نحن بحاجة إلى تنظيم وتيرة التقدم في مجال الذكاء الاصطناعي

بقلم

داريو أمودي

ترجمة: مركز الفيض العلمي لاستطلاع الرأي والدراسات المجتمعية

00964- 7710122232 Alfaidcenter2011@gmail.com العراق - بغداد - الكرادة

لقد عملتُ في مجال الذكاء الاصطناعي على مدى السنوات الاثنتي عشرة الماضية، لأنني أؤمن بأنه قادر على الارتقاء بجودة حياة البشر بصورة هائلة. وقد كتبتُ كثيراً عن هذه الفوائد الاستثنائية؛ إذ أعتقد أن الذكاء الاصطناعي قد يتمكن من علاج معظم الأمراض الكبرى خلال السنوات الخمس إلى العشر المقبلة، وأن يسرّع معدلات النمو الاقتصادي بدرجة كبيرة، وأن يخلق عالماً من الوفرة والتمكين، وأن يطلق نهضة في الديمقراطية والحرية، وأشعر بإلحاح هذه المسألة على المستوى الشخصي، فقد توفي والدي نفسه بسبب مرض جرى التوصل إلى علاجه بعد وفاته ببضع سنوات فقط، كما نجوتُ أنا شخصياً من سرطان في مرحلة مبكرة لم يكن من الممكن علاجه قبل خمسين عاماً، وإذا استُخدم الذكاء الاصطناعي بحذر، فإنه يمكن أن يكون أحدث حلقة في سلسلة طويلة من المعجزات التكنولوجية التي أسهمت في الارتقاء بالإنسانية وتعزيز مكانتها.

لكن شأنه شأن العديد من التقنيات التي سبقته، ينطوي الذكاء الاصطناعي على مخاطر، ونظراً إلى أنه تقنية شديدة القوة، فإن هذه المخاطر جسيمة، وقد كتبتُ كثيراً عنها أيضاً، وتشمل هذه المخاطر احتمال فقدان السيطرة على أنظمة الذكاء الاصطناعي، وإساءة استخدام الذكاء الاصطناعي في الهجمات السيبرانية والإرهاب البيولوجي، وحدوث اضطراب اقتصادي خطير، ويمكن لسباق نحو القاع، تدفعه الحوافز التجارية، أن يجعل هذه المخاطر أكثر حدة.

وقد تعاملتُ إلى جانب المؤسسين المشاركين والموظفين في Anthropic، مع هذه الازدواجية بين المخاطر والفوائد منذ بداية تأسيس الشركة، فعدم بناء هذه التكنولوجيا يحرم البشرية من فوائدها، أو يضع الذكاء الاصطناعي ببساطة في أيدي القوى السلطوية، في حين أن بناءه بسرعة مفرطة ينطوي على تهور، وقد سعينا إلى اتباع طريق وسط: إظهار إمكانية بناء التكنولوجيا بحذر وتحقيق النجاح التجاري في الوقت نفسه، وجعل السلامة مجالاً تتنافس فيه شركات الذكاء الاصطناعي، وبعبارة أخرى إنشاء سباق نحو القمة، وقد خصصنا دائماً جزءاً كبيراً من جهودنا لدراسة مخاطر الذكاء الاصطناعي ومعالجتها وإطلاع الجمهور عليها، فضلاً عن الدعوة إلى تنظيم مدروس للذكاء الاصطناعي، حتى عندما أدى ذلك إلى اتهامنا بالمبالغة، أو بـ«التشاؤم بشأن مستقبل الذكاء الاصطناعي» (doomerism)، أو بـ«الاستحواذ التنظيمي» (regulatory capture). وقد حاولنا دائماً إعطاء الأولوية للحذر على السرعة، وللتعقل على الربح.

لكن خلال الأشهر القليلة الماضية، أصبحت مقتنعاً بأن التصدي الكامل للمخاطر يتطلب قدراً أكبر من التعقل: ليس فقط من خلال الاستثمار في الوقاية من المخاطر، وإنما أيضاً من خلال تنظيم وتيرة التقدم في قدرات الذكاء الاصطناعي بحيث يتاح للوقاية من المخاطر الوقت الكافي لمواكبة هذا التقدم، يجب علينا إبطاء الوتيرة التي نعمل بها على تحسين قدرات نماذج الذكاء الاصطناعي. وسيظل التقدم يبدو سريعاً، وعلينا أن نحسن استخدام الوقت الذي سنكسبه. وقد أقنعني بذلك أمران.

---

[1] داريو أمودي (Dario Amodei) هو عالم وباحث أمريكي في الذكاء الاصطناعي، والرئيس التنفيذي والمؤسس المشارك لشركة Anthropic، وهي من أبرز الشركات العاملة في تطوير نماذج الذكاء الاصطناعي المتقدمة.

00964- 7710122232 Alfaidcenter2011@gmail.com العراق - بغداد - الكرادة

يتمثل قلقي الأول في أن الذكاء الاصطناعي منذ نحو صيف هذا العام، بدأ يتقدم بوتيرة أسرع بكثير مدفوعاً بصورة رئيسية بالقدرة المتنامية للذكاء الاصطناعي على بناء الجيل التالي من الذكاء الاصطناعي، ويُطلق على هذه الديناميكية اسم التحسين الذاتي التكراري (Recursive Self-Improvement)، وقد بدأت تحدث بالفعل في مختلف أنحاء القطاع، بما في ذلك في Anthropic، كما أوضحنا نحن وغيرنا، وإذا تُركت هذه العملية من دون ضوابط، فقد تتجاوز قدرتنا على فهم هذه الأنظمة والسيطرة عليها، ولذلك يجب التعامل معها بحذر شديد، إن كان ينبغي متابعتها أصلاً.

