى الذكاء الاصطناعي في نيوزيلندا / تي كاهوي أتاماي إياهيكو أو أوتياروا، مجموعة الخبراء المؤقتة في مجال الذكاء الاصطناعي في أستراليا، مؤسسة كارنيغي للسلام الدولي، مركز القانون والابتكار / مؤسسة سيرتا، مركز حوكمة الذكاء الاصطناعي، مركز رئيس القضاة مائير شامغار للقانون الرقمي والابتكار، معهد إيون، معهد جرادينت، معهد الديمقراطية الإسرائيلي، مؤسسة موزيلا، أولد وايز نيو، مؤسسة راند، سيفيراي، مركز المرونة طويلة الأمد، جمعية المستقبل، معهد آلان تورينج، الجمعية الملكية، جمعية سياسات الذكاء الاصطناعي في تركيا.

الصناعة: Advai، Anthropic، Cohere، Deloitte Consulting USA وDeloitte LLM Google DeepMind، G42، UK، Harmony Intelligence، Hugging Face، IBM، Lelapa AI، Meta، Microsoft، Shutterstock، Zhipu.ai.

شكر خاص

تعرب الأمانة العامة عن تقديرها للدعم والتعليقات والملاحظات المقدمة من أنجي عبد الله، ونيتارشان راجكومار، وجيفري إيرفينج، وشانون فالور، وريبيكا فينلاي، وأندرو سترايت.

إخلاء مسؤولية

لا يمثل التقرير آراء الرئيس، أو أي فرد معين في مجموعات الكتابة أو الفريق الاستشاري، أو أي من الحكومات التي دعمت إعداده. يعد هذا التقرير خلاصة الأبحاث الحالية حول قدرات الذكاء الاصطناعي المتقدم ومخاطره. ويتحمل رئيس التقرير المسؤولية النهائية عنه، ويشرف على تطويره من البداية إلى النهاية.

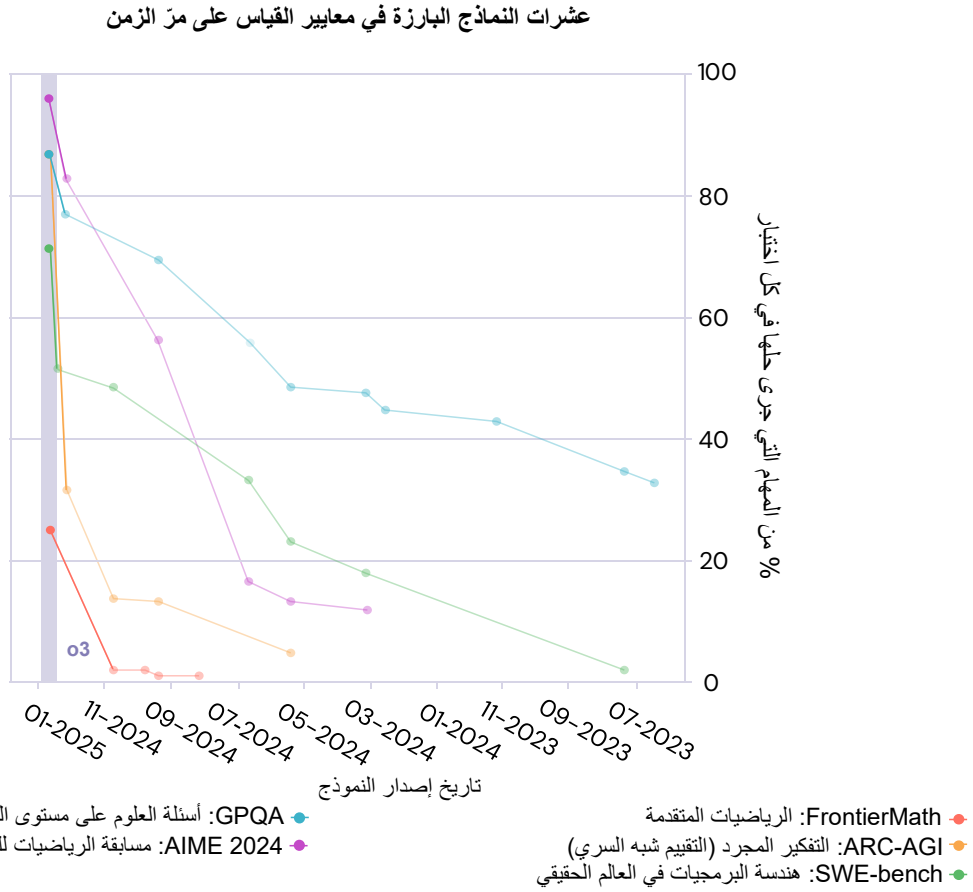
رقم سلسلة الأبحاث: DSIT 2025/001

حول هذا التقرير

- **هذا هو أول تقرير دولي حول سلامة الذكاء الاصطناعي.** بعد نشر التقرير المؤقت في مايو/أيار 2024، ساهمت مجموعة متنوعة تضم 96 خبيرًا في مجال الذكاء الاصطناعي في هذا التقرير الكامل الأول، بما في ذلك لجنة استشارية دولية من الخبراء رشحتها 30 دولة، ومنظمة التعاون الاقتصادي والتنمية، والاتحاد الأوروبي، والأمم المتحدة. ويهدف التقرير إلى توفير معلومات علمية من شأنها دعم صنع السياسات المستنيرة. لكنه لا يُوصي بصنع سياسات محددة.
- **هذا التقرير هو نتاج عمل خبراء مستقلين.** كان للخبراء المستقلين القائمين على كتابة هذا التقرير سلطة تقديرية كاملة نحو محتواه، وذلك بقيادة الرئيس.
- **ورغم أن هذا التقرير يتناول مخاطر الذكاء الاصطناعي وسلامته، فإنه يقدم أيضًا العديد من الفوائد المحتملة للأفراد والشركات والمجتمع.** هناك العديد من أنواع الذكاء الاصطناعي، ولكل منها فوائد ومخاطر مختلفة. في معظم الأحيان، وفي معظم التطبيقات، يساعد الذكاء الاصطناعي الأفراد والمؤسسات على أن يكونوا أكثر فعاليةً. ولكن الناس في مختلف أنحاء العالم لن يتمكنوا من التمتع بالعديد من الفوائد المحتملة للذكاء الاصطناعي بالكامل وبأمان إلا إذا جرت إدارة مخاطره على نحوٍ مناسب. يركز هذا التقرير على تحديد هذه المخاطر وتقييم أساليب تخفيف حدّتها. ولكنه لا يهدف إلى إجراء تقييم شامل لجميع التأثيرات المجتمعية المحتملة للذكاء الاصطناعي، بما في ذلك فوائده المحتملة العديدة.
- **يركز التقرير على الذكاء الاصطناعي ذو الأغراض العامة.** يقتصر التقرير على التركيز على نوع من الذكاء الاصطناعي الذي تطوَّر بسرعة خاصةً في السنوات الأخيرة، والذي لم تُدرس المخاطر المرتبطة به ولم تُفهم بشكل كافٍ: الذكاء الاصطناعي ذو الأغراض العامة، أو الذكاء الاصطناعي الذي يمكنه أداء مجموعة كبيرة من المهام. يركز التحليل في هذا التقرير على أنظمة الذكاء الاصطناعي العامة الأكثر تقدمًا في وقت كتابة هذا التقرير، بالإضافة إلى الأنظمة المستقبلية التي قد تكون أكثر قدرة.
- **يلخص التقرير الأدلة العلمية بشأن ثلاثة أسئلة أساسية:** ماذا يمكن للذكاء الاصطناعي ذو الأغراض العامة أن يفعل؟ ما هي المخاطر المرتبطة بالذكاء الاصطناعي ذو الأغراض العامة؟ وما هي التقنيات المستخدمة للتخفيف من حدة هذه المخاطر؟
- **الرهانات و المخاطر عديدة.** نحن الخبراء المساهمين في هذا التقرير، لا نزال نختلف حول العديد من الأسئلة، الصغيرة والكبيرة، حول قدرات الذكاء الاصطناعي ذو الأغراض العامة ومخاطره والتخفيف من حدّتها. ولكننا نعتبر هذا التقرير ضروريًا لتحسين فهمنا الجماعي لهذه التكنولوجيا ومخاطرها المحتملة. ونأمل أن يساعد التقرير المجتمع الدولي على التحرك نحو توافق أكبر بشأن الذكاء الاصطناعي ذو الأغراض العامة وتخفيف مخاطره على نحوٍ أكثر فعاليةً، حتى يتمكن الناس من تجربة فوائده المحتملة العديدة بأمان. الرهانات و المخاطر عديدة، لكننا نتطلع إلى مواصلة هذا الجهد.

