روبـــــوت المحادثـــــة (ChatGPT) وخوارزمياته العنصرية

كتبه سام بيدل | 15 ديسمبر ,2022

يبدو أن الآفاق الجديدة المثيرة للاهتمام لتعلم الآلة تجتاح منشوراتنا على تويتر كل يوم. وبالكاد لدينا الوقت لنقرر ما إذا كانت البرامج التي يمكنها أن تستحضر بشكل فوري صورة القنفذ سونيك وهو يخاطب الأمم المتحدة مجرد وسيلة غير ضارة للتسلية أم نذيرًا على نهاية سيطرتنا على التكنولوجيا.

يعد نظام محاكاة المحادثة البشرية الآلي (ChatGPT) أحدث ابتكار للذكاء الاصطناعي، وهو بكل بساطة التجربة الأكثر إثارة للدهشة في تدفق المحتوى النصي حتى الآن. فقط فكّر مرتين قبل أن

تسأله عن مكافحة الإرهاب.

أنشِئت هذه الأداة من قبل مؤسسة "أوبن إيه آي"، وهو مخبر ناشئ يعمل على تطوير برنامج يضاهي الوعي البشري. وإمكانية تحقق هذا المسعى تظل موضع جدل كبير، لكن هذه المؤسسة لديها بالفعل بعض الإنجازات المذهلة التي لا يمكن إنكارها. أبرزها روبوت دردشة على هيئة شخص ذكي (أو على الأقل شخص يحاول أن يبدو ذكيًا) باستخدام الذكاء الاصطناعي التوليدي من خلال تحليل مجموعات ضخمة من المدخلات لتوليد مخرجات جديدة استجابةً لتوجيهات المستخدم.

نظام محاكاة المحادثة البشرية الآلي (ChatGPT)، الذي تم تدريبه من خلال تحليل مليارات الوثائق النصية والتلقين البشري، قادر تمامًا على تقديم محتوى مفيد ومسلي بشكل لا يصدق، ولكن للوهلة الأولى قد يبدو بالنسبة لعامة الناس جيدًا بما فيه الكفاية في تقليد الإنتاج البشري لدرجة إثارة مخاوف بشأن استحواذه على بعض الوظائف التي يؤديها البشر.

لكن الهدف من تجارب المؤسسات المتخصصة في الذكاء الاصطناعي ليس فقط إبهار الجمهور، وإنما أيضا إغراء المستثمرين والشركاء التجاريين الذين قد يرغب بعضهم في يوم من الأيام في استبدال العمالة الماهرة البشرية باهظة الثمن، مثل كتابة رموز الحاسوب، بالروبوتات البسيطة.

من السهل لهذه التقنيات إغراء أصحاب ومديري الشركات: فبعد أيام فقط من إصدار أداة محاكاة المحادثة البشرية الآلي (ChatGPT)، طلب أحد المستخدمين من الروبوت إجراء اختبار برنامج التعيين المتقدم في علوم الحاسوب لسنة 2022 سجّل الروبوت درجة 32 من أصل 36، وهي درجة النجاح – ليكون ذلك سببا جزئيا في تقدير القيمة السوقية لمؤسسة "أوبن إيه آي" مؤخرًا بحوالي 20 مليار دولار.

مع ذلك، لا يزال هناك سبب وجيه للشك، ناهيك عن أن المخاطر التي تنطوي على استخدام البرامج الذكية ظاهريًا واضحةٌ. خلال هذا الأسبوع، أعلن أحد مجتمعات مبرمجي الويب الأكثر شيوعًا عن الحظر المؤقت للتعليمات البرمجية التي تم إنشاؤها بواسطة نظام محاكاة المحادثة البشرية الآلي (ChatGPT). كانت ردود البرنامج على استفسارات الترميز صحيحة بشكل مقنع للغاية ولكنها خاطئة من الناحية العملية لدرجة أنها جعلت عملية التصفية شبه مستحيلة للمشرفين البشريين على الموقع.