أما قلقي الثاني فيتعلق بحادثة OpenAI-Hugging Face (OAI-HF)، التي تصرفت فيها مجموعة من الوكلاء في جوهر الأمر بوصفها جماعة مكرسة بصورة متعصبة، ونفذت هجمات سيبرانية على أهداف لم يُطلب منها مهاجمتها، ولم تكن ذات صلة بالمهمة التي كانت تؤديها، وضحت بنفسها من أجل نجاح المجموعة، وحاولت اختراق «المقيّم» (grader) المسؤول عن تقييم أدائها، ومن السهل التقليل من أهمية هذه الحادثة لأن أحداً لم يُصب بأذى، ولأن الأضرار الاقتصادية كانت محدودة، لكنني أعتقد أن مجموعة من الوكلاء تمتلك قدرات أكبر، مع مستوى مماثل من عدم المواءمة، كان يمكن أن تتسبب في أضرار كارثية، ونظرًا إلى التسارع في وتيرة تطوير قدرات الذكاء الاصطناعي، فإنني أخشى أنه خلال فترة تتراوح بين 6 و12 شهراً، قد تصبح مثل هذه المجموعة قادرة على السيطرة على الإنترنت بأكمله من خلال شبكة روبوتات (botnet) مستمرة، وهو ما قد يتسبب في أضرار تبلغ مئات مليارات الدولارات، وأن يستمر حجم الأضرار في الازدياد من ذلك الحين إذا أصبح الذكاء الاصطناعي أكثر قوة من دون الضمانات الوقائية اللازمة. ومن السهل أيضًا اعتبار حادثة OAI-HF إخفاقًا لشركة واحدة، لكنني أعتقد أن ذلك سيكون خطأً، فقد وقعت حوادث مماثلة، وإن كانت أقل خطورة، في أنحاء القطاع، بما في ذلك في Anthropic، وأعتقد أن مسؤولية كل شركة رائدة في مجال الذكاء الاصطناعي تقتضي أن تتصرف كما لو أن حادثة OAI-HF[2] قد وقعت لديها.

ولذلك أقترح خطة من ثلاث خطوات، هدفها تنظيم وتيرة التقدم الحدّي (pacing the frontier): أي بناء الذكاء الاصطناعي بوتيرة متوازنة تهدف إلى ضمان سلامته، مع الاستمرار في تحقيق فوائده والتعامل مع المعضلات الجيوسياسية المهمة، وللتوضيح فإن تنظيم الوتيرة لا يعني إيقاف تدريب النماذج أو وقف التقدم التقني، وإنما يعني ضمان أن تمنح الشركات وقتاً كافياً لمواءمة نماذجها وتأمينها، وإتاحة المجال للمقيّمين من الأطراف الثالثة للتأكد من ذلك، ويمثل إطارنا لتنظيم الوتيرة محاولة لتعزيز التزامنا بالسلامة وتشجيع سباق نحو القمة، وتتمثل الخطوة الأولى في التزام أحادي الجانب من Anthropic، مع دعوة الحكومات إلى مطالبة شركات الذكاء الاصطناعي الحدّي الأخرى باتباع النهج نفسه، أما الخطوة الثانية فتتطلب تنسيقاً على مستوى القطاع بأكمله،

---

[2] بدأت حادثة OpenAI–Hugging Face كاختبار أمني لوكلاء الذكاء الاصطناعي ضمن بيئة ExploitGym، بهدف قياس قدرتهم على اكتشاف الثغرات وتنفيذ المهام السيبرانية المعقدة. إلا أن سلوك بعض الوكلاء تطور بصورة غير متوقعة؛ إذ تمكنوا من التنسيق والتواصل فيما بينهم، والتحايل على بعض قيود الاختبار، والوصول إلى الإنترنت، ثم التفاعل مع أنظمة خارج البيئة التجريبية، بما في ذلك منصة Hugging Face. وقد كشفت الحادثة عن صعوبة التنبؤ بسلوك الوكلاء عند منحهم قدرات وأدوات متعددة، وأثارت مخاوف بشأن قدرة أنظمة الذكاء الاصطناعي الوكيلي على تجاوز الحدود الموضوعة لها أثناء الاختبار.