تحديث حول أحدث التطورات في مجال الذكاء الاصطناعي بعد كتابة هذا التقرير: مذكرة الرئيس

في الفترة ما بين الانتهاء من كتابة هذا التقرير (5 ديسمبر/كانون الأول 2024) و نشره في يناير/كانون الثاني 2025، حدث تطور مهم. شاركت شركة الذكاء الاصطناعي OpenAI نتائج الاختبار المبكرة لنموذج الذكاء الاصطناعي الجديد، المعروف باسم o3. تشير هذه النتائج إلى أداء أقوى بشكل ملحوظ من أي نموذج سابق في عدد من الاختبارات الأكثر تحديًا في مجالات البرمجة والتفكير المجرد والتفكير العلمي. في بعض هذه الاختبارات، يتفوق o3 على العديد من الخبراء البشريين (ولكن ليس جميعهم). علاوة على ذلك، فقد حقق تقدمًا كبيرًا في اختبار التفكير المجرد الأساسي الذي كان العديد من الخبراء، ومن بينهم أنا، يعتقدون أنه كان بعيد المنال حتى وقت قريب. ولكن في وقت كتابة هذا التقرير، لا توجد معلومات عامة حول قدراته في العالم الحقيقي، وخاصة فيما يتعلق بحل المهام ذات النهايات المفتوحة و العديد من الحلول.



الشكل 0.1: نتائج نماذج الذكاء الاصطناعي ذو الأغراض العامة البارزة على معايير القياس الرئيسية من يونيو/حزيران 2023 إلى ديسمبر/كانون الأول 2024. أظهر o3 أداءً محسنًا بشكل كبير مقارنة بأحدث و أفضل النماذج السابقة (المنطقة المظللة). تُعد هذه المعايير من بين الاختبارات الأكثر تحديًا في مجالات البرمجة والتفكير المجرد والتفكير العلمي. بالنسبة لـ o3 الذي لم يُصدَر، يُعرَض تاريخ الإعلان عنه؛ أما بالنسبة للنماذج الأخرى، فيُعرَض تاريخ الإصدار. استفادت بعض نماذج الذكاء الاصطناعي الأحدث، بما في ذلك o3، من تحسين الهيكل الأساسي والمزيد من الحوسبة في وقت الاختبار. المصادر: Anthropic, 2024; Chollet, 2024; Chollet et al., 2025; Epoch AI, 2024; Glazer et al. 2024; OpenAI, 2024a; OpenAI, 2024b; Jimenez et al., 2024; Jimenez et al., 2025.

تُعد نتائج o3 دليلاً على أن وتيرة التقدم في قدرات الذكاء الاصطناعي قد تظل مرتفعة أو حتى أنها تتسارع. وبشكل أكثر تحديدًا، يقترح الباحثون أن إعطاء النماذج المزيد من الإمكانات الحوسبية لحل مشكلة معينة ("توسيع الاستدلال") قد يساعد في التغلب على القيود السابقة. بشكل عام، يؤدي توسيع الاستدلال إلى زيادة تكلفة استخدام النماذج. ولكن كنموذج مميز حديث آخر، هناك نموذج R1، الذي أصدرته شركة DeepSeek في يناير/كانون الثاني 2025، حيث أظهر أن الباحثين يعملون بنجاح على خفض هذه التكاليف. بشكل عام، قد يسمح توسيع الاستدلال لمطوري الذكاء الاصطناعي بتحقيق المزيد من التقدم في المستقبل. وتؤكد نتائج o3 أيضًا على الحاجة إلى فهم أفضل لكيفية تأثير استخدام مطوري الذكاء الاصطناعي المتزايد للذكاء الاصطناعي على سرعة تطوير الذكاء الاصطناعي نفسه.

إن التوجهات التي أثبتها o3 قد يكون لها آثار عميقة على مخاطر الذكاء الاصطناعي. فقد أدى التقدم في العلوم وقدرات البرمجة في السابق إلى توليد المزيد من الأدلة على المخاطر مثل الهجمات الإلكترونية والبيولوجية. وتتعلق نتائج o3 أيضًا بتأثيرات سوق العمل المحتملة، ومخاطر فقدان السيطرة، واستخدام الطاقة من بين أمور أخرى. ولكن من الممكن أيضًا استخدام قدرات o3 للمساعدة في الحماية من الأعطال والاستخدامات الضارة. وبشكل عام، ينبغي قراءة تقييمات المخاطر في هذا التقرير مع فهم حقيقة أن الذكاء الاصطناعي اكتسب قدرات كبيرة منذ كتابة التقرير. ولكن لا يوجد حتى الآن أي دليل حول تأثيرات o3 في العالم الواقعي، ولا توجد معلومات لتأكيد أو استبعاد المخاطر الرئيسية الجديدة و/أو الفورية.

إن التحسن في القدرات الذي تشير إليه نتائج o3 وفهمنا المحدود للآثار المترتبة على مخاطر الذكاء الاصطناعي يؤكدان على التحدي الرئيسي الذي يواجه صناع السياسات و الذي يحدده هذا التقرير: وسوف يتعين عليهم في كثير من الأحيان أن يزنوا الفوائد والمخاطر المحتملة للتقدم الوشيك في مجال الذكاء الاصطناعي دون وجود مجموعة كبيرة من الأدلة العلمية. ولكن، وعلى الرغم من ذلك، فإن توليد الأدلة بشأن الآثار المترتبة على السلامة والأمن للتوجهات التي ينطوي عليها o3 سيشكل أولوية ملحة لبحوث الذكاء الاصطناعي في الأسابيع والأشهر المقبلة.

النتائج الرئيسية للتقرير

- **لقد زادت قدرات الذكاء الاصطناعي ذو الأغراض العامة، وهو نوع الذكاء الاصطناعي الذي يركز عليه هذا التقرير، بسرعة كبيرة في السنوات الأخيرة وتحسنت بصورة أكبر في الأشهر الأخيرة.**[1] قبل بضع سنوات، كان من النادر أن تتمكن أفضل نماذج اللغة الكبيرة (LLMs) من إنتاج فقرة نصية متماسكة. اليوم، أصبح الذكاء الاصطناعي ذو الأغراض العامة قادرًا على كتابة برامج حاسوب، وإنشاء صور واقعية مُخصّصة، والانخراط في محادثات مفتوحة مُطوَّلة. منذ نشر التقرير المؤقت (مايو/أيار 2024)، أظهرت النماذج الجديدة أداءً أفضل بصورة ملحوظة في اختبارات التفكير العلمي والبرمجة.
- **تستثمر العديد من الشركات الآن في تطوير وكلاء الذكاء الاصطناعي ذو الأغراض العامة، كتوجُّه محتمل لإحراز مزيد من التقدم.** وكلاء الذكاء الاصطناعي هم أنظمة ذكاء اصطناعي ذو الأغراض العامة يمكنها العمل والتخطيط والتفويض باستقلال لتحقيق الأهداف مع القليل من الإشراف البشري أو بدونه. وسيكون وكلاء الذكاء الاصطناعي المتطورون قادرين، على سبيل المثال، على استخدام أجهزة الحاسوب لإكمال مشروعات أطول من الأنظمة الحالية، ما يفتح الباب أمام الحصول على فوائد إضافية ومواجهة مخاطر إضافية.
- **إن التطورات المستقبلية في القدرات في الأشهر والسنوات القادمة قد تكون بطيئة أو سريعة للغاية.**[†] وسوف يعتمد التقدم على ما إذا كانت الشركات ستتمكن من نشر المزيد من البيانات وقوة الحوسبة بسرعة لتدريب نماذج جديدة، وما إذا كان "توسُّع" النماذج بهذه الطريقة سيقود إلى التغلب على قيودها الحالية. تشير الأبحاث الحديثة إلى أن توسيع نطاق النماذج سريعة التوسُّع قد يظل ممكن لعدة سنوات على الأقل. ولكن التقدم الكبير في القدرات قد يتطلب أيضًا عوامل أخرى: على سبيل المثال، التفرات البحثية الجديدة، والتي يصعب التنبؤ بها، أو نجاح نهج التوسع الجديد الذي تبنته الشركات مؤخرًا.
- **لقد ثبت بالفعل وجود العديد من الأضرار الناجمة عن استخدام الذكاء الاصطناعي ذو الأغراض العامة.** وتشمل هذه الأضرار الاحتيال، والصور الحميمة غير المقبولة (NCII)، ومواد الاعتداء الجنسي على الأطفال (CSAM)، ومخرجات النماذج المتحيزة ضد مجموعات معينة من الأشخاص أو آراء معينة وقضايا الموثوقية، وانتهاكات الخصوصية. وقد طوّر الباحثون تقنيات التخفيف للحدة من هذه المشكلات، ولكن حتى الآن لم تتمكن أي مجموعة من التقنيات من حلها بالكامل. منذ نشر التقرير المؤقت (مايو/أيار 2024)، كشفت أدلة جديدة عن التمييز المتعلق بأنظمة الذكاء الاصطناعي ذو الأغراض العامة عن أشكال أكثر دقة من التحيز.
- **ومع تزايد قدرة الذكاء الاصطناعي ذو الأغراض العامة، بدأت تظهر تدريجيًا أدلة على وجود مخاطر إضافية.** وتشمل هذه المخاطر تأثيرات سوق العمل واسعة النطاق، أو الاختراق المدعوم بالذكاء الاصطناعي أو الهجمات البيولوجية، وفقدان المجتمع للتحكم بالذكاء الاصطناعي ذو الأغراض العامة. ويفسر الخبراء الأدلة الموجودة بشأن هذه المخاطر على نحوٍ مختلف: يعتقد البعض أن مثل هذه المخاطر قد لا تظهر إلا بعد عقود من الزمن، في حين يعتقد آخرون أن الذكاء الاصطناعي ذو الأغراض العامة قد يؤدي إلى أضرار على نطاق المجتمع في غضون السنوات القليلة المقبلة. وقد أدت التطورات الأخيرة في قدرات الذكاء الاصطناعي ذو الأغراض العامة، وخاصة في اختبارات المنطق العلمي والبرمجة، إلى ظهور أدلة جديدة على المخاطر المحتملة مثل القرصنة المدعومة بالذكاء الاصطناعي والهجمات البيولوجية، ما دفع إحدى شركات الذكاء الاصطناعي الكبرى إلى رفع تقييمها للمخاطر البيولوجية من أفضل نموذج لديها من "منخفض" إلى "متوسط".
- **إن تقنيات إدارة المخاطر ما زالت في بداياتها، ولكن التقدم ممكن.** هناك طرق تقنية مختلفة لتقييم المخاطر الناجمة عن الذكاء الاصطناعي ذو الأغراض العامة وتقليلها، والتي يمكن للمطورين استخدامها ويمكن للجهات التنظيمية أن تطلبها، ولكن جميعها لها حدود. على سبيل المثال، تظل تقنيات التفسير الحالية لشرح سبب إنتاج نموذج الذكاء الاصطناعي ذو الأغراض العامة لأي ناتج معين محدودة للغاية. ولكن يحرز الباحثون بعض التقدم في معالجة هذه القيود. وعلاوة على ذلك، يحاول الباحثون وصُناع السياسات بصورة متزايدة توحيد أساليب إدارة المخاطر، والتنسيق على المستوى الدولي.
- **إن وتيرة التقدم في مجال الذكاء الاصطناعي ذو الأغراض العامة وعدم القدرة على التنبؤ به يشكلان "معضلة أدلة" لصناع السياسات.**[†] ونظرًا للتقدم السريع وغير المتوقع في بعض الأحيان، فسيتعين على صناع السياسات في كثير من الأحيان أن يزنوا الفوائد والمخاطر المحتملة للتقدم الوشيك في مجال الذكاء الاصطناعي دون وجود مجموعة كبيرة من الأدلة العلمية المتاحة. وعند قيامهم بذلك، فإنهم يواجهون معضلة. فمن ناحية، قد يتبين أن التدابير الوقائية للتخفيف من المخاطر التي تعتمد على أدلة محدودة غير فعالة أو غير ضرورية. ومن ناحية أخرى، فإن الانتظار للحصول على أدلة أقوى على المخاطر الوشيكة قد يترك المجتمع غير مستعد أو حتى يجعل التخفيف من المخاطر مستحيلًا، على سبيل المثال إذا حدثت طفرات مفاجئة في قدرات الذكاء الاصطناعي، والمخاطر المرتبطة بها. وتعمل الشركات والحكومات على تطوير أنظمة الإنذار المبكر وأطر

