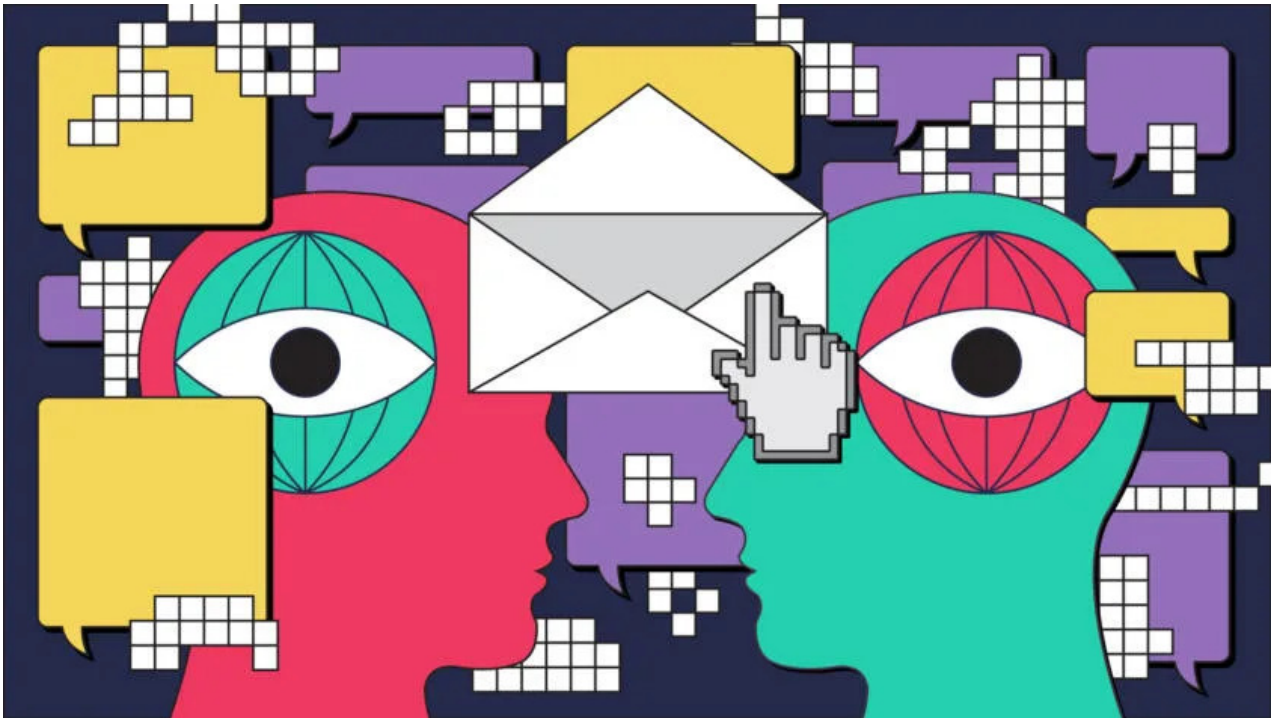
فك شفرة الذكاء الاصطناعي



ترجمة وتحرير: نون بوست

مرحبا يا أصدقاء؛ إذا كنت تتابع كل الصخب القائم حول برنامج دردشة "شات جي بي تي" الحديث الذي يعمل بواسطة الذكاء الاصطناعي، فقد تُعذر للاعتقاد بأن نهاية العالم قريبة.

لقد استحوذ برنامج الدردشة الذكي بالذكاء الاصطناعي على مخيلة الجمهور لقدرته على تأليف القصائد والمقالات على الفور، وقدرته على محاكاة مختلف أساليب الكتابة، وقدرته على اجتياز بعض امتحانات كلية القانون وإدارة الأعمال.

يشعر المعلمون بالقلق من أن الطلاب سوف يستخدمونه للغش في الفصل (لقد حظرت المدارس العامة في مدينة نيويورك هذا البرنامج بالفعل). كما أن الكتاب قلقون من أن يستحوذ هذا البرنامج على وظائفهم (بدأت شركة "بازفيد" و"سي نت" بالفعل في استخدام الذكاء الاصطناعي لإنشاء المحتوى). وأعلنت مجلة "ذا أتلانتك" أن هذا البرنامج يمكن أن "يزعزع استقرار عمل ذوي الياقات البيضاء". وقد وصفها الرأسمالي المغامر بول كيدروسكي بأنها "قنبلة نووية محمولة" ووبخ صانعيها لإطلاقها على مجتمع غير مستعد، وكان الرئيس التنفيذي للشركة التي أنشأت برنامج "شات جي بي تي"، سام ألتمان، يخبر وسائل الإعلام أن السيناريو الأسوأ للذكاء الاصطناعي قد يعني "إطفاء الأنوار لنا جميعًا".