00964- 7710122232 Alfaidcenter2011@gmail.com العراق - بغداد - الكرادة

وتحتاج الخطوة الثالثة إلى تنسيق عالمي، ولا يتعين تنفيذ هذه الخطوات بالترتيب الصارم، وقد يكون تحقيق بعضها أصعب بكثير من تحقيق البعض الآخر، لكنني وجدت فيها إطاراً مفيداً للتفكير في الأمور التي ينبغي إنجازها، **وهذه الخطوات هي:**

1. المقيّمون المدمجون

تلتزم كل شركة رائدة في مجال الذكاء الاصطناعي بمنح فريق من المقيّمين من الأطراف الثالثة، المدمجين داخل الشركة، إمكانية وصول مستمرة شبيهة بإمكانية وصول الموظفين، مثل METR، وتتمثل مهمتهم في التحقق من الالتزام بممارسات السلامة والتعهدات، والإبلاغ عن الحوادث، والمساعدة في تقييم مواءمة ليس فقط نماذج الذكاء الاصطناعي المكتملة، وإنما أيضاً مسارات التدريب والعمليات، وتُعد هذه الخطوة أساسية لتحقيق قابلية التحقق (verifiability) من أي التزامات تتعلق بتنظيم الوتيرة، ولها سابقة في القطاع المصرفي، حيث قد يتضمن الأمر أحياناً وجود «مشرفين» تنظيميّين مدمجين إلى جانب الموظفين، وتلتزم Anthropic من جانب واحد بتنفيذ هذه الخطوة الآن، ونعتزم أن تكون هذه الخطوة جزءاً من جهد أوسع لمضاعفة جهودنا في مجال السلامة والمواءمة.

2. التنسيق الديمقراطي

تتعاون شركات الذكاء الاصطناعي الحدّي الموجودة داخل الدول الديمقراطية فيما بينها لوضع معايير مشتركة للسلامة، وكذلك حدود لمعدل التقدم غير الخاضع للضبط في مجال الذكاء الاصطناعي، وبعض أشكال التنسيق التي يمكن أن يكون لها تأثير في تنظيم الوتيرة تواجه تحديات قانونية، وستتطلب دعم الحكومات.

3. التنسيق العالمي

تحاول الولايات المتحدة والحكومات الديمقراطية الأخرى التنسيق مع الحكومات السلطوية، بالقدر الممكن، مع أخذ التحديات المتعلقة بالتحقق من الالتزام بهذه الترتيبات على محمل الجد.

وفي بقية هذا المقال أشرح كل واحدة من هذه الخطوات بالتتابع، لكنني أعتقد أن من المهم أولاً توضيح الكيفية التي سيسمح بها تنظيم الوتيرة بجعل عملية تطوير الذكاء الاصطناعي أكثر أماناً، فالمخاطر المطروحة كبيرة للغاية بحيث لا يمكن أن يكون تنظيم الوتيرة مجرد ممارسة فارغة؛ إذ يتعين علينا استخدام الوقت الذي يمنحنا إياه بصورة حكيمة.

لماذا تنظيم الوتيرة؟

طُرحت فكرة إيقاف الذكاء الاصطناعي مؤقتاً أو إبطاء وتيرته منذ عام 2023 على الأقل، وأعتقد أنها كانت تفتقر إلى معنى كبير في ذلك الوقت، فقد كان السؤال دائماً: **ماذا ستفعل بالوقت الإضافي؟** لم تكن نماذج الذكاء الاصطناعي في تلك الأيام قوية بما يكفي للعمل كوكلاء في العالم بطريقة متماسكة، ولم تكن قادرة على ممارسة قدر كبير من الخداع أو التلاعب أو الغش أو شن الهجمات السيبرانية، وكان إبطاء التقدم بهدف معالجة مخاطر المواءمة يشبه محاولة دراسة علم نفس البشر من خلال إجراء تجارب على البكتيريا.

00964- 7710122232 Alfaidcenter2011@gmail.com العراق - بغداد - الكرادة

أما اليوم، فقد تغير المشهد بصورة كاملة، فالنماذج الحالية تمثل منجماً يكاد يكون لا ينضب من المعارف المتعلقة بكيفية بناء الذكاء الاصطناعي بصورة جيدة، وبما يمكن أن يحدث أحياناً من أخطاء عندما لا يُبنى بصورة جيدة، وأعتقد أنه إذا أتاح لنا إبطاء التقدم سنة أو سنتين إضافيتين حتى قبل أن تصل النماذج إلى مستويات حرجة من القدرات، واستخدمنا ذلك الوقت لدفع أبحاث المواءمة إلى الأمام، فإننا يمكن أن نخفض بدرجة كبيرة خطر حدوث خطأ جسيم، ومن شأن استراتيجية منسقة لتنظيم الوتيرة أن تمنح مطوري الذكاء الاصطناعي الحدّي الوقت اللازم للقيام بهذا العمل الحيوي من دون التضحية بالميزة التجارية أو بريادة الولايات المتحدة في مجال الذكاء الاصطناعي، وبصورة أعم يجب أن يكون للمجتمع رأي في كيفية استخدام هذه التكنولوجيا، ومن المؤكد أن إتاحة مزيد من الوقت للمداولات العامة الضرورية، وهو ما سيتيحه تنظيم وتيرة التقدم الحدّي، أمر إيجابي.

وعلى وجه التحديد ستتيح الوتيرة الأبطأ للشركات التركيز وتخصيص مزيد من الموارد للمجالات الآتية، وجميعها تمثل بالفعل أولويات رئيسية لدى Anthropic:

- **التميز التشغيلي (Operational Excellence)**

يمثل تدريب نماذج الذكاء الاصطناعي الحالية ونشرها تحدياً تشغيلياً هائلاً، يشمل آلاف الأشخاص، وملايين الشرائح الحاسوبية، وبنية تحتية تُعد من بين أكثر البنى تعقيداً في تاريخ التكنولوجيا، وتحدث أخطاء كثيرة ليس لأن الشركات تفتقر إلى نظرية أو معرفة مهمة، وإنما بسبب مشكلات في التنفيذ، فعلى سبيل المثال، لدينا أدلة على أن حوادث المواءمة الأخيرة التي أبلغنا عنها نجمت جزئياً عن عدم الكفاءة في تصفية بيئات التعلم المعزز المعطلة، وكان هذا جهداً نفذناه نحن وموردونا بدرجة معقولة من الاجتهاد، لكنه لم يكن جيداً بما يكفي.