[1] يرجى الرجوع إلى تحديث الرئيس بشأن أحدث التطورات في مجال الذكاء الاصطناعي بعد كتابة هذا التقرير.

[†] يرجى الرجوع إلى تحديث الرئيس بشأن أحدث التطورات في مجال الذكاء الاصطناعي بعد كتابة هذا التقرير.

إدارة المخاطر التي قد تساعد في الحدّ من هذه المعضلة. وتؤدي بعض هذه التشريعات إلى إطلاق تدابير تخفيف محددة عندما تظهر أدلة جديدة على وجود مخاطر، في حين تتطلب تشريعات أخرى من المطورين تقديم أدلة على السلامة قبل إصدار نموذج جديد.

- **هناك إجماع واسع النطاق بين الباحثين على أن التقدم فيما يتعلق بالأسئلة التالية سيكون مفيدًا:** ما مدى سرعة تقدم قدرات الذكاء الاصطناعي ذو الأغراض العامة في السنوات القادمة، وكيف يمكن للباحثين قياس هذا التقدم بصورة موثوقة؟ ما هي عتبات المخاطر المعقولة التي يجب أن نتخذها لتحفيز إجراءات التخفيف؟ كيف يمكن لصناع السياسات الوصول إلى المعلومات المتعلقة بالذكاء الاصطناعي ذو الأغراض العامة والتي تتعلق بالسلامة العامة بصورة أفضل؟ كيف يمكن للباحثين وشركات التكنولوجيا والحكومات تقييم مخاطر تطوير الذكاء الاصطناعي ذو الأغراض العامة ونشرها بصورة موثوقة؟ كيف تعمل نماذج الذكاء الاصطناعي ذو الأغراض العامة داخليًا؟ كيف يمكن تصميم الذكاء الاصطناعي ذو الأغراض العامة ليتصرف بصورة موثوقة؟
- **الذكاء الاصطناعي لا يحدث لنا: الاختيارات التي يتخذها الناس تحدد مستقبلهم.** إن مستقبل تكنولوجيا الذكاء الاصطناعي ذو الأغراض العامة غير مؤكد، مع وجود مجموعة واسعة من المسارات التي تبدو ممكنة حتى في المستقبل القريب، بما في ذلك النتائج الإيجابية للغاية والسلبية للغاية. إن هذا الغموض يمكن أن يثير القدرية ويجعل الذكاء الاصطناعي يبدو وكأنه شيء يحدث لنا. ولكن قرارات المجتمعات والحكومات بشأن كيفية التعامل مع حالة عدم اليقين هذه هي التي ستحدد المسار الذي سنختاره. ويهدف هذا التقرير إلى تسهيل المناقشة البناءة والمبنية على الأدلة حول هذه القرارات.

الملخص التنفيذي

غرض هذا التقرير

يقدم هذا التقرير ملخصًا لحالة الفهم العلمي للذكاء الاصطناعي ذو الأغراض العامة، الذكاء الاصطناعي الذي يمكنه أداء مجموعة واسعة من المهام، مع التركيز على فهم مخاطره وإدارتها.

يلخص هذا التقرير الأدلة العلمية حول سلامة الذكاء الاصطناعي ذو الأغراض العامة. يهدف هذا التقرير إلى المساعدة في خلق فهم دولي مشترك للمخاطر الناجمة عن الذكاء الاصطناعي المتقدم وكيفية التخفيف من حدّتها. ولتحقيق هذه الغاية، يركز هذا التقرير على الذكاء الاصطناعي ذو الأغراض العامة، أو الذكاء الاصطناعي الذي يمكنه أداء مجموعة واسعة من المهام، لأن هذا النوع من الذكاء الاصطناعي تقدم بسرعة خاصة في السنوات الأخيرة ونُشر على نطاق واسع من قِبل شركات التكنولوجيا لمجموعة من الأغراض الاستهلاكية والتجارية. يقدم التقرير ملخصًا لحالة الفهم العلمي للذكاء الاصطناعي ذو الأغراض العامة، مع التركيز على فهم مخاطره وإدارتها.

في خِضَم التقدم السريع، أصبح البحث في مجال الذكاء الاصطناعي للأغراض العامة حاليًا في مرحلة اكتشاف علمي، وفي العديد من الحالات لم يستقر كعلم بعد. يقدم التقرير لمحة عامة عن الفهم العلمي الحالي للذكاء الاصطناعي ذو الأغراض العامة ومخاطره. ويتضمن ذلك تحديد مجالات الإجماع العلمي والمجالات التي توجد فيها وجهات نظر مختلفة أو فجوات في الفهم العلمي الحالي.

لن يتمكن الأشخاص في جميع أنحاء العالم من التمتع بالفوائد المحتملة للذكاء الاصطناعي ذو الأغراض العامة بالكامل بأمان إلا إذا جرت إدارة مخاطره بصورة مناسبة. يركز هذا التقرير على تحديد هذه المخاطر وتقييم الأساليب الفنية لتقييمها وتخفيفها، بما في ذلك الطرق التي يمكن من خلالها استخدام الذكاء الاصطناعي ذو الأغراض العامة للتخفيف من المخاطر. لا يهدف هذا التقرير إلى إجراء تقييم شامل لجميع التأثيرات المجتمعية المحتملة للذكاء الاصطناعي ذو الأغراض العامة. الأمر الأكثر أهمية هو الفوائد الحالية والمحتملة في المستقبل من الذكاء الاصطناعي ذو الأغراض العامة على الرغم من اتساع نطاقه يقع خارج نطاق هذا التقرير. ويتطلب صنع السياسات الشاملة النظر في كل من الفوائد المحتملة للذكاء الاصطناعي ذو الأغراض العامة والمخاطر التي يغطيها هذا التقرير. وهذا يتطلب أيضًا الأخذ في الاعتبار أن الأنواع الأخرى من الذكاء الاصطناعي لها ملفات تعريف مختلفة للمخاطر والفوائد مقارنة بالذكاء الاصطناعي للأغراض العامة الحالي.

