Helen Toner

Com regular la IA, encara que sigui difícil de predir

TED Talk celebrat a Vancouver el 16 d'abril de 2024

Fotografia: Gilberto Tadday / TED

Helen Toner és australiana i la directora d'estratègia i beques del Centre de Seguretat i Tecnologia Emergent de Georgetown. És investigadora i experta en Intel·ligència Artificial. Ha format part de diferents iniciatives altruistes i va ser una de les responsables de l'acomiadament de Sam Altman d'OpenAI, una decisió que finalment va comportar el seu cessament de la direcció. Toner defensa més control en la IA per poder seguir-ne el desenvolupament.

Aquest és un estat de coses força estrany. Normalment, les persones que basteixen una nova tecnologia entenen com funciona per dins i per fora. Però amb la IA, una tecnologia que està remodelant radicalment el món que ens envolta, no és així. Els experts saben força de com crear i executar els sistemes d'IA, és clar. Però quan es tracta de com funciona per dins, hi ha serioses limitacions del que en sabem.

I importa, perquè sense entendre profundament la IA, és molt difícil saber de què serà capaç més endavant, i fins i tot què pot fer ara. I el fet que ens costi tant d'entendre el que passa amb la tecnologia i predir cap a on anirà després, és un dels majors obstacles als quals ens enfrontem per determinar com regular la IA. Però la IA ja està a tot arreu, de manera que no podem quedar-nos asseguts i esperar que les coses s'aclareixin soles. Hem de traçar, de totes maneres, alguna mena de camí cap a endavant.

He estat treballant uns vuit anys en aquestes qüestions de polítiques i regulació de la IA, primer a San Francisco, i ara a Washington DC. Al llarg d'aquest temps, m'he fet una idea de primera mà de com s'està treballant des dels governs per gestionar aquesta tecnologia. I a la indústria també n'he vist unes quantes coses. Així que compartiré un parell d'idees sobre com podria ser el nostre camí a seguir per regular la IA. Però, primer, parlem del que realment provoca que la IA sigui tan difícil d'entendre i predir.

Un gran repte a l'hora de construir la IA és que ningú es posa d'acord pel que fa a què significa exactament ser intel·ligent. Aquesta és una situació ben estranya a l'hora de bastir una nova tecnologia. Quan els germans Wright van començar a experimentar amb els avions, no sabien com construir-ne, però tothom sabia què significava volar. Amb la IA, és el contrari, els diferents experts tenen intuïcions totalment diferents de què hi ha al cor d'aquesta intel·ligència. És el fet de resoldre problemes? Implica d'alguna manera l'aprenentatge i l'adaptació, les emocions o tenir un cos físic? El fet és que no ho sabem. Però les respostes diverses condueixen a expectatives radicalment diferents d'on va la tecnologia i amb quina celeritat hi arribarà.

Un gran repte a l'hora de construir la IA és que ningú es posa d'acord pel que fa a què significa exactament ser intel·ligent

Un exemple de com n'estem de confosos és de com parlàvem abans de la IA estreta en oposició a la general. Durant molt de temps, parlàvem en termes de dues sitges. Molts pensaven que simplement havíem de dividir-les entre una IA estreta, entrenada per a una tasca específica, com per exemple recomanar el següent vídeo de YouTube, a diferència de la Intel·ligència General Artificial, o AGI [sigles en anglès], que sabria fer tot el que sap fer una persona humana. Pensàvem en aquesta diferenciació, entre estreta i general, com una divisió nuclear entre allò que podríem construir a la pràctica i allò que realment seria intel·ligent.