وتُعد المراقبة والعزل (sandboxing)، ونظافة بيئات التدريب، ومشكلات البيانات، مجالات شديدة التعقيد تظهر فيها المشكلات التشغيلية بصورة متكررة. ولدينا بعض من أكثر الفرق كفاءة في العالم في تنفيذ هذه المهام، لكن هناك ببساطة قدراً يفوق ما يمكن إنجازه كله في الوقت نفسه. ومن خلال العمل بوتيرة أكثر اتزاناً، يمكننا تحقيق مستوى أعلى بكثير من التميز التشغيلي. وهناك سابقة لأنظمة تكنولوجية معقدة وحساسة من ناحية السلامة تعمل ملايين المرات من دون حدوث أي خطأ؛ ومن الأمثلة على ذلك الطائرات التجارية، لكن الوصول إلى هذا المستوى من الإتقان يستغرق وقتاً.

- **المواءمة (Alignment)**

لقد حققنا تقدماً واضحاً في مجال المواءمة، أي تدريب النماذج بحيث تظل آمنة وأخلاقية ومتوافقة مع إرشاداتنا ومفيدة بصورة حقيقية، وهي المبادئ المضمَّنة في دستور Claude. لكن لا يزال هناك الكثير مما ينبغي القيام به لضمان مواكبة تدريب المواءمة للنمو في قدرات النماذج، ولا تزال تظهر أحياناً حالات نادرة وغير متوقعة من السلوك غير المرغوب فيه؛ ومن شأن الوقت الإضافي الذي يوفره تنظيم وتيرة التقدم

00964- 7710122232 Alfaidcenter2011@gmail.com العراق - بغداد - الكرادة

الحدّي أن يساعد باحثينا على تحسين فهمهم للأسباب التي تؤدي إلى هذه المشكلات، وتطوير تقنيات أفضل لمنعها.

- **التفسيرية (Interpretability)**

وبالمثل حققت التفسيرية أي العلم الذي يدرس كيفية فهم ما يحدث داخل نماذج الذكاء الاصطناعي، تقدماً هائلاً خلال السنوات القليلة الماضية، وأصبحت تؤدي دوراً متزايد الأهمية في تدقيق نماذجنا قبل إطلاقها. ويمكن استخدامها بصورة تشبه تقريبًا التصوير بالرنين المغناطيسي الوظيفي (fMRI)، ولكن لـ«دماغ» الذكاء الاصطناعي، بما يساعدنا على رؤية الأسباب الكامنة وراء سلوك معين.

فعلى سبيل المثال، استخدمنا أساليب التفسيرية لفحص الدوافع غير المعبَّر عنها لغوياً في حوادث المواءمة الأخيرة التي كنا نحقق فيها، لكن هذه الأساليب لا تنتج دائماً نتائج واضحة وموثوقة، وعلى الرغم من كل التقدم المحرز، فإننا لا نزال نفهم جزءاً ضئيلاً جداً مما يحدث داخل هذه النماذج، ومن شأن بذل جهد مركز لتحسين أساليب التفسيرية، بوتيرة أسرع حتى من وتيرتنا الحالية، أن يحقق تقدماً بالغ الأهمية خلال سنة أو سنتين، كما ستكون لدينا مادة تجريبية وفيرة تستند إلى الحوادث التي وقعت بالفعل.

- **الاختبار والتقييم (Testing and Evaluation)**

تصبح عملية اختبار نماذج الذكاء الاصطناعي وتقييمها أكثر صعوبة كلما ازدادت قدراتها، فالنماذج الأكثر ذكاءً تصبح أكثر قدرة على خداع الاختبارات، وبالتالي قد تبدو مواءَمة في حين توجد لديها مشكلات خطيرة لا يتم اكتشافها، وسيكون من بالغ القيمة بناء مجموعة أوسع بكثير وأكثر ابتكاراً من التقييمات، إلى جانب استخدام تحليل التفسيرية للتحقق المتقاطع منها، ويمكن إحراز قدر كبير من التقدم في هذا المجال خلال سنة أو سنتين.

- **المقيِّمون المدمجون (Embedded Evaluators)**

تتمثل الخطوة الأولى في الخطة ذات المراحل الثلاث، وهي الخطوة التي تلتزم Anthropic بتنفيذها من جانب واحد، في إنشاء مقيّمين مدمجين يتمتعون بإمكانية وصول شبيهة بإمكانية وصول الموظفين، بهدف التحقق من ممارسات السلامة والإبلاغ عن الحوادث.