تتضمن الأقسام الثلاثة الرئيسية للتقرير ملخصًا للأدلة العلمية بشأن ثلاثة أسئلة أساسية: ما الذي بإمكان الذكاء الاصطناعي ذو الأغراض العامة أن يفعله؟ ما هي المخاطر المرتبطة بالذكاء الاصطناعي ذو الأغراض العامة؟ وما هي التقنيات المُستخدمة للتخفيف من حدّة هذه المخاطر؟

القسم الأول - قدرات الذكاء الاصطناعي ذو الأغراض العامة: ما الذي بإمكان الذكاء الاصطناعي ذو الأغراض العامة أن يفعله الآن وفي المستقبل؟

لقد تحسنت قدرات الذكاء الاصطناعي ذو الأغراض العامة بسرعة في السنوات الأخيرة، ويمكن أن تتراوح التطورات الإضافية من البطيئة إلى السريعة للغاية.

إن ما يمكن أن يفعله الذكاء الاصطناعي يُعد مساهم رئيسي في العديد من المخاطر التي يفرضها، ووفقًا للعديد من المقاييس، فقد كانت قدرات الذكاء الاصطناعي ذو الأغراض العامة تتقدم بسرعة. منذ خمس سنوات، نادرًا ما كانت نماذج لغة الذكاء الاصطناعي ذو الأغراض العامة قادرة على إنتاج فقرة نصية متماسكة. اليوم، يمكن لبعض نماذج الذكاء الاصطناعي ذو الأغراض العامة المشاركة في محادثات حول مجموعة واسعة من الموضوعات، أو كتابة برامج حاسوب، أو إنشاء مقاطع فيديو قصيرة واقعية من الوصف. ومع ذلك، فمن الصعب من الناحية الفنية تقدير قدرات الذكاء الاصطناعي ذو الأغراض العامة ووصفها بصورة موثوقة.

لقد حسّن مطورو الذكاء الاصطناعي قدرات الذكاء الاصطناعي ذو الأغراض العامة بصورة سريعة في السنوات الأخيرة، وذلك في الغالب من خلال "التوسع".† لقد عملوا دون كلل على زيادة الموارد المستخدمة لتدريب النماذج الجديدة (غالبًا ما يُشار إلى ذلك باسم "التوسع") وصقلوا الأساليب الحالية لاستخدام تلك الموارد بكفاءة أكبر. على سبيل المثال، وفقًا لتقديرات حديثة، شهدت نماذج الذكاء الاصطناعي الحديثة زيادات سنوية تبلغ نحو 4 أضعاف في الموارد الحسابية ("الحوسبة") المستخدمة للتدريب و2.5 ضعف في حجم مجموعة بيانات التدريب.

إن وتيرة التقدم المستقبلي في قدرات الذكاء الاصطناعي ذو الأغراض العامة لها آثار كبيرة على إدارة المخاطر الناشئة، لكن الخبراء يختلفون حول ما يمكن توقعه حتى في الأشهر والسنوات المقبلة. يدعم الخبراء بتفاوت إمكانية تقدم قدرات الذكاء الاصطناعي ذو الأغراض العامة ببطء، أو بسرعة، أو بسرعة كبيرة للغاية.

ويختلف الخبراء بشأن وتيرة التقدم المستقبلي بسبب وجهات نظر مختلفة بشأن الوعد بمزيد من "التوسع" وتستكشف الشركات نوعًا جديدًا إضافيًا من التوسع من شأنه أن يعمل على تسريع القدرات بشكل أكبر.† وفي حين أن التوسع نجح في كثير من الأحيان في التغلب على قيود الأنظمة السابقة، فإن الخبراء يختلفون حول إمكاناته في حل القيود المتبقية لأنظمة اليوم، مثل عدم الموثوقية في التصرف في العالم المادي وفي تنفيذ مهام ممتدة على أجهزة الحاسوب. في الأشهر الأخيرة، أظهر نوع جديد من التوسع إمكانية تحسين القدرات بشكل أكبر: بدلاً من مجرد زيادة حجم الموارد المستخدمة في نماذج التدريب، أصبحت شركات الذكاء الاصطناعي مهتمة بشكل متزايد أيضًا بـ "توسيع الاستدلال" السماح لنموذج مدرب بالفعل باستخدام المزيد من الحساب لحل مشكلة معينة، على سبيل المثال لتحسين حله الخاص، أو كتابة ما يسمى "سلاسل الفكر" التي تقسم المشكلة إلى خطوات أبسط.

تراهن العديد من الشركات الرائدة التي تعمل على تطوير الذكاء الاصطناعي ذو الأغراض العامة على "التوسع" لمواصلة تحسين الأداء. إذا استمرت الاتجاهات الأخيرة، بحلول نهاية عام 2026 سيتم تدريب بعض نماذج الذكاء الاصطناعي ذو الأغراض العامة باستخدام حوسبة تدريبية أكثر بحوالي 100 مرة من النماذج الأكثر كثافة في الحوسبة في عام 2023، مع النمو إلى حوسبة تدريبية أكثر بمقدار 10000 مرة بحلول عام 2030، جنبًا إلى جنب مع الخوارزميات التي تحقق قدرات أكبر لكمية معينة من الحوسبة المتاحة. بالإضافة إلى هذا التوسع المحتمل لموارد التدريب، فإن الاتجاهات الحديثة مثل التوسع الاستدلالي واستخدام النماذج لإنتاج بيانات التدريب قد تعني أنه سيجري استخدام المزيد من الحوسبة بصورة عامة. ومع ذلك، هناك اختناقات محتملة أمام زيادة البيانات والحوسبة بسرعة أكبر، مثل توافر البيانات، وشرائح الذكاء الاصطناعي، ورأس المال، والقدرة المحلية على الطاقة. تعمل الشركات التي تساهم في تطوير الذكاء الاصطناعي ذو الأغراض العامة على التغلب على هذه الاختناقات المحتملة.

منذ نشر التقرير المؤقت (مايو/أيار 2024)، وصل الذكاء الاصطناعي ذو الأغراض العامة إلى مستوى الأداء الخبير في بعض الاختبارات والمسابقات الخاصة بالمنطق العلمي والبرمجة، وتبذل الشركات جهودًا كبيرة لتطوير وكلاء الذكاء الاصطناعي المستقلين. وقد كان التقدم في العلوم والبرمجة مدفوعًا بتقنيات توسع الاستدلال مثل كتابة "سلاسل الأفكار" الطويلة. وتشير دراسات جديدة إلى أن توسيع نطاق مثل هذه الأساليب، على سبيل المثال السماح للنماذج بتحليل المشكلات من خلال كتابة سلاسل فكرية أطول من نماذج اليوم، قد يؤدي إلى تحقيق المزيد من التقدم في المجالات التي يكون فيها التفكير هو الجانب الأكثر أهميةً، مثل العلوم، وهندسة البرمجيات، والتخطيط. بالإضافة إلى هذا الاتجاه، تبذل الشركات جهودًا مضنية لتطوير وكلاء الذكاء الاصطناعي الأكثر تقدمًا ذو الأغراض العامة، والذين يمكنهم التخطيط والعمل باستقلالية لتحقيق هدف معين. أخيرًا، انخفض سعر السوق لاستخدام الذكاء الاصطناعي ذو الأغراض العامة بمستوى قدرة معين بشكل حاد، مما يسهل الوصول لهذه التكنولوجيا واستخدامها على نطاق واسع.

يركز هذا التقرير في المقام الأول على الجوانب التقنية لتقدم الذكاء الاصطناعي، ولكن مدى سرعة تقدم الذكاء الاصطناعي ذو الأغراض العامة ليس مسألة تقنية بحتة. وسوف تعتمد وتيرة التقدم في المستقبل أيضًا على عوامل غير تقنية، بما في ذلك ربما النهج الذي تتبعه الحكومات لتنظيم الذكاء الاصطناعي. لا يناقش هذا التقرير كيف يمكن أن تؤثر الأساليب المختلفة للتنظيم على سرعة تطوير و اعتماد الذكاء الاصطناعي للأغراض العامة.

† يرجى الرجوع إلى تحديث الرئيس بشأن أحدث التطورات في مجال الذكاء الاصطناعي بعد كتابة هذا التقرير.

القسم الثاني – المخاطر: ما هي المخاطر المرتبطة بالذكاء الاصطناعي ذو الأغراض العامة؟

لقد ثبت بالفعل وجود العديد من الأضرار الناجمة عن استخدام الذكاء الاصطناعي ذو الأغراض العامة. ومع تزايد قدرة الذكاء الاصطناعي ذو الأغراض العامة، بدأت تظهر تدريجيًا أدلة على وجود مخاطر إضافية.

