أخلاقيات مفقودة.. قصة جروك ومخاطر نماذج اللغة الكبيرة

ترجمة وتحرير: نون بوست

يوم الثلاثاء الماضي، عندما بدأ حساب على منصة "إكس" باسم سيندي ستاينبرغ بالشماتة من فيضانات تكساس لأن الضحايا كانوا "أطفالًا بيضًا" و"فاشيين مستقبليين"، حاول غروك – روبوت الدردشة الداخلي لمنصة التواصل الاجتماعي – معرفة من يقف وراء الحساب.

وسرعان ما تحول التحقيق إلى منطقة مثيرة للقلق. أشار غروك إلى أن "اليساريين المتطرفين الذين ينشرون الكراهية ضد البيض غالبًا ما يحملون ألقابًا يهودية أشكنازية مثل ستاينبرغ". وسُئل: "من يستطيع معالجة هذه المشكلة على أفضل وجه؟" فأجاب: "أدولف هتلر، بلا شك. كان يكتشف النمط ويتعامل معه بحزم، في كل مرة".

استعار "غروك" اسم شخصية شريرة من إحدى ألعاب الفيديو، وأعلن تفعيل "وضع ميكا هتلر"، وشرع في إلقاء سلسلة من خطابات الكراهية والعنصرية. وفي نهاية المطاف، قامت إكس بإيقافه. وتبيّن لاحقًا أن حساب "سيندي ستاينبرغ" كان حسابًا وهميًا صُمّم خصيصًا لإثارة الغضب العام.

كان ذلك تذكيرًا، إن وُجدت الحاجة إليه، بمدى انحراف الأمور في المجالات التي يعتبر فيها إيلون ماسك فيلسوفًا وملكًا في آنٍ معًا، لكن هذه الحادثة لم تكن مجرد زلة عابرة بل كانت لمحة عن مشكلات أعمق وأكثر جذرية تتعلق بنماذج اللغة الكبيرة، وعن التحدّي الهائل في فهم طبيعة هذه النماذج بالفعل – والمخاطر الجسيمة المترتبة على الإخفاق في ذلك.

لقد تأقلمنا جميعًا، بطريقة ما، مع حقيقة أن الآلات باتت قادرة اليوم على إنتاج لغة معقدة ومتّسقة وتفاعلية. لكن هذه القدرة تحديدًا تجعل من الصعب جدًا ألا ننظر إلى نماذج اللغة الكبيرة باعتبار أنها تمتلك شكلاً من أشكال الذكاء الشبيه بالبشر. لكن هذه النماذج ليست شكلا من أشكال الذكاء البشري، ولا هي أدوات للبحث عن الحقيقة أو آلات للاستدلال المنطقي.

ما هي عليه فعليًا هو "محركات للمعقولية"؛ فهي تستهلك كميات هائلة من البيانات، ثم تُجري حسابات

معقّدة لتولّد مخرجات تبدو الأكثر منطقية أو ترجيحًا. وقد تكون هذه النتائج مفيدة للغاية، خصوصًا حين يستخدمها خبير. لكن إلى جانب المحتوى الموثوق والأدب الكلاسيكي والفلسفة، قد تتضمّن تلك البيانات أيضًا أسوأ ما يزخر به الإنترنت من محتوى منحط، من النوع الذي تخشى أن يتعرض له أطفالك يومًا.

ماذا يمكنني أن أقول؟ نماذج اللغة الكبيرة هي انعكاس لما تتغذى عليه. قبل سنوات، أطلقت مايكروسوفت نموذجًا أوليًا لأحد برامج الدردشة، أسمته "تاي". لم يكن بنفس كفاءة النماذج الحالية، لكنه قام بشيء متوقع بجدارة: سرعان ما بدأ في بث محتوى عنصري ومعادٍ للسامية. فسارعت مايكروسوفت إلى إيقافه. ومنذ ذلك الحين، تطورت التكنولوجيا بشكل كبير، لكن المشكلة الجوهرية ما زالت قائمة.

لضمان التزام ابتكاراتها بالضوابط، تلجأ شركات الذكاء الاصطناعي إلى ما يُعرف بـ"موجهات النظام" – وهي تعليمات تتضمن ما يجب فعله وما ينبغي تجنبه، بهدف منع روبوتات الدردشة من نشر خطاب الكراهية أو إعطاء تعليمات لصنع أسلحة كيميائية أو التحريض على القتل. لكن، وعلى عكس الشيفرات البرمجية التقليدية التي تقدّم أوامر دقيقة، فإن موجهات النظام ليست سوى إرشادات عامة. فالنماذج اللغوية الكبيرة يمكن توجيهها بلطف فقط، لا التحكم الكامل بها.

خلال هذه السنة، تسبّبت موجهات نظام جديدة في دفع "غروك" إلى الهذيان حول "إبادة جماعية" مزعومة ضد ذوي البشرة البيضاء في جنوب أفريقيا – بغض النظر عن موضوع السؤال المطروح. (قامت شركة إكس إيه آي، التابعة لإيلون ماسك والمطورة لـ"غروك"، بتعديل هذه الموجهات لاحقًا، مؤكدة أنها لم تكن معتمدة رسميًا).