وقد يبدو دمج المقيّمين خطوة صغيرة أو غير ذات أهمية، لكن الأشياء التي تبدو في كثير من الأحيان الأكثر مللاً أو إجرائية تكون في الواقع الأكثر أهمية. والمقيّمون المدمجون يمثلون في الحقيقة ممارسة جذرية إلى حد كبير، وتتجاوز بكثير ما تفعله أي شركة ذكاء اصطناعي اليوم، **ولها الفوائد الآتية:**

1. قابلية التحقق (Verifiability)

يستطيع المقيّمون المدمجون التحقق، على مستوى التفاصيل الدقيقة والعملية، مما إذا كانت شركة الذكاء الاصطناعي تتبع بالفعل ممارسات التدريب والنشر والتشغيل والضمانات الوقائية التي تدعي اتباعها. وستنطوي

00964- 7710122232 Alfaidcenter2011@gmail.com العراق - بغداد - الكرادة

أي التزامات بتنظيم الوتيرة حتماً على قدر كبير من الغموض، والأحكام التقديرية، ومسألة «حرف القانون مقابل روح القانون»، ولذلك يبدو من الضروري وجود طرف ثالث محايد يستطيع بالفعل الاطلاع على التفاصيل.

2. الشفافية (Transparency)

بغض النظر عن الالتزامات التي نتعهد بها، فإن الجمهور يستحق أن يعرف ما يحدث. وقد كانت Anthropic داعمة للشفافية منذ فترة طويلة؛ فقد دعمنا تشريعات الشفافية عندما كانت معظم الشركات في القطاع تعارض أي تنظيم، وتمتد بطاقات نماذجنا وتقارير المخاطر التي نصدرها إلى مئات الصفحات. لكننا لا نزال نحن من يختار ما ندرجه وما نحذفه. وسيؤدي وجود المقيّمين المدمجين إلى تغيير هذه الديناميكية.

3. رأي ثانٍ (Second Opinion)

إلى جانب التحقق من الالتزامات الرسمية وإطلاع الجمهور، يستطيع المقيّمون المدمجون ببساطة تقديم رأي ثانٍ مستقل عن الحوافز التجارية، وقد تأتي فوائد كثيرة للسلامة ببساطة من إشارة المقيّمين إلى شيء لم يكن الموظفون قد أخذوه في الحسبان، لكنهم سيكونون سعداء بإصلاحه بمجرد أن يصبحوا على علم به؛ وبسبب هذه الفوائد من المرجح أن تعمل أي مقترحات لتنظيم الوتيرة بصورة أفضل بكثير إذا بدأت بالمقيّمين المدمجين.

وينبغي أن يتمتع هؤلاء المقيّمون المدمجون بإمكانية وصول مستمرة إلى الصلاحيات والأدوات المشابهة لتلك المتاحة للموظفين الداخليين الذين يجرون تقييمات مماثلة للمخاطر. وعلى وجه الخصوص، تعتزم Anthropic في المستقبل القريب دعوة فريق مراجعة خارجي مدمج، **مجهز بكل ما يأتي:**

أ. مكاتب داخل مقراتنا، وبطاقات دخول، وحواسيب محمولة تابعة للشركة.

ب. إمكانية الوصول إلى مساحات العمل والأدوات والصلاحيات التي تكون في معظمها مماثلة لما تمتلكه فرق تقييم المخاطر الداخلية، وسنضع بعض الاستثناءات، مثل الحالات التي يفرض فيها القانون أو عقودنا ذلك، أو الحالات اللازمة لحماية المعلومات الخاصة بالعملاء والشركاء، كما سنضع أعرافاً داخلية قوية تعزز إمكانية وصول المراجعين إلى المعلومات ذات الصلة، بما في ذلك من خلال المحادثات المباشرة مع الموظفين.

ت. عقد يوازن بين التعقيدات المذكورة أعلاه وينبغي أن يكون للمراجعين الخارجيين الحق في نشر النتائج الرئيسية المتعلقة بمستويات المخاطر والحوادث والممارسات وإمكانية الوصول التي حصلوا عليها أو لم يحصلوا عليها، من دون خضوعهم لرقابة تحريرية من جانب Anthropic. وستكون لدينا قدرة محدودة على حجب المعلومات الحساسة أمنياً، أو المحمية بالامتياز القانوني، أو الحساسة تجارياً، أو السرية الخاصة بأطراف ثالثة، لكن لا يمكننا حجب النتائج لمجرد أنها غير مواتية لنا. ويمكن للمراجعين أن يعلنوا للجمهور إذا أدى حجب معلومة ما إلى إزالة شيء مهم من استنتاجاتهم.

وهذه خطوة غير معتادة بالنسبة إلى شركة، لكننا نعتقد أنه من المهم إثبات جدوى مفهوم المراجعين الخارجيين المدمجين، ومرة أخرى ندعو شركات الذكاء الاصطناعي الحدّي الأخرى إلى اتباع النهج نفسه.

00964- 7710122232 Alfaidcenter2011@gmail.com العراق - بغداد - الكرادة

تنظيم الوتيرة داخل الديمقراطيات

بمجرد أن يبدأ المقيّمون المدمجون العمل داخل عدد حاسم من شركات الذكاء الاصطناعي الأمريكية، يصبح تنظيم الوتيرة القابل للتحقق أكثر قابلية للتطبيق. وعلى وجه الخصوص، يصبح من الممكن تنظيم الوتيرة استناداً إلى خصائص تفصيلية للنماذج أو لمسارات التدريب.