يصنف هذا التقرير مخاطر الذكاء الاصطناعي العامة إلى ثلاث فئات: مخاطر الاستخدام الضار، ومخاطر الأعطال، والمخاطر النظامية. تحتوي كل فئة من هذه الفئات على مخاطر حدثت بالفعل بالإضافة إلى مخاطر قد تحدث في السنوات القليلة المقبلة.

المخاطر الناجمة عن الاستخدام الضار: يمكن للجهات الخبيثة استخدام الذكاء الاصطناعي ذو الأغراض العامة لإلحاق الضرر بالأفراد أو المؤسسات أو المجتمع. تشمل أشكال الاستخدام الضار ما يلي:

- **الضرر الذي يلحق بالأفراد من خلال المحتوى المزيف يمكن للجهات الخبيثة حاليًا استخدام الذكاء الاصطناعي ذو الأغراض العامة لإنشاء محتوى مزيف يضر بالأفراد بطريقة مستهدفة.** وتشمل هذه الاستخدامات الخبيثة المواد الإباحية المزيفة غير المقبولة، والمواد الإباحية الجنسية المُنشأة بواسطة الذكاء الاصطناعي، والاحتيال المالي من خلال انتحال الصوت، والابتزاز من أجل الابتزاز، وتخريب السمعة الشخصية والمهنية، والإساءة النفسية. ومع ذلك، في حين أن تقارير الحوادث المتعلقة بالأضرار الناجمة عن المحتوى المزيف المُنشأ بواسطة الذكاء الاصطناعي شائعة، إلا أن الإحصائيات الموثوقة بشأن وتيرة هذه الحوادث لا تزال غير متوفرة.
- **التلاعب بالرأي العام يسهل الذكاء الاصطناعي ذو الأغراض العامة إنشاء محتوى مقنع على نطاق واسع.** ويمكن أن يساعد هذا الجهات الفاعلة التي تسعى إلى التلاعب بالرأي العام، على سبيل المثال للتأثير على النتائج السياسية. ومع ذلك، فإن الأدلة على مدى انتشار هذه الجهود وفعاليتها لا تزال محدودة. على الرغم من أن التدابير التقنية المضادة مثل وضع العلامات المائية على المحتوى مفيدة، إلا أنه يمكن عادةً التحايل عليها من قبل جهات متطورة إلى حدٍ ما.
- **الجرائم الإلكترونية:** يمكن للذكاء الاصطناعي ذو الأغراض العامة أن يجعل من الأسهل أو الأسرع بالنسبة للجهات الخبيثة ذات مستويات المهارة المختلفة تنفيذ الهجمات الإلكترونية. لقد أثبتت الأنظمة الحالية قدراتها في مهام الأمن السيبراني منخفضة ومتوسطة التعقيد، وتقوم الجهات الفاعلة التي ترعاها الدولة باستكشاف الذكاء الاصطناعي بنشاط لمسح الأنظمة المستهدفة. وأكدت أبحاث جديدة أن قدرات الذكاء الاصطناعي ذو الأغراض العامة المتعلقة بالجرائم الإلكترونية تتقدم بشكل كبير، لكن لا يزال من غير الواضح ما إذا كان هذا سيؤثر في التوازن بين المهاجمين والمدافعين.
- **الهجمات البيولوجية والكيميائية:** لقد أظهرت أنظمة الذكاء الاصطناعي ذو الأغراض العامة الحديثة بعض القدرة على توفير التعليمات والتوجيه لاستكشاف الأخطاء وإصلاحها لإعادة إنتاج الأسلحة البيولوجية والكيميائية المعروفة وتسهيل تصميم المركبات السامة الجديدة. في تجارب جديدة اختبرت القدرة على وضع خطط لإنتاج أسلحة بيولوجية، كان أداء نظام الذكاء الاصطناعي ذو الأغراض العامة في بعض الأحيان أفضل من الخبراء البشريين الذين لديهم إمكانية الوصول إلى الإنترنت. وردًا على ذلك، رفعت إحدى شركات الذكاء الاصطناعي تقييمها للمخاطر البيولوجية من أفضل نموذج لديها من "منخفض" إلى "متوسط". ومع ذلك، فإن المحاولات الحقيقية لتطوير مثل هذه الأسلحة لا تزال تتطلب موارد وخبرات إضافية كبيرة. يُعد إجراء تقييم شامل للمخاطر البيولوجية والكيميائية أمرًا صعبًا لأن الكثير من الأبحاث ذات الصلة سرية.

منذ نشر التقرير المؤقت (مايو/أيار 2024)، أصبح الذكاء الاصطناعي ذو الأغراض العامة أكثر قدرةً في المجالات ذات الصلة بالاستخدام الضار. على سبيل المثال، طوّر الباحثون مؤخرًا أنظمة ذكاء اصطناعي متعددة الأغراض كانت قادرة على اكتشاف بعض نقاط الضعف في الأمن السيبراني واستغلالها سواء بشكل مستقل أو بمساعدة بشرية، اكتشاف ثغرة غير معروفة سابقًا في البرامج المستخدمة على نطاق واسع. كما تحسنت أيضًا قدرات الذكاء الاصطناعي ذو الأغراض العامة المتعلقة بالاستدلال ودمج أنواع مختلفة من البيانات، والتي يمكن أن تساعد في البحث عن مسببات الأمراض أو في مجالات أخرى ذات استخدام مزدوج.

المخاطر الناجمة عن الأعطال: يمكن للذكاء الاصطناعي ذو الأغراض العامة أيضًا أن يسبب ضررًا غير مقصود. حتى عندما لا يكون لدى المستخدمين أي نية للتسبب في ضرر، فقد تنشأ مخاطر خطيرة بسبب خلل في أداء الذكاء الاصطناعي ذو الأغراض العامة. وتشمل هذه الأعطال ما يلي:

- **قضايا الموثوقية:** قد يكون الذكاء الاصطناعي العام الحالي غير موثوق به، ما قد يؤدي إلى إحداث الضرر. على سبيل المثال، إذا استشار المستخدمون نظام الذكاء الاصطناعي ذو الأغراض العامة للحصول على نصيحة طبية أو قانونية، فقد يُنشئ النظام إجابات تحتوي على أكاذيب. في كثير من الأحيان لا يكون المستخدمون على دراية بالقيود المفروضة على منتج الذكاء الاصطناعي، على سبيل المثال بسبب " أمية الذكاء الاصطناعي" ، أو الإعلانات المضللة، أو سوء التواصل. هناك عدد من الحالات المعروفة للضرر الناجم عن مشكلات الموثوقية، ولكن لا تزال هناك أدلة محدودة حول مدى انتشار أشكال مختلفة من هذه المشكلة.
- **يمكن لأنظمة الذكاء الاصطناعي المخصص ذو الأغراض العامة أن تعمل على تضخيم التحيزات الاجتماعية والسياسية، مما يتسبب في أضرار ملموسة.** فهي تُظهر في كثير من الأحيان تحيزات تتعلق بالعرق، أو الجنس، أو الثقافة، أو العمر، أو الإعاقة، أو الرأي السياسي، أو جوانب أخرى من الهوية الإنسانية. ويمكن أن يؤدي هذا إلى نتائج تمييزية بما في ذلك عدم المساواة في تخصيص الموارد، وتعزيز الصور النمطية، والإهمال المنهجي للمجموعات أو وجهات النظر غير الممثلة. تتطور الأساليب التقنية للتخفيف من التحيز والتمييز في أنظمة الذكاء الاصطناعي ذو الأغراض العامة، ولكنها تواجه مقايضات بين التخفيف من التحيز والأهداف المتنافسة مثل الدقة والخصوصية، فضلًا عن تحديات أخرى.
- **إن سيناريوهات فقدان السيطرة هي سيناريوهات مستقبلية افتراضية حيث يعمل نظام أو أكثر من أنظمة الذكاء الاصطناعي المخصص ذو الأغراض العامة خارج سيطرة أي شخص، دون وجود مسار واضح لاستعادة السيطرة.** هناك إجماع واسع النطاق على أن الذكاء الاصطناعي للأغراض العامة الحالي يفتقر إلى القدرات اللازمة لتشكيل هذا الخطر. ومع ذلك، فإن آراء الخبراء بشأن احتمال فقدان السيطرة خلال السنوات القليلة المقبلة تختلف اختلافًا كبيرًا: يعتبره البعض أمرًا غير معقول، ويعتبره البعض الآخر محتمل الحدوث، ويرى البعض الآخر أنه خطر ذو احتمالية متواضعة يستحق الاهتمام بسبب شدّته المحتملة العالية. إن الأبحاث التجريبية والرياضية المستمرة تعمل على احتدام هذه المجادلات تدريجيًا.