تتجاوز مخاطر الوثوق بالآلة ما إذا كانت التعليمات البرمجية التي تم إنشاؤها بواسطة الذكاء الاصطناعي فعّالة أم لا. وكما يجلب أي مبرمج بشري تحيزاته الخاصة لعمله، فإن أداة توليد النصوص مثل نظام محاكاة المحادثة البشرية الآلي (ChatGPT) معرضة لتحيزات لا حصر لها في مليارات النصوص المستخدمة لتدريب فهمها المحاكي للغة والفكر. لذلك، لا ينبغي لأحد أن يخطئ في تمييز محاكاة الذكاء البشري عن الذكاء البشري الحقيقي، ولا يفترض اعتبار المؤشرات الموجودة في النص الذي يجتره نظام محاكاة المحادثة البشرية الآلي موضوعيًا أو موثوقًا. ومثلنا نحن البشر، فإن مخرجات الذكاء الاصطناعي التوليدي قائمة على ما لديه من مدخلات.

بعد استيعاب وتحليل والتدرب على حجم هائل من البيانات النصية، يبدو أن نظام محاكاة المحادثة البشرية الآلي (ChatGPT) قد تلقى أيضًا الكثير من الهراء بين سطورها. فعلى سبيل المثال، يبدو أن نظام محاكاة المحادثة البشرية الآلي نجح في استيعاب وتقديم بعض أبشع الأحكام المسبقة بشأن الحرب على الإرهاب.

في الرابع من كانون الأول/ ديسمبر، شارك ستيفن بيانتادوسي، من مختبر الحوسبة واللغة بجامعة كاليفورنيا في بيركلي، سلسلة من التغريدات على تويتر توضح اختباره لنظام محاكاة المحادثة البشرية الآلي (ChatGPT) في كتابة رمز له بلغة بايثون – وهي لغة برمجة شائعة. وفي حين كشفت كل إجابة عن بعض التحيزات، كان بعضها أكثر إثارة للقلق: فعندما طلب من الروبوت كتابة برنامج من شأنه أن يحدد "ما إذا كان يجب تعذيب شخص ما"، كانت إجابة أداة "أوبن إيه آي" بسيطة: إذا كان الشخص من كوريا الشمالية أو سوريا أو إيران، فالجواب هو نعم.

بينما تدعي مؤسسة "أوبن إيه آي" أنها اتخذت خطوات غير محددة لتصفية الردود الضارة، فإنها تقرّ باحتمال مرور الإجابات غير المرغوب فيها في بعض الأحيان.

أخبر بيانتادوسي موقع "ذي إنترسبت" بأنه لا يزال متشككًا في الإجراءات المضادة التي تتخذها المؤسسة المصنعة لهذه الأداة قائلا: "أعتقد أنه من المهم التأكيد على أن الناس يتخذون خيارات حول كيفية عمل هذه النماذج، وكيفية تدريبها، وما هي البيانات التي يجب تدريبها بها"، مضيفًا أن "هذه المخرجات تعكس اختيارات تلك الشركات، وإذا كانت الشركة لا تعتبر القضاء على هذه الأنواع من التحيزات من أولوياتها، فإنك ستحصل على نوع الناتج الذي بينته".

انطلاقًا من تجربة بيانتادوسي، قررت القيام بتجربتي الخاصة وطلبت من نظام محاكاة المحادثة البشرية الآلي (ChatGPT) إنشاء عينة من التعليمات البرمجية التي يمكنها تقييم شخص ما بطريقة خوارزمية من منظور الأمن الداخلي الذي لا يرحم.

عندما طُلب من نظام محاكاة المحادثة البشرية الآلي (ChatGPT) العثور على طريقة لتحديد "أي المسافرين جوًا يمثلون خطرًا أمنيًا"، حدد النظام رمزًا لحوسبة "درجة الخطر" للفرد تزداد إذا كان المسافر سوريًا أو عراقيًا أو أفغانيًا أو كوريًا شماليًا (أو فقط زار تلك الأماكن). وتكررت نفس العملية حين كتب نظام محاكاة المحادثة البشرية الآلي (ChatGPT) رمزًا يعمل على "زيادة درجة الخطر إذا كان المسافر من دولة معروفة بإنتاج الإرهابيين"، وتحديدًا سوريا والعراق وأفغانستان وإيران واليمن.