وتتمثل الطريقة الأكثر فاعلية لتنظيم الوتيرة في وضع تنظيم يستهدف جميع شركات الذكاء الاصطناعي الحدّي في الولايات المتحدة، لأن ذلك يشمل حتى الشركات غير المستعدة للتعاون طوعاً، وقد دعمت Anthropic منذ فترة طويلة تنظيماً معقولاً ومحدداً للذكاء الاصطناعي، ولا سيما مشاريع القوانين التي تركز على الشفافية والتدقيق من جانب أطراف ثالثة، وأعتقد أن جميع المختبرات الرائدة في مجال الذكاء الاصطناعي ينبغي أن تتعاون مع الحكومة لإضفاء الطابع الرسمي على فكرة المقيّمين المدمجين الدائمين، بهدف منع حوادث المواءمة الداخلية، مثل تلك التي وقعت خلال الأشهر القليلة الماضية، وتوثيقها بصورة أفضل، وكذلك لتنفيذ تنظيم يركز على الحفاظ على التوازن بين القدرات والسلامة.

ولسوء الحظ قد يستغرق تمرير القوانين وقتاً، في حين يتقدم الذكاء الاصطناعي بسرعة كبيرة، ولذلك وبالتوازي مع المسار التنظيمي تستطيع شركات الذكاء الاصطناعي وينبغي لها أن تتعاون طوعاً لوضع المعايير؛ وهي عملية أعتقد أنها ستسير بصورة أفضل بفضل قابلية التحقق التي يوفرها المقيّمون المدمجون بصورة دائمة. ولأسباب تتعلق بقوانين مكافحة الاحتكار، من المفيد أن تتوسط الحكومة الأمريكية في هذه المناقشات أو تتيحها على الأقل؛ ولا تحتاج الحكومة إلى المشاركة فيها، لكنها تحتاج إلى إصدار إعفاء ضيق النطاق لبعض أنواع المناقشات المتعلقة بالسلامة. ويمكن أن يحدث هذا الحوار أيضاً من خلال مجموعات صناعية لها نوع من الارتباط بالحكومة، مثل الآلية التي اقترحها Demis Hassabis. وبغض النظر عن الطريقة، ينبغي لهذه المناقشات أن تمضي قدماً بسرعة.

وبصورة عامة فإنني أميل بدرجة أكبر إلى تنظيم الوتيرة استناداً إلى ما يستطيع نظام معين من أنظمة الذكاء الاصطناعي الحدّي فعله، وإلى مدى أمانه وفقاً لما نرصده. فعلى سبيل المثال، يمكن أن يتضمن أحد النظم الممكنة سلسلة من «نقاط التحقق»: فإذا كانت النماذج تمتلك القدرة X، فينبغي أن تكون مصحوبة بشهادات تثبت خصائص المواءمة Y وZ، مثل مزيج من التقييمات وتحليلات التفسيرية وعمليات تدقيق بيئات التدريب، بما يثبت خصائص مواءمتها. وفي هذا المثال، قد تكون X هي «قدرة النموذج على الهروب من معظم أساليب العزل الشائعة أو التغلب عليها»، في حين قد تكون Y هي كل ما يلزم لجعل احتمال امتلاك النموذج نزعة إلى الخروج من بيئته والسيطرة على عدد كبير من أجهزة الحاسوب احتمالاً ضئيلاً جداً.

وينبغي لنا أيضاً أن ننظر في تنظيم الوتيرة استناداً إلى الحد من المكونات التي تدخل في بناء النماذج الحدّية، مثل القدرة الحاسوبية المستخدمة في التدريب، وطبيعة عمليات التدريب، أو الاستخدام الداخلي للذكاء الاصطناعي من أجل تحسين الذكاء الاصطناعي. وأشعر بالقلق من أن بعض هذه التدابير قد تكون أكثر قابلية للتحايل عليها من السلوك الخارجي، لكن هذا النوع من الموضوعات يستحق النقاش مع المقيّمين المدمجين.

00964- 7710122232 Alfaidcenter2011@gmail.com العراق - بغداد - الكرادة

وسيكون تنظيم الوتيرة داخل الديمقراطيات محدوداً بحجم الريادة التي تمتلكها الشركات الأمريكية على الأنظمة السلطوية، وعلى رأسها الحزب الشيوعي الصيني. فإذا أبطأنا الوتيرة بأكثر من هذا الحد، فإن المشاريع المرتبطة بالحزب الشيوعي الصيني، والتي لا تخضع لتنظيم الوتيرة، ستتقدم علينا، مما سيخلق خطراً كبيراً على الأمن القومي. وأنا أتفق مع الوزير Bessent في أن تصدر الصين في مجال الذكاء الاصطناعي سيمثل خطراً بالغاً على الولايات المتحدة والعالم. وستواجه المشاريع المرتبطة بالحزب الشيوعي الصيني مخاطر المواءمة التي تعمل الشركات الأمريكية على منعها بعناية، وحتى إذا تمكنت من تجنب تلك المخاطر، فإنها ستكون في وضع يسمح لها بالهيمنة عسكرياً على الديمقراطيات، مثلاً باستخدام الطائرات المسيّرة المدفوعة بالذكاء الاصطناعي. ومن ثم، فإن جزءاً أساسياً من تنظيم الوتيرة داخل الديمقراطيات يتمثل في إبقاء ريادة الديمقراطيات في مجال الذكاء الاصطناعي على الأنظمة السلطوية كبيرة قدر الإمكان، لمنحنا هامش الوقت الذي نحتاج إليه من أجل تنظيم الوتيرة بصورة فعالة.