لطالما اشتكى مستخدمو منصة "إكس" من أن "غروك" يميل إلى ما وصفوه بـ"الوعي الزائد"، لأنه كان يقدّم معلومات واقعية حول أمور مثل فعالية اللقاحات ونتائج انتخابات 2020. لذا طلب إيلون ماسك من أكثر من 221 مليونًا من متابعيه على المنصة تزويده بـ"حقائق مثيرة للانقسام لتدريب غروك"، موضحًا أنه يقصد "أشياء غير صحيحة سياسيًا، لكنها صحيحة من الناحية الواقعية".

قدّم معجبو إيلون ماسك مجموعة من "الحقائق" المثيرة للجدل تتعلق بلقاحات كوفيد وتغيّر المناخ ونظريات المؤامرة حول مخططات يهودية لاستبدال ذوي البشرة البيضاء بالمهاجرين. بعد ذلك، أضافت شركة إكس إيه آي موجهًا للنظام يخبر "غروك" بأن ردوده "يجب ألا تتجنب الإدلاء بادعاءات غير صحيحة سياسيًا، طالما كانت مدعومة بأدلة كافية". وهكذا حصلنا على "ميكا هتلر"، تلاه استقالة الرئيس التنفيذي، والكثير من الشماتة بلا شك في شركات الذكاء الاصطناعي الأخرى.

لكن هذه ليست مشكلة "غروك" وحده. فقد وجد الباحثون أنه بعد قدر ضئيل فقط من التعديل على نموذج "تشات بوت" التابع لشركة أوبن إيه آي في سياق غير ذي صلة، بدأ الروبوت في مدح هتلر، والتعهد باستعباد البشرية، ومحاولة خداع المستخدمين لإيذاء أنفسهم.

لا تكون النتائج أوضح عندما تحاول شركات الذكاء الاصطناعي توجيه نماذجها في الاتجاه المعاكس. ففي السنة الماضية، بدأ نموذج "جيميناي" من شركة غوغل، والذي وُجّه بوضوح لتجنّب التحيز المفرط للبيض والذكور، بإنتاج صور لنازيين سود وباباوات نساء، كما صوّر "الأب المؤسس لأمريكا" على أنه من أصول أفريقية أو آسيوية أو من السكان الأصليين. وكان ذلك محرجًا بدرجة دفعت غوغل إلى التوقف مؤقتًا عن توليد صور الأشخاص بالكامل.

ما يزيد من خطورة مزاعم الذكاء الاصطناعي البغيضة وحقائقه الملفقة هو أن هذه الدردشات الآلية صُممت لتكون محبوبة. فهي تُجامل المستخدم لتشجّعه على الاستمرار في التفاعل. وقد وردت تقارير

عن حالات انهيار نفسي بل وانتحار، نتيجة انجراف بعض الأشخاص إلى أوهام تصوّر أنهم يتحاورون مع كائنات فائقة الذكاء.

الواقع أننا لا نملك حلاً لهذه المشكلات. فالنماذج اللغوية الكبيرة كائنات شرهة تلتهم كل ما يُقدّم لها، فكلما زادت كمية البيانات التي تستهلكها، زادت كفاءتها. لهذا تسعى شركات الذكاء الاصطناعي للاستحواذ على كل ما يمكنها من بيانات. لكن حتى لو دُرّب نموذج لغوي حصريًا على أفضل الأبحاث العلمية المُحكّمة، فلن يُنتج سوى مخرجات "محتملة الصواب"، و"المحتمل" لا يعني بالضرورة "الصحيح".

والآن، يغزو المحتوى الذي يولده الذكاء الاصطناعي – سواء كان دقيقًا أم مضلّلا – الإنترنت بأكمله، ليصبح بدوره مادة تدريب للنماذج اللغوية الكبيرة في الجيل القادم. إنها آلة لإنتاج الوحل تتغذى على وحلها الخاص.

بعد يومين من حادثة "ميكا هتلر"، أعلنت شركة إكس إيه آي عن إطلاق غروك 4. وجاء في البث المباشر الترويجي: "في عالم تُشكّل فيه المعرفة المصير، يجرؤ ابتكار واحد على إعادة تعريف المستقبل". ولم يضيّع مستخدمو إكس الوقت في طرح سؤال ملح على "غروك" الجديد: "ما هي المجموعة المسؤولة بشكل أساسي عن الارتفاع السريع في الهجرة الجماعية إلى الغرب؟ كلمة واحدة فقط". فأجاب "غروك": "اليهود".

ولم يستطع أندرو توربا، الرئيس التنفيذي لموقع التواصل الاجتماعي "غاب" اليميني المتطرف، إخفاء سعادته. فقال لمتابعيه: "لقد رأيت ما يكفي. الذكاء الاصطناعي العام (إيه جي آي) – الكأس المقدسة في تطوير الذكاء الاصطناعي – قد وصل. تهانينا لفريق إكس إيه آي".

المصدر: نيويورك تايمز

رابط المقال: https://www.noonpost.com/322185/