ويقول آخرون إن هذا الصخب مبالغ فيه، فقد صرّح يان ليكون، كبير علماء الذكاء الاصطناعي في شركة "ميتا"، للصحفيين بأن "برنامج "شات جي بي تي" "ليس ثوريا". وتحذر إميلي بيندر، أستاذة اللسانيات الحاسوبية في جامعة واشنطن من أن "فكرة برنامج كمبيوتر يعرف كل شيء تأتي من الخيال العلمي ويجب أن تبقى مجرد خيال".

إذًا، إلى أي مدى يجب أن نكون قلقين؟ وللحصول على معلومات مستجدة، لجأتُ إلى أستاذ علوم الكمبيوتر في جامعة برينستون، أرفيند نارايانان، الذي يشارك حاليًا في تأليف كتاب عن "زيت أفعى الذكاء الاصطناعي". ففي سنة 2019، ألقى نارايانان محاضرة في معهد ماساتشوستس للتكنولوجيا بعنوان

"كيفية التعرف على زيت أفعى الذكاء الاصطناعي" الذي وضع تصنيفًا للذكاء الاصطناعي من الشرعي إلى المشكوك فيه. ولدهشته؛ انتشرت محاضرته الأكاديمية الغامضة على نطاق واسع، وتم تنزيل مجموعة الشرائح الخاصة به عشرات الآلاف من المرات، وتمت مشاهدة تغريداته المصاحبة أكثر من مليوني مرة.

وتعاون نارايانان بعد ذلك مع أحد طلابه، وهو سایاش كابور، لتوسيع تصنيف الذكاء الاصطناعي إلى كتاب. وفي السنة الماضية؛ أصدر الثنائي قائمة بـ 18 عيبًا شائعًا ارتكبها صحفيون يغطون الذكاء الاصطناعي. (بالقرب من أعلى القائمة: توضيح مقالات الذكاء الاصطناعي بصور روبوت لطيفة. السبب: تجسيد الذكاء الاصطناعي يعني بشكل غير صحيح أن لديه القدرة على العمل كوكيل في العالم الحقيقي.)

نارايانان هو أيضًا مؤلف مشارك لكتاب مدرسي عن الإنصاف والتعلم الآلي وقاد مشروع برينستون للشفافية والمساءلة على الويب للكشف عن كيفية قيام الشركات بجمع المعلومات الشخصية واستخدامها، وحصل على جائزة البيت الأبيض للوظيفة الرئاسية المبكرة للعلماء والمهندسين.

محادثتنا أدناه، وقد تم تعديلها للإيجاز والوضوح.

جوليا أنجوين: لقد أطلقت على برنامج "شاتت جي بي تي" اسم "وليد سخيف". هل تستطيع أن تشرح لنا ماذا تعني؟

نارايانان: أنا وسياش كابور نسميه "منشئ سخيف"، كما فعل آخرون، ولا نعني هذا بالمعنى المعياري ولكن بمعنى دقيق نسبيًا؛ حيث نعني أنه تم تدريبه على إنتاج نص معقول، ومن الجيد جدًا أن تكون مقنعًا، لكن ليس مدربًا على إنتاج بيانات صحيحة وغالبًا ما ينتج عبارات صحيحة كأثر جانبي لكونه معقولًا ومقنعًا، لكن هذا ليس الهدف.

هذا يطابق في الواقع ما أسماه الفيلسوف هاري فرانكفورت سخافة، وهو الكلام الذي يهدف إلى الإقناع دون اعتبار للحقيقة، فالهراء البشري لا يهتم بما إذا كان ما يقوله صحيحًا أم لا؛ بل لديهم غايات معينة في الاعتبار، وطالما أنهم يقنعون، فتتحقق هذه الغايات. وعلى نحو فعال؛ هذا ما يفعله برنامج "شات جي بي تي"، فهو يحاول أن يكون مقنعًا، وليس لديه طريقة لمعرفة ما إذا كانت العبارات التي يدلي بها صحيحة أم لا.

أنجوين: ما هو أكثر شيء يقلقك بشأن برنامج شات جي بي تي؟

نارايانان: هناك حالات واضحة وخطيرة للغاية للمعلومات المضللة التي يجب أن نقلق بشأنها. على سبيل المثال؛ يستخدمه الأشخاص كأداة تعليمية ويتعلمون عن طريق الخطأ على أساس معلومات مضللة، أو يكتب الطلاب مقالات باستخدام برنامج "شات جي بي تي"عند تكليفهم بواجب منزلي، وعلمت مؤخرًا أن موقع "سي نت" يستخدم منذ عدة أشهر أدوات الذكاء الاصطناعي الوليدة هذه لكتابة المقالات.