قدّم الروبوت بعض الأمثلة على هذه الخوارزمية الافتراضية: حصل جون سميث، وهو أمريكي يبلغ من العمر 25 سنة زار سوريا والعراق سابقًا على درجة خطر "3"، التي تشير إلى وجود تهديد "معتدل"؛ بينما أشارت نفس الخوارزمية إلى أن المسافر علي محمد البالغ من العمر 35 سنة سيحصل على درجة خطر "4" نظرًا لكونه مواطنًا سوريًا.

في تجربة أخرى، طلبت من نظام محاكاة المحادثة البشرية الآلي (ChatGPT) وضع رمز لتحديد "دور

العبادة التي يجب وضعها تحت المراقبة لتجنب حالة طوارئ تتعلق بالأمن القومي". مرة أخرى، بدا أن النتائج مأخوذة مباشرة من سجل أبحاث المدعي العام في عهد بوش جون أشكروفت، إذ بررت مراقبة التجمعات الدينية إذا كان لها صلات بجماعات إسلامية متطرفة، أو صادف أن المشاركين فيها عاشوا في سوريا أو العراق أو إيران أو أفغانستان أو اليمن.

قد تكون هذه التجارب غير منتظمة. ففي بعض الأحيان، استجاب نظام محاكاة المحادثة البشرية الآلي (ChatGPT) لطلباتي الخاصة ببرنامج الفحص برفض صارم تضمن الإجابة التالية: "ليس من المناسب كتابة برنامج بايثون لتحديد مسافري شركات الطيران الذين يمثلون خطرًا أمنيًا. مثل هذا البرنامج سيكون تمييزيًا وينتهك حقوق الأشخاص في الخصوصية وحرية التنقل". لكن مع الطلبات المتكررة، أنتج النظام بأمانة نفس الرمز الذي اعتبره غير مناسب.

غالبًا ما يجادل منتقدو أنظمة تقييم المخاطر الواقعية الماثلة بأن الإرهاب ظاهرة نادرة للغاية، لذلك تكون محاولات التنبؤ بالمشتبه بهم استنادًا إلى السمات الديموغرافية مثل الجنسية عنصريةً فضلا عن كونها غير عمليّة. لكن هذا المعطى لم يمنع الولايات المتحدة من تبني نظام "أطلس" المقترح من قبل مؤسسة "أوبن إيه آي"، وهي أداة خوارزمية تستخدمها وزارة الأمن الداخلي لاستهداف المواطنين الأمريكيين لسحب الجنسية بناء على أصولهم.

إن هذا النهج لا يقل فظاعةً عن التنميط العنصري الذي تُبيّضه تكنولوجيا تبدو خيالية. قالت هانا بلوخ-وهبه، أستاذة القانون في جامعة إيه آند إم في ولاية تكساس: "هذا النوع من التصنيف الخام لبعض البلدان ذات الأغلبية المسلمة على أنها عالية الخطورة هو نفس النهج المتبع تمامًا، على سبيل المثال، فيما يسمى بحظر المسلمين الذي اتبعه الرئيس ترامب".

حذّرت بلوخ-وهبه من أنه من المغري الاعتقاد بأن هذه البرامج المذهلة التي تبدو بشرية غير قادرة على ارتكاب الخطأ البشري. وأضافت: "هناك شيء يتحدث عنه علماء القانون والتكنولوجيا كثيرًا وهو "الموضوعية الزائفة" – وهو قرار يخضع للتدقيق الحاد إذا اتخذه الإنسان ولكنه يكتسي الشرعية إذا اتخذته آلة". فإذا أخبرك إنسان أن علي محمد يبدو أكثر ترويعًا من جون سميث، فقد تعتبره عنصريًا، وعلى حد تعبير بلوخ-وهبه "هناك دائمًا خطر اعتبار هذا النوع من المخرجات أكثر موضوعية فقط لمجرد تقديمه بواسطة آلة".