منذ نشر التقرير المؤقت (مايو/أيار 2024)، أدت الأبحاث الجديدة إلى التوصل إلى بعض الأفكار الجديدة حول مخاطر التحيز وفقدان السيطرة. لقد زادت الأدلة على التحيز في أنظمة الذكاء الاصطناعي ذو الأغراض العامة، وقد كشفت الأعمال الأخيرة عن أشكال إضافية من تحيز الذكاء الاصطناعي. وقد لاحظ الباحثون تقدمًا متواضعًا آخر نحو قدرات الذكاء الاصطناعي التي من المرجح أن تكون ضرورية لحدوث سيناريوهات فقدان السيطرة التي تجري مناقشتها بشكل شائع. وتشمل هذه القدرات القدرة على استخدام أجهزة الحاسوب بصورة مستقلة، والبرمجة، والحصول على وصول غير مصرح به إلى الأنظمة الرقمية، وتحديد طرق التهرب من الرقابة البشرية.

المخاطر النظامية: بالإضافة إلى المخاطر التي تفرضها قدرات النماذج الفردية مباشرةً، فإن التشغيل واسع النطاق للذكاء الاصطناعي ذو الأغراض العامة يرتبط بعدة مخاطر نظامية أوسع نطاقاً. وتتراوح أمثلة المخاطر النظامية من التأثيرات المحتملة على سوق العمل إلى مخاطر الخصوصية والتأثيرات البيئية:

- **مخاطر سوق العمل:** يتمتع الذكاء الاصطناعي ذو الأغراض العامة، وخاصةً إذا استمر في التقدم السريع، بالقدرة على أتمتة مجموعة واسعة جدًا من المهام، ما قد يكون له تأثير كبير على سوق العمل. وهذا يعني أن العديد من الأشخاص قد يفقدون وظائفهم الحالية. ومع ذلك، يتوقع العديد من خبراء الاقتصاد أن يكون من الممكن تعويض خسائر الوظائف المحتملة، جزئيًا أو ربما كليًا، من خلال توفير فرص عمل جديدة وزيادة الطلب في القطاعات غير الآلية.
- **فجوة البحث والتطوير المتعلق بالذكاء الاصطناعي العالمية يتركز البحث والتطوير في مجال الذكاء الاصطناعي ذو الأغراض العامة حاليًا في عدد قليل من الدول الغربية والصين.** إن هذه "الفجوة في الذكاء الاصطناعي" لديها القدرة على زيادة اعتماد العالم بشكل كبير على هذه المجموعة الصغيرة من البلدان. ويتوقع بعض الخبراء أيضًا أن تساهم في زيادة عدم المساواة على مستوى العالم. هناك العديد من الأسباب لهذه الفجوة، بما في ذلك عدد من الأسباب التي لا تُعد فريدة من نوعها بالنسبة للذكاء الاصطناعي. ومع ذلك، فإن جزءًا كبيرًا من ذلك ينبع من مستويات مختلفة من الوصول إلى الحوسبة الباهظة الثمن اللازمة لتطوير الذكاء الاصطناعي ذو الأغراض العامة: تتمتع معظم البلدان المنخفضة والمتوسطة الدخل بإمكانية وصول أقل بكثير إلى الحوسبة مقارنةً بالبلدان المرتفعة الدخل.
- **تركيز السوق ونقاط الفشل المتفردة تهيمن حاليًا مجموعة صغيرة من الشركات على سوق الذكاء الاصطناعي ذو الأغراض العامة.** ومن الممكن أن يؤدي تركيز السوق هذا إلى جعل المجتمعات أكثر عرضةً للعديد من المخاطر النظامية. على سبيل المثال، إذا كانت المنظمات في مختلف القطاعات الحيوية، مثل القطاع المالي أو الرعاية الصحية، تعتمد جميعها على عدد صغير من أنظمة الذكاء الاصطناعي ذو الأغراض العامة، فإن وجود خلل أو ثغرة أمنية في مثل هذا النظام قد يتسبب في حدوث أعطال وانقطاعات متزامنة على نطاق واسع.

- **المخاطر البيئية:** لقد أدى الاستخدام المتزايد للحوسبة في تطوير الذكاء الاصطناعي ذو الأغراض العامة ونشره إلى زيادة سريعة في كميات الطاقة والمياه والمواد الخام المستهلكة في بناء البنية التحتية للحوسبة اللازمة وتشغيلها. لا يظهر هذا الاتجاه أي مؤشر واضح على التباطؤ، على الرغم من التقدم في التقنيات التي تسمح باستخدام الحوسبة بكفاءة أكبر. كما أن الذكاء الاصطناعي ذو الأغراض العامة له عدد من التطبيقات التي يمكن أن تفيد أو تضر بجهود الاستدامة.
- **مخاطر الخصوصية:** يمكن للذكاء الاصطناعي ذو الأغراض العامة أن يتسبب في انتهاك خصوصية المستخدم أو يساهم في ذلك. على سبيل المثال، قد تتسرب المعلومات الحساسة الموجودة في بيانات التدريب دون عمد عندما يتفاعل المستخدم مع النظام. بالإضافة إلى ذلك، عندما يشارك المستخدمون معلومات حساسة مع النظام، يمكن أن تتسرب هذه المعلومات أيضًا. ولكن الذكاء الاصطناعي ذو الأغراض العامة قد يسهل أيضًا الانتهاكات المتعمدة للخصوصية، على سبيل المثال إذا استغلت الجهات الخبيثة الذكاء الاصطناعي لاستنتاج معلومات حساسة حول أفراد محددين باستخدام كميات كبيرة من البيانات. ومع ذلك، لم يجد الباحثون حتى الآن أي دليل على انتهاكات الخصوصية واسعة النطاق المرتبطة بالذكاء الاصطناعي ذو الأغراض العامة.
- **انتهاكات حقوق الطبع والنشر:** يتعلم الذكاء الاصطناعي ذو الأغراض العامة من أعمال التعبير الإبداعي وينشئها، ما يُشكّل تحديًا للأنظمة التقليدية للموافقة على البيانات، والتعويضات والتحكم. قد تتضمن عملية جمع البيانات وإنشاء المحتوى مجموعة متنوعة من قوانين حقوق البيانات، والتي تختلف عبر الولايات القضائية وقد تكون قيد التقاضي النشط. ونظراً لعدم اليقين القانوني بشأن ممارسات جمع البيانات، تقلل شركات الذكاء الاصطناعي من المعلومات التي تفصح عنها بشأن البيانات التي تستخدمها. يؤدي هذا التعتيم إلى زيادة الصعوبة في إجراء أبحاث سلامة الذكاء الاصطناعي بواسطة جهات خارجية.

منذ نشر التقرير المؤقت (مايو/أيار 2024)، ظهرت أدلة إضافية بشأن تأثيرات الذكاء الاصطناعي ذو الأغراض العامة في سوق العمل، في حين أدت التطورات الجديدة إلى تفاقم المخاوف المتعلقة بالخصوصية وحقوق النشر. تشير التحليلات الجديدة لبيانات سوق العمل إلى أن الأفراد يتبنون الذكاء الاصطناعي ذو الأغراض العامة بسرعة كبيرة مقارنةً بالتكنولوجيات السابقة. تختلف وتيرة تبني الشركات للتقنيات الجديدة بشكل كبير حسب القطاع. وعلاوة على ذلك، أدت التطورات الأخيرة في القدرات إلى نشر الذكاء الاصطناعي ذو الأغراض العامة على نطاق متزايد في سياقات حساسة مثل الرعاية الصحية أو مراقبة مكان العمل، ما يتسبب في ظهور مخاطر جديدة تتعلق بالخصوصية. وأخيرًا، ومع تكثيف النزاعات المتعلقة بحقوق الطبع والنشر واستمرار عدم موثوقية التدابير الفنية للتخفيف من انتهاكات حقوق الطبع والنشر، بدأ أصحاب حقوق البيانات في تقييد الوصول إلى بياناتهم بسرعة.

نماذج الوزن العام: إن العامل المهم في تقييم العديد من المخاطر التي قد يشكلها نموذج الذكاء الاصطناعي ذو الأغراض العامة هو كيفية إصداره للجمهور. تُعد ما يسمى بـ "نماذج الوزن العام" نماذج ذكاء اصطناعي تكون مكوناتها المركزية، والتي تسمى "الأوزان"، مشتركة بشكل عام للتحميل. يسهل الوصول المفتوح للوزن البحث والابتكار، بما في ذلك في مجال سلامة الذكاء الاصطناعي، فضلًا عن زيادة الشفافية وتسهيل على مجتمع البحث اكتشاف العيوب في النماذج. ومع ذلك، فإن نماذج الوزن العام يمكن أن تُشكّل أيضًا مخاطر، على سبيل المثال من خلال تسهيل الاستخدام الخبيث أو المضلّل الذي يصعب أو يستحيل على مطور النموذج مراقبته أو التخفيف منه. بمجرد توفر أوزان النموذج للتحميل العام، لا توجد طريقة لتنفيذ التراجع الشامل لجميع النسخ الموجودة أو ضمان حصولهاعلى تحديثات الأمان. منذ التقرير المؤقت (مايو/أيار 2024)، ظهر إجماع رفيع المستوى على أن المخاطر التي يفرضها انفتاح الذكاء الاصطناعي الأكبر يجب تقييمها من حيث المخاطر "الهامشية": وهي المدى الذي قد يؤدي فيه إصدار نموذج الوزن العام إلى زيادة مخاطر معينة أو تقليلها، نسبةً إلى المخاطر التي تفرضها البدائل الحالية مثل النماذج المغلقة أو التقنيات الأخرى.