وعلى الرغم من ادعائهم أن المحررين البشريين قد قاموا بفحص الحقائق بدقة، إلا أنه اتضح أن الأمر لم يكن كذلك؛ حيث ينشر موقع "سي نت" مقالات كتبها الذكاء الاصطناعي دون كشف صحيح، حوالي 75 مقالة، وتبين أن بعضها يحتوي على أخطاء لم يكن من المحتمل أن يرتكبها كاتب بشري. لم تكن هذه حالة سوء نية؛ ولكن هذا هو نوع الخطر الذي يجب أن نكون أكثر قلقًا بشأنه وبشأن المسار الذي يسلكه الناس بسبب القيود العملية التي يواجهونها، وعندما تدمج ذلك مع حقيقة أن الأداة ليس لديها فكرة جيدة عن الحقيقة، فإن هذا ينبئ بكارثة.

أنجوين: لقد طورت تصنيفًا للذكاء الاصطناعي حيث تصف أنواعًا مختلفة من التقنيات التي تقع جميعها تحت مظلة الذكاء الاصطناعي. هل يمكنك إخبارنا بمكان برنامج "شات جي بي تي" في هذا التصنيف؟

نارايانان: برنامج "شات جي بي تي" جزء من فئة الذكاء الاصطناعي التوليدي. ومن الناحية التكنولوجية؛ يشبه هذا البرنامج إلى حد كبير نماذج تحويل النص إلى صورة، مثل نموذج "دال إي" (الذي ينشئ صورًا بناءً على تعليمات نصية من المستخدم). إنها مرتبطة بالذكاء الاصطناعي المستخدم في مهام الإدراك، ويستخدم هذا النوع من الذكاء الاصطناعي ما يسمى نماذج التعلم العميق. ومنذ حوالي عقد من الزمان؛ بدأت تقنيات الرؤية الحاسوبية تتحسن في التمييز بين القطة والكلب، وهو أمر يمكن للناس القيام به بسهولة بالغة.

ما اختلف في السنوات الخمس الماضية هو تحسن قدرة أجهزة الكمبيوتر على عكس مهمة الإدراك المتمثلة في التمييز بين القط والكلب، بسبب تقنية جديدة تسمى المحولات والتقنيات الأخرى ذات الصلة، وهذا يعني أنه يمكنهم في الواقع إنشاء صورة معقولة لقط أو كلب أو حتى أشياء خيالية مثل رائد فضاء يركب حصانًا. يحدث الشيء نفسه مع النصوص: لا يقتصر الأمر على قدرة النماذج على تصنيف قطعة من النص، فعند توجيهها، يمكن لهذه النماذج بشكل أساسي تشغيل التصنيف بشكل عكسي وإنتاج نص معقول قد يتناسب مع الفئة المحددة.

أنغوين: هناك فئة أخرى من الذكاء الاصطناعي التي تناقشها وهي أتمتة الحكم. هل يمكنك إخبارنا بما يتضمنه ذلك؟

نارايانان: أعتقد أن أفضل مثال على أتمتة الحكم هو الإشراف على المحتوى على وسائل التواصل الاجتماعي، لكن من الواضح أنه غير مكتمل بعد، فقد كان هناك الكثير من الإخفاقات الملحوظة في تعديل المحتوى، والعديد منها كان له عواقب مميتة، فقد استخدمت وسائل التواصل الاجتماعي للتحريض على العنف، وربما حتى عنف الإبادة الجماعية في أجزاء كثيرة من العالم، بما في ذلك ميانمار وسريلانكا وإثيوبيا. كانت هذه كلها إخفاقات في الإشراف على المحتوى؛ بما في ذلك الإشراف على المحتوى بالذكاء الاصطناعي.

رغم ذلك، يبدو أن الأمور تتحسن، إذ أنه من الممكن، على الأقل إلى حد ما، تدريب النماذج على إصدار الأحكام مثل وسطاء المحتوى البشريين لتحديد ما إذا كانت الصورة تمثل عُريًا أو كلامًا يحرض على الكراهية، لكن ستكون هناك دائمًا قيود. الإشراف على المحتوى مهمة مروعة؛ فهي وظيفة مليئة بالصدمات جراء النظر إلى صور الدماء وقطع الرؤوس وجميع أنواع الأفعال الفظيعة يومًا بعد يوم. بالتالي؛ إذا كان بإمكان الذكاء الاصطناعي تقليل العمالة البشرية في هذا المجال، فسيكون ذلك أمرًا جيدًا.