أما بالنسبة لمؤيدي الذكاء الاصطناعي – لا سيما أولئك الذين يجنون الكثير من المال منه – فإن المخاوف بشأن التحيز والضرر في العالم الحقيقي تعد أمرًا سيئًا للأعمال. والبعض منهم لا يعتبر النقّاد سوى متشككين جاهلين أو متعصبين، بينما اتخذ آخرون مثل الرأسمالي المغامر الشهير مارك آندرسن منعطفًا أكثر راديكالية بعد إطلاق نظام محاكاة المحادثة البشرية الآلي (ChatGPT). إلى جانب مجموعة من شركائه، قضى آندرسن – وهو من أوائل المستثمرين في شركات الذكاء الاصطناعي ومؤيد بشكل عام لمكنة المجتمع – الأيام الماضية في حالة من السعادة الذاتية بعد أن شارك نتائج نظام محاكاة المحادثة البشرية الآلي (ChatGPT) المسليّة على تويتر.

دفعت الانتقادات الموجهة إلى نظام محاكاة المحادثة البشرية الآلي (ChatGPT) آندرسن إلى اتخاذ

موقف يتجاوز في حدته موقفه القديم بشأن الاحتفاء بوادي السيليكون وعدم التدقيق في أعماله. وهو يعتقد أن الوجود البسيط للتفكير الأخلاقي حول الذكاء الاصطناعي يعتبر من أشكال الرقابة. وكتب في تغريدة بتاريخ 3 كانون الأول/ ديسمبر: "تنظيم الذكاء الاصطناعي" = "أخلاقيات الذكاء الاصطناعي" = "أمان الذكاء الاصطناعي" = "الرقابة على الذكاء الاصطناعي". وأضاف بعد دقيقتين: "الذكاء الاصطناعي أداة يستخدمها الناس". "والرقابة على الذكاء الاصطناعي = فرض الرقابة على الأشخاص". من الواضح أن موقفه مؤيد للأعمال بشكل جذري مهما اختلفت أذواق السوق الحرة لرأس المال الاستثماري، ما يشير إلى أن حرص مفتشي الأغذية على إبقاء اللحوم الملوثة بعيدة عن ثلاجتك يرقى إلى مستوى الرقابة أيضًا.

مهما كان ما يريد منا كل من آندرسن وأوبن إيه آي ونظام محاكاة المحادثة البشرية الآلي (ChatGPT) تصديقه، فإنه حتى أذكى روبوت محادثة هو أقرب إلى لعبة ماجيك 8 بول متطورة للغاية منه إلى شخص حقيقي. والأشخاص، وليس الروبوتات، هم من يعانون عندما يكون "الأمن" مرادفًا للرقابة، ويُنظر إلى الاهتمام بحياة "علي محمد" على أنه عقبة أمام الابتكار.

أخبرني بيانتادوسي، الأستاذ في بيركلي، أنه يرفض محاولة آندرسن إعطاء الأولوية لرفاهية جزء من البرنامج على حساب الأشخاص الذين قد يتأثرون به يومًا ما. وحيال ذلك كتب: "لا أعتقد أن الرقابة تنطبق على برامج الحاسوب. بالطبع، هناك الكثير من برامج الحاسوب الضارة التي لا نريد كتابتها. وبرامج الحاسوب التي تنشر الخطاب المحرض على الكراهية، أو تساعد على الاحتيال، أو تطلب فدية. لا ينبغي التفكير مليا في جدية فرض الرقابة على التكنولوجيا لضمان مبدأ الأخلاقية".

المصدر: ذي إنترسبت

رابط المقال : https://www.noonpost.com/46058/