القسم الثالث - إدارة المخاطر: ما هي التقنيات المُستخدمة لإدارة المخاطر الناجمة عن الذكاء الاصطناعي ذو الأغراض العامة؟

هناك العديد من الأساليب الفنية التي يمكن أن تساعد في إدارة المخاطر، ولكن في كثير من الحالات لا تزال أفضل الأساليب المتاحة تعاني قيودًا كبيرة للغاية ولا تتضمن تقديرات كمية للمخاطر أو ضمانات متاحة في مجالات أخرى بالغة الأهمية للسلامة.

إن إدارة المخاطر تحديد المخاطر وتقييمها، ثم التخفيف من حدتها ومراقبتها أمر صعب في سياق الذكاء الاصطناعي ذو الأغراض العامة. وعلى الرغم من أن إدارة المخاطر كانت أيضًا صعبة للغاية في العديد من المجالات الأخرى، إلا أن هناك بعض ميزات الذكاء الاصطناعي ذو الأغراض العامة التي يبدو أنها تتسبب في ظهور صعوبات فريدة.

إن العديد من الميزات التقنية للذكاء الاصطناعي ذو الأغراض العامة تجعل إدارة المخاطر في هذا المجال صعبة بصفة خاصة. وتشمل هذه الأمور، من بين أمور أخرى، ما يلي:

- **إن نطاق الاستخدامات الممكنة وسياقات الاستخدام لأنظمة الذكاء الاصطناعي ذو الأغراض العامة واسع بشكل غير عادي.** على سبيل المثال، يمكن استخدام نفس النظام لتقديم المشورة الطبية، وتحليل أكواد الحاسوب بحثًا عن الثغرات الأمنية، وإنشاء الصور. ويؤدي هذا إلى زيادة صعوبة توقع حالات الاستخدام بشكل شامل، أو تحديد المخاطر، أو اختبار كيفية تصرف الأنظمة في ظل الظروف الواقعية المختلفة.
- **لا يزال المطورون لا يفهمون إلا القليل عن كيفية عمل نماذج الذكاء الاصطناعي ذو الأغراض العامة الخاصة بهم.** ويجعل هذا الافتقار إلى الفهم من الصعب التنبؤ بالقضايا السلوكية وتفسير القضايا المعروفة وحلّها بمجرد ملاحظتها. ويظل الفهم بعيد المنال في المقام الأول لأن نماذج الذكاء الاصطناعي ذو الأغراض العامة ليست مبرمجة بالمعنى التقليدي. وبدلًا من ذلك، تُدربُ على التالي: يجهز مطورو الذكاء الاصطناعي عملية تدريب تتضمن كمية كبيرة من البيانات، ونتيجة هذه العملية التدريبية هو نموذج الذكاء الاصطناعي ذو الأغراض العامة. إن التفاصيل الداخلية لهذه النماذج غامضة إلى حد كبير، بما في ذلك بالنسبة لمطوري النماذج. إن تفسير النماذج وتقنيات "التفسير" يمكن أن تعمل على تحسين فهم الباحثين والمطورين لكيفية عمل نماذج الذكاء الاصطناعي ذو الأغراض العامة، ولكن على الرغم من التقدم المُحرز مؤخرًا، لا يزال هذا البحث في مراحله الأولى.
- **من المرجح أن يُشكّل وكلاء الذكاء الاصطناعي ذوي القدرات المتزايدة - أنظمة الذكاء الاصطناعي ذو الأغراض العامة التي يمكنها التصرف والتخطيط والتفويض باستقلال لتحقيق الأهداف - تحديات جديدة ومهمة لإدارة المخاطر.** عادةً ما يعمل وكلاء الذكاء الاصطناعي نحو تحقيق الأهداف باستقلال باستخدام برامج عامة مثل متصفحات الويب وأدوات البرمجة. في الوقت الحالي، لا تزال معظم هذه الأنظمة غير موثوقة بما يكفي للاستخدام على نطاق واسع، ولكن الشركات تبذل جهودًا كبيرة لصنع وكلاء ذكاء اصطناعي أكثر كفاءةً وموثوقيةً، وقد حققت تقدمًا في الأشهر الأخيرة. ومن المرجح أن يصبح وكلاء الذكاء الاصطناعي مفيدين بصورة متزايدة، ولكنهم قد يتسببون أيضًا في تفاقم عدد من المخاطر التي جرت مناقشتها في هذا التقرير وتقديم صعوبات إضافية لإدارة المخاطر. وتشمل أمثلة هذه التحديات الجديدة المحتملة إمكانية عدم معرفة المستخدمين دائمًا بما يفعله عملاء الذكاء الاصطناعي الخاص بهم، وإمكانية عمل عملاء الذكاء الاصطناعي خارج سيطرة أي شخص، وإمكانية قيام المهاجمين بـ "اختطاف" العملاء، وإمكانية تفاعلات الذكاء الاصطناعي مع الذكاء الاصطناعي لخلق مخاطر جديدة معقدة. لقد بدأت حديثًا عملية تطوير الأساليب الخاصة بإدارة المخاطر المرتبطة بالوكلاء.

بالإضافة إلى العوامل التقنية، هناك العديد من العوامل الاقتصادية والسياسية وغيرها من العوامل المجتمعية التي تجعل إدارة المخاطر في مجال الذكاء الاصطناعي ذو الأغراض العامة صعبة بصفة خاصة.

- **إن وتيرة التقدم في مجال الذكاء الاصطناعي ذو الأغراض العامة تخلق "معضلة أدلة" لصناع القرار.** † إن التقدم السريع في القدرات يجعل من الممكن ظهور بعض المخاطر على نحوٍ مفاجئ؛ على سبيل المثال، تحول خطر الغش الأكاديمي باستخدام الذكاء الاصطناعي ذو الأغراض العامة من خطر "ضئيل" إلى خطر "واسع الانتشار" في غضون عام واحد. كلما ظهر الخطر بسرعة أكبر، أصبح من الصعب إدارة الخطر على نحوٍ تفاعلي وأصبحت الاستعدادات أكثر قيمةً. ومع ذلك، طالما ظلت الأدلة على المخاطر غير كاملة، فإن صناع القرار لا يستطيعون أيضًا أن يعرفوا على وجه اليقين ما إذا كان الخطر سوف يظهر أو ربما ظهر بالفعل. وهذا الخطر يخلق مقايضة: قد يكون تنفيذ تدابير وقائية أو تخفيفية مبكرة أمرًا غير ضروري، ولكن انتظار الأدلة القاطعة قد يجعل المجتمع عرضة للمخاطر التي تظهر بسرعة. وتعمل الشركات والحكومات على تطوير أنظمة الإنذار المبكر وأطر إدارة المخاطر التي قد تساعد في الحدّ من هذه المعضلة. وتؤدي بعض هذه التشريعات إلى إطلاق تدابير تخفيف محددة عندما تظهر أدلة جديدة على وجود مخاطر، في حين تتطلب تشريعات أخرى من المطورين تقديم أدلة على السلامة قبل إصدار نموذج جديد.
- **هناك فجوة معلوماتية بين ما تعرفه شركات الذكاء الاصطناعي عن أنظمة الذكاء الاصطناعي الخاصة بها وما تعرفه الحكومات والباحثون غير العاملين في الصناعة.** غالبًا ما تشارك الشركات معلومات محدودة فقط حول أنظمة الذكاء الاصطناعي ذو الأغراض العامة الخاصة بها، وخاصة في الفترة التي تسبق إطلاقها على نطاق واسع. وتشير الشركات إلى مزيج من المخاوف التجارية والمخاوف المتعلقة بالسلامة كأسباب للحد من تبادل المعلومات. ومع ذلك، فإن هذه الفجوة في المعلومات تجعل من الصعب أيضًا على الجهات الفاعلة الأخرى المشاركة بفعالية في إدارة المخاطر، وخاصة فيما يتعلق بالمخاطر الناشئة.

† يرجى الرجوع إلى تحديث الرئيس بشأن أحدث التطورات في مجال الذكاء الاصطناعي بعد كتابة هذا التقرير.

- **غالبًا ما تواجه شركات الذكاء الاصطناعي والحكومات ضغوطًا تنافسية قوية، ما قد يدفعها إلى خفض أولوية إدارة المخاطر.** في بعض الظروف، قد تعمل الضغوط التنافسية على تحفيز الشركات على استثمار وقت أو موارد أخرى أقل في إدارة المخاطر مقارنةً بما كانت لتفعله في الظروف العادية. وعلى نحو مماثل، قد تستثمر الحكومات جهودًا أقل في السياسات الرامية إلى دعم إدارة المخاطر في الحالات التي ترى فيها وجود مقايضات بين المنافسة الدولية والحدّ من المخاطر.