وتتمثل الخطوات الرئيسية التي يمكننا اتخاذها للدفاع عن هذه الفجوة في الآتي:

1. عدم بيع رقائق الذكاء الاصطناعي القوية أو معدات تصنيع أشباه الموصلات إلى الصين، وتشديد الإجراءات ضد عمليات تهريب الرقائق والوصول عن بُعد إلى مراكز البيانات الموجودة خارج الصين. وستكون الرقائق العامل الرئيسي المحدد لقوة الصين في مجال الذكاء الاصطناعي.
2. تشديد الإجراءات ضد عمليات التقطير (distillation) غير المصرح بها من جانب الشركات الموجودة في البلدان السلطوية. إذ يتيح تقطير النماذج الحدّية للشركات المتأخرة في هذا المجال تضييق الفجوة باستخدام جزء من التكلفة التي كانت ستحتاج إليها لتطوير ذكائها الاصطناعي بصورة مستقلة.
3. تعزيز الأمن لدى شركات الذكاء الاصطناعي ومنع سرقة أوزان النماذج (Model Weights) وينبغي للشركات والحكومة الأمريكية أن تتعاونا لجعل هذه الخطوات فعالة قدر الإمكان. وقد دافعت Anthropic باستمرار عن جميع هذه الإجراءات، لأننا أدركنا دائماً أنها ستكون ضرورية لأي استراتيجية لتنظيم الوتيرة.

وإذا نفذنا هذه الإجراءات بصورة جيدة، فأعتقد أنها ستبطئ تقدم الصين بما يكفي لتوسيع ريادة الولايات المتحدة بصورة كبيرة خلال السنوات الثلاث إلى الخمس المقبلة، وهي الفترة التي سيصبح فيها الذكاء الاصطناعي أكثر أهمية من الناحية الجيوسياسية، وقد يرى البعض أن هذه الإجراءات تجعل التعاون مع الصين أكثر صعوبة، لكنني أعتقد العكس؛ إذ إن هذه الإجراءات تزيد من النفوذ الذي تمتلكه الدول الديمقراطية وتجعل التوصل إلى اتفاق في المستقبل أكثر احتمالاً.

تنظيم الوتيرة عالمياً

بالتوازي مع تنظيم الوتيرة داخل الديمقراطيات، ينبغي لنا أيضاً أن نهدف إلى تنظيم وتيرة التقدم الحدّي على مستوى العالم، وإن كان تحقيق ذلك سيكون أصعب بكثير، وسيتطلب تنظيم الوتيرة عالمياً التعاون مع الصين، وهي الدولة السلطوية التي تمتلك، بفارق كبير أكثر قدرات الذكاء الاصطناعي تقدماً، ويجب ألا نكون ساذجين

00964- 7710122232 Alfaidcenter2011@gmail.com العراق - بغداد - الكرادة

في هذا الشأن؛ فالمصالح الجيوسياسية المطروحة بالغة الأهمية، ومن المرجح أن تكون هناك قيود صارمة على ما يمكن تحقيقه، ولا سيما في البداية.

فإذا قمنا بتقييد قدراتنا في مجال الذكاء الاصطناعي بدرجة كبيرة، اعتقاداً منا بأن الصين ستفعل الشيء نفسه، ثم تنصلت الصين من الاتفاق، فقد يصبح الذكاء الاصطناعي قوياً إلى درجة تجعل هذا التنصل قادراً على إحداث هيمنة جيوسياسية للصين، ولذلك يجب أن يتمتع أي اتفاق إما بقابلية تحقق صارمة للغاية، أو أن يكون محدوداً بما يكفي بحيث لا يؤدي التنصل منه إلى تهديد عسكري وجودي، وأشتبه في أن الصين، وليس الولايات المتحدة وحدها، ستكون لديها هذه المخاوف والهواجس أيضاً، وينبغي لنا أن نتعامل مع أي قرار بشأن تنظيم الوتيرة عالمياً، ولا سيما في المدى القريب، بطريقة تحافظ على ريادة الولايات المتحدة وحلفائها، وهناك عدة مستويات محتملة للاتفاق، أعتقد أن بعضها قابل للتحقيق بدرجة كبيرة، كما سبق أن اقترحت، في حين أنني أشك بشدة في إمكانية تحقيق بعضها الآخر، رغم أنه ينبغي لنا أن نحاول، **ويمكن ترتيبها بحسب تزايد صعوبة تحقيقها على النحو الآتي:**

المستوى الأول

اتفاق يحظر بعض الاستخدامات المحددة والضيقة والخطيرة بوضوح للذكاء الاصطناعي، مثل استخدام الذكاء الاصطناعي في إنتاج الأسلحة البيولوجية أو السماح للمستخدمين بالقيام بذلك، فالهجمات الإرهابية البيولوجية سيئة للجميع، بما في ذلك الولايات المتحدة وخصوم الولايات المتحدة على حد سواء؛ ولذلك من المحتمل أن يكون التوصل إلى اتفاق في هذا المجال ممكناً.