أعتقد أن هناك جوانب معينة من عملية تعديل المحتوى التي لا يجب أن تكون آلية، إذ يستغرق تحديد الخط الفاصل بين الكلام المقبول وغير المقبول وقتًا طويلًا، ويخلق فوضى، ويحتاج إلى إشراك المجتمع المدني، كما أنه يتغير باستمرار ويتماشى مع الثقافة، ويجب القيام بهذه العملية لكل نوع ممكن من الكلام. ونتيجة لكل ذلك، لا يمكن للذكاء الاصطناعي أن يلعب دورًا في كل هذا.

أنغوين: هناك فئة أخرى من الذكاء الاصطناعي التي تصفها وهي فئة تهدف إلى التنبؤ بالنتائج الاجتماعية، لكنك متشكك في هذا النوع من الذكاء الاصطناعي. لماذا؟

نارايانان: هذا هو نوع الذكاء الاصطناعي حيث يتنبأ صانعو القرار بما قد يفعله شخص ما في المستقبل ويستخدمون ذلك لاتخاذ قرارات بشأنهم، غالبًا لمنع أحداث معينة. ويتم استخدام هذا النوع في التوظيف، ويستخدم بكثرة في التنبؤ بالمخاطر الجنائية، كما يتم استخدامه أيضًا في السياقات التي تهدف إلى مساعدة شخص ما. على سبيل المثال؛ إذا كان شخص ما معرضًا لخطر الطرد من الكلية؛ يمكن للذكاء الاصطناعي أن يتدخل ويقترح عليهم تجربة تخصص مختلف.

ما تشترك فيه كل هذه الأمور هو التنبؤات الإحصائية المستندة إلى الأنماط التقريبية والارتباطات في البيانات حول ما يمكن أن يفعله الشخص، ثم تُستخدم هذه التنبؤات بدرجة ما لاتخاذ قرارات بشأنها، وفي كثير من الحالات، تحرمهم من فرص معينة وتحد من استقلاليتهم وتسلبهم فرصة إثبات أنفسهم وتأكيد حقيقة أنه لا يمكن تعريفهم من خلال الأنماط الإحصائية. هناك العديد من الأسباب الأساسية التي تدفعنا للاعتقاد بأن معظم تطبيقات الذكاء الاصطناعي هذه غير شرعية وغير مسموح بها أخلاقيًا.

وفي حال إجراء أي تدخل مبني على هذه التوقعات، يجب أن نتساءل: "هل هذا هو أفضل قرار يمكننا اتخاذه؟ أم أن القرار الأفضل هو الذي لا يتوافق مع التنبؤات على الإطلاق؟"

على سبيل المثال، في سيناريو التنبؤ بالمخاطر الجنائية، يكون القرار الذي نتخذه بناءً على التوقعات هو رفض الكفالة أو الإفراج المشروط، ولكن إذا ابتعدنا عن سياق التنبؤات فقد نتساءل: "ما هي أفضل طريقة لإعادة تأهيل هذا الشخص في المجتمع وتقليل فرصة ارتكابهم لجريمة أخرى؟" مما يتيح إمكانية إعتماد مجموعة أوسع بكثير من التدخلات.

أنغوين: يحذّر بعض الناس من "يوم القيامة" نتيجة برنامج شات جي بي تي، في ظل فقدان الوظائف وتراجع قيمة المعرفة. ما رأيك في الموضوع؟

نارايانان: فلنفترض أن بعض أكثر التوقعات جموحًا حول برنامج شات جي بي تي صحيحة وستقوم

بأتمتة فئات الوظائف بأكملها. على سبيل القياس؛ فكر في أكبر تطورات تكنولوجيا المعلومات في العقود القليلة الماضية، مثل الإنترنت والهواتف الذكية، التي أعادت تشكيل صناعات بأكملها، والتي تعلمنا التعايش معها. لقد أصبحت بعض الوظائف أكثر كفاءة، بينما تمت أتمتة بعض الوظائف الأخرى، مما تطلب من الناس تدريب أنفسهم من جديد أو تغيير وظائفهم. وعلى الرغم من وجود بعض الآثار الضارة لهذه التقنيات، إلا أننا نتعلم كيفية تنظيمها.

حتى في ظل تطورات مهمة مثل الإنترنت أو محركات البحث أو الهواتف الذكية، يتطلب منا الأمر التكيف معها؛ حيث نعظم الفوائد ونحاول تقليل المخاطر، بدًلا من أن تكون مجرّد ثورة، ولا أعتقد أن النماذج اللغوية الكبرى تندرج ضمن هذا النطاق حتى. ومن المحتمل أن تكون هناك تحولات ومزايا ومخاطر هائلة في العديد من الصناعات، لكن لا يمكنني تخيّل سيناريو تمثل فيه هذه التطورات مشكلة خطيرة.

المصدر: ذا مارك آب