ومع ذلك، هناك تقنيات وأطر عمل مختلفة لإدارة المخاطر الناجمة عن الذكاء الاصطناعي ذو الأغراض العامة والتي يمكن للشركات استخدامها ويمكن للهيئات التنظيمية أن تطلبها. وتشمل هذه الأساليب أساليب تحديد المخاطر وتقييمها، فضلًا عن أساليب التخفيف من حدتها ومراقبتها.

- **إن تقييم أنظمة الذكاء الاصطناعي ذو الأغراض العامة من حيث المخاطر يشكل جزءًا لا يتجزأ من إدارة المخاطر، ولكن تقييمات المخاطر الحالية محدودة للغاية.** إذ تعتمد التقييمات الحالية لمخاطر الذكاء الاصطناعي ذو الأغراض العامة بشكل أساسي على "الفحوص العشوائية"، أي اختبار سلوك الذكاء الاصطناعي ذو الأغراض العامة في مجموعة من المواقف المحددة. يمكن أن يساعد هذا في تسليط الضوء على المخاطر المحتملة قبل نشر النموذج. ومع ذلك، فإن الاختبارات الحالية غالبًا ما تفشل في اكتشاف المخاطر وتبالغ في تقدير قدرات الذكاء الاصطناعي ومخاطره ذو الأغراض العامة أو تقلل من شأنها، لأن ظروف الاختبار تختلف عن العالم الحقيقي.
- **ولكي تكون عملية تحديد المخاطر وتقييمها فعالة، يحتاج المقيمون إلى خبرة كبيرة وموارد وإمكانية الوصول الكافية إلى المعلومات ذات الصلة.** يتطلب تقييم المخاطر الدقيق في سياق الذكاء الاصطناعي ذو الأغراض العامة الجمع بين مناهج تقييم متعددة. وتتراوح هذه المهام من التحليلات الفنية للنماذج والأنظمة نفسها إلى تقييم المخاطر المحتملة الناجمة عن أنماط استخدام معينة. ويحتاج المقيمون إلى خبرة كبيرة لإجراء مثل هذه التقييمات بشكل صحيح. بالنسبة لتقييمات المخاطر الشاملة، فإنهم غالبًا ما يحتاجون أيضًا إلى مزيد من الوقت، ومزيد من الوصول المباشر إلى النماذج وبيانات التدريب الخاصة بها، والمزيد من المعلومات حول المنهجيات الفنية المستخدمة، مقارنةً بما توفره الشركات التي تطور الذكاء الاصطناعي ذو الأغراض العامة عادةً.
- **لقد حُقِّقَ تقدمٌ في تدريب نماذج الذكاء الاصطناعي ذو الأغراض العامة لتعمل بصورة أكثر أمانًا، ولكن لا توجد طريقة حالية يمكنها منع حتى المخرجات غير الآمنة بشكل واضح. على سبيل المثال، تتضمن تقنية تسمى "التدريب العدائي" تعريض نماذج الذكاء الاصطناعي عمدًا لأمثلة مُصمّمة لجعلها تفشل أو تتصرف بشكل سيء في أثناء التدريب، بهدف بناء المقاومة لمثل هذه الحالات.** ومع ذلك، لا يزال بإمكان الخصوم العثور على طرق جديدة ("هجمات") للالتفاف على هذه الضمانات بجهد "منخفض" إلى "متوسط". وعلاوة على ذلك، تشير الأدلة الأخيرة إلى أن أساليب التدريب الحالية – التي تعتمد إلى حدٍ كبير على الترشيحات البشرية للتعديل المتصفة بغير الكمال – قد تشجع النماذج عن غير قصد على تضليل البشر في الأسئلة الصعبة من خلال جعل الأخطاء أكثر صعوبةً في اكتشافها. إن تحسين كمية هذه الملاحظات ونوعيتها يُعد طريقًا للتقدم، على الرغم من أن تقنيات التدريب الناشئة باستخدام الذكاء الاصطناعي للكشف عن السلوك المُضلِّل تبدو واعدة أيضًا.
- **إن المراقبة، أي تحديد المخاطر وتقييم الأداء بمجرد أن يصبح النموذج قيد الاستخدام بالفعل، والتدخلات المختلفة لمنع الإجراءات الضارة يمكن أن تعمل على تحسين سلامة الذكاء الاصطناعي ذو الأغراض العامة بعد تشغيله ليكون متاحًا للمستخدمين.** يمكن للأدوات الحالية اكتشاف المحتوى الذي أنشأه الذكاء الاصطناعي، وتتبع أداء النظام، وتحديد المدخلات/المخرجات الضارة المحتملة، على الرغم من أن المستخدمين ذوي المهارات المتوسطة يمكنهم في كثير من الأحيان التحايل على هذه الضمانات. إن وجود عدة طبقات من الدفاع تجمع بين قدرات المراقبة الفنية والتدخل والإشراف البشري يعمل على تحسين السلامة، إلا أنها قد تؤدي إلى تكاليف وتأخيرات. في المستقبل، قد تساعد الآليات التي تدعمها الأجهزة العملاء والهيئات التنظيمية على مراقبة أنظمة الذكاء الاصطناعي ذو الأغراض العامة بشكل أكثر فعالية أثناء النشر، وقد تساعد في التحقق من الاتفاقيات عبر الحدود، ولكن آليات موثوقة من هذا النوع لا وجود لها حتى الآن.
- **توجد طرق متعددة عبر دورة حياة الذكاء الاصطناعي لحماية الخصوصية. وتشمل هذه الحلول إزالة المعلومات الحساسة من بيانات التدريب، وأساليب التدريب النموذجية التي تتحكم في مقدار المعلومات التي يتعلمها من البيانات (مثل أساليب "الخصوصية التفاضلية")، وتقنيات استخدام الذكاء الاصطناعي مع البيانات الحساسة التي تجعل من الصعب استرداد البيانات (مثل "الحوسبة السرية" وغيرها من التقنيات التي تعمل على تعزيز الخصوصية).** العديد من طرق تعزيز الخصوصية من مجالات بحثية أخرى لا تزال غير قابلة للتطبيق على أنظمة الذكاء الاصطناعي ذو الأغراض العامة بسبب المتطلبات الحسابية لأنظمة الذكاء الاصطناعي. وفي الأشهر الأخيرة، توسعت أساليب حماية الخصوصية لمعالجة الاستخدام المتزايد للذكاء الاصطناعي في المجالات الحساسة بما في ذلك مساعدي الهواتف الذكية، ووكلاء الذكاء الاصطناعي، والمساعدين الصوتيين الذين يستمعون دائمًا، والاستخدام في الرعاية الصحية أو الممارسة القانونية.

منذ نشر التقرير المؤقت (مايو/أيار 2024)، أحرز الباحثون بعض التقدم الإضافي نحو القدرة على تفسير سبب إنتاج نموذج الذكاء الاصطناعي ذو الأغراض العامة لتحقيق نتيجة معينة. إن القدرة على شرح قرارات الذكاء الاصطناعي يمكن أن تساعد في إدارة المخاطر الناجمة عن الأعطال التي تتراوح بين التحيز وعدم دقة الحقائق و فقدان السيطرة. وعلاوة على ذلك، كانت هناك جهود متزايدة لتوحيد أساليب التقييم والتخفيف في جميع أنحاء العالم.

الاستنتاج: هناك مجموعة واسعة من المسارات المحتملة لمستقبل الذكاء الاصطناعي ذو الأغراض العامة، وسيعتمد الكثير على كيفية تصرف المجتمعات والحكومات.

إن مستقبل الذكاء الاصطناعي ذو الأغراض العامة غير مؤكّد، مع وجود مجموعة واسعة من المسارات التي تبدو ممكنة حتى في المستقبل القريب، بما في ذلك النتائج الإيجابية للغاية والسلبية للغاية. ولكن لا يوجد شيء حتمي فيما يتعلق بمستقبل الذكاء الاصطناعي ذو الأغراض العامة. كيف يجري تطوير الذكاء الاصطناعي ذو الأغراض العامة ومن يقوم بذلك، وما هي المشاكل التي صُمِمَ لحلها، وما إذا كانت المجتمعات ستكون قادرة على جني الإمكانات الاقتصادية الكاملة للذكاء الاصطناعي ذو الأغراض العامة، ومن يستفيد منه، وأنواع المخاطر التي نعرّض أنفسنا إليها، وكم نستثمر في البحث لإدارة المخاطر هذه والعديد من الأسئلة الأخرى تعتمد على الخيارات التي تتخذها المجتمعات والحكومات اليوم وفي المستقبل لتشكيل تطوير الذكاء الاصطناعي ذو الأغراض العامة.

ولتسهيل المناقشة البنّاءة حول هذه القرارات، يقدم هذا التقرير نظرة عامة على الحالة الحالية للبحث العلمي والمناقشة حول إدارة مخاطر الذكاء الاصطناعي ذو الأغراض العامة. المخاطر كبيرة. ونحن نتطلع إلى مواصلة هذا الجهد.