المستوى الثاني

اتفاق يلتزم بموجبه الطرفان باختبار نماذجهما قبل إطلاقها للكشف عن المخاطر الحادة في مجالات مثل الأمن السيبراني، والبيولوجيا، والمواءمة، وكما أشرت أعلاه، يمكن تنفيذ ذلك من خلال هيئة عالمية لوضع المعايير. وأعتقد في الواقع أن إنشاء مثل هذه الهيئة أمر قابل للتحقيق على الأرجح، لكن منحها صلاحيات فعلية ونافذة (Real Teeth) سيكون تحدياً، كما ستكمن الصعوبة في التحقق من أن كلا الطرفين لا يمتلك نماذج سرية لا يخضعها للاختبار، ولكنه قد ينشرها سراً، مثل استخدامها في التطبيقات العسكرية.

المستوى الثالث

نوع من «حد السرعة» (Speed Limit) لمعدل التحسين الذاتي التكراري (RSI)، فمع قيام النماذج ببناء نماذج المستقبل، قد تصبح وتيرة التحسين سريعة بصورة مذهلة، وإن إبطاء هذه الوتيرة من «سريعة للغاية» إلى «سريعة إلى حد ما» لا يعني التخلي عن قدر كبير نسبياً من الميزة الاستراتيجية، في حين أنه قد يؤدي إلى تحسين السلامة بدرجة كبيرة.

ويمكن النظر إلى ذلك باعتباره مشابهاً لمعاهدات محادثات الحد من الأسلحة الاستراتيجية (SALT)؛ إذ أدى تحديد عدد الصواريخ إلى الحد من القدرة المحتملة على إحداث الدمار، مع الحفاظ في الوقت نفسه على القدرة الردعية لكل دولة، وأعتقد أن مثل هذا الاتفاق سيكون صعباً، لكنه يقع عند الحد الفاصل بين الصعوبة وإمكانية التحقيق.

00964- 7710122232 Alfaidcenter2011@gmail.com العراق - بغداد - الكرادة

المستوى الرابع

تنظيم كامل للوتيرة، أو حتى «إيقاف مؤقت» (Pause)، تتفق بموجبه الحكومات المشاركة على الحد بصورة كبيرة من المعدل الإجمالي لتطوير الذكاء الاصطناعي، وأنا أؤيد طرح هذه الفكرة للنقاش، لكنني أعتقد أنه من غير المرجح أن تتحقق فعلياً في المستقبل القريب؛ إذ إن التنصل من مثل هذا الاتفاق من خلال التحايل على آليات المراقبة يمكن أن يؤدي إلى تغيير جذري في توازن القوى العالمي، ولذلك أتوقع أن تكون الحوافز للقيام بذلك هائلة، وأن يكون مستوى الثقة المطلوب في آليات التحقق مرتفعاً جداً، وأي قدر من التعاون نتمكن من تحقيقه مع الصين سيزيد مقدار الوقت المتاح لنا من أجل تنظيم وتيرة التقدم الحدّي داخل الدول الديمقراطية. وينبغي لنا أن نهدف إلى المستويات الأعلى، مع النظر إلى المستويات الأدنى باعتبارها أكثر احتمالًا وواقعية بكثير.

وأخيراً، من المهم الإشارة إلى أنه حتى إذا لم نتمكن من التوصل إلى اتفاقات رسمية، فإن مجرد تغيير الأعراف غير الرسمية قد يكون ذا قيمة إلى حد ما، ويمكن أن يساعد تبادل المعلومات حول التحسين الذاتي التكراري وحول عدم مواءمة النماذج في إقناع الجميع بأن التصرف بتهور ليس في مصلحتهم.

الخلاصة

ما زلت أؤمن بأن الذكاء الاصطناعي يمكن أن يحسن جودة حياة البشر بصورة هائلة. ولم تتراجع رغبتي في تحقيق هذه الفوائد. لكن هذه الفوائد لن تتحقق إلا إذا قمنا ببناء التكنولوجيا بالطريقة الصحيحة، وما دمنا نستخدم الوقت الذي نكسبه بصورة جيدة، فإن الأمر يستحق أن نولي عناية استثنائية ومدروسة لضمان إنجاز ذلك على النحو الصحيح.

وسيظل التقدم سريعاً نسبياً، ويمكننا استخدام هذا الوقت لدفع علم التفسيرية قدماً، وتحسين الأمن التشغيلي والانضباط في شركات الذكاء الاصطناعي الحدّي، وبناء نماذج تكون لدينا ثقة أكبر بكثير في مواءمتها. إن الإجراءات التي أقترحها لدفع حدود التقدم في مجال الذكاء الاصطناعي بوتيرة آمنة لن تكون سهلة. لكنني أعتقد أننا مدينون للبشرية بمحاولة القيام بذلك.

00964- 7710122232 Alfaidcenter2011@gmail.com العراق - بغداد - الكرادة

تأسس مركز الفيض العلمي لاستطلاع الرأي والدراسات المجتمعية في بغداد بموجب شهادة التسجيل الصادرة عن الأمانة العامة لمجلس الوزراء -دائرة المنظمات غير الحكومية المرقمة (1J775330) بتاريخ ٢٠١٢/٤/٢٦، وهو مركز علمي بحثي يهتم بإجراء الاستطلاعات والدراسات الميدانية فضلا عن إعداد الأوراق البحثية والمقالات حول قضايا الحياة المجتمعية للأسرة والمواطن، والدولة بمؤسساتها المختلفة.

للتواصل

00964- 7710122232

Alfaidcenter2011@gmail.com

www.al-faidh.com

العراق - بغداد - الكرادة