Però, aleshores, fa uns anys, apareix el ChatGPT. I us podeu preguntar, és una IA estreta, entrenada per a una tasca específica? O és AGI i és capaç de fer tot el que sap fer una persona humana? És evident que la resposta és, cap de les dues. Sens dubte, és de funció general. Sap codificar, escriure poesia, analitzar problemes empresarials, ajudar-te a arreglar el cotxe. Però està molt lluny de poder fer-ho tot tan bé com ho sabríem fer tu o jo. Llavors, resulta que aquesta idea de generalitat no sembla ser la línia divisòria correcta entre intel·ligent i no-intel·ligent. I aquesta mena de cosa és, ara mateix, el repte enorme per a tot el camp de la IA. No ens posem d'acord del que estem intentant construir, ni com ha de ser el full de ruta a partir d'aquí. Ni tan sols entenem clarament els sistemes d'IA que tenim avui.

I com és això? Els investigadors de vegades descriuen les xarxes neuronals profundes, la mena d'IA que principalment s'està bastint avui, com una caixa negra. Però el que volen dir amb això no és que sigui intrínsecament misteriós i que no tenim manera de veure què hi ha a dins de la caixa. El problema és que quan hi mirem, el que hi trobem són milions, milers de milions o, fins i tot, bilions de xifres que se sumen i es multipliquen d'una manera específica. El que fa difícil que els experts sàpiguen el que està passant, és bàsicament que hi ha massa números i encara no tenim cap manera de treure l'entrellat del que estan fent. Hi ha algun detall més, però no gaire més.

Així que, com hem de regular aquesta tecnologia que ens costa tant d'entendre i predir? Us plantejaré dues idees. Una per a tots nosaltres, i una per als responsables polítics.

Primer, no us deixeu intimidar, sigui per la mateixa tecnologia o per les persones i les empreses que la basteixen. Pel que fa a la tecnologia, la IA pot ser confusa, però no és màgia. Hi ha algunes parts dels sistemes d'IA que sí que entenem bé, i fins i tot les parts que no entenem no seran opaques per sempre. Una àrea de recerca que es coneix com a «interpretabilitat de la IA» ha avançat bastant en els últims anys a l'hora de treure l'entrellat de què fan tots aquests milers de milions de números. Un equip d'investigadors, per exemple, ha trobat la manera d'identificar diferents parts d'una xarxa neuronal que poden graduar a més o a menys per fer que les respostes de la IA siguin de més felicitat o d'enuig, més honestes, més maquiavèl·liques... Si podem tirar endavant aquesta mena de recerca, potser d'aquí a cinc o deu anys tindrem una comprensió molt més clara del que passa dins de l'anomenada caixa negra.

I pel que fa a aquells que basteixen la tecnologia, els tecnòlegs, de vegades

es comporten com si no tinguessis cap dret a una opinió sobre què n'hauríem de fer si no estàs ficat en tots els detalls tècnics. L'expertesa té el seu lloc, és clar, però la història ens mostra com és d'important que les persones afectades per una nova tecnologia tinguin un paper en la configuració de com la fem servir. Com les persones treballadores de les fàbriques del segle XX, que van lluitar per la seguretat laboral, o els defensors de les persones discapacitades que s'empenyien que la Internet havia de ser accessible. No cal ser científic o enginyer per tenir veu.

En segon lloc, hem de centrar-nos en l'adaptabilitat, no en la certesa. Moltes converses sobre com fer polítiques d'IA s'empantaneguen en discussions entre, d'una banda, qui diu: «*Hem de regular la IA molt rígidament ara mateix perquè és tan arriscada*» i, d'altra banda, qui diu: «*Però la regulació matarà la innovació, i de totes maneres, aquests riscos són imaginaris*». Però, tal com ho veig jo, no és només una opció entre frenar en sec o trepitjar l'accelerador. Si estàs conduint per una carretera amb girs i revolts inesperats, dues coses t'ajudaran molt, i són tenir la visió clara pel parabrisa i un sistema de direcció excel·lent. En IA, això significa tenir una imatge clara d'on es troba la tecnologia i cap a on va, i tenir plans establerts de què fer en diferents escenaris. Concretament, això significa coses com invertir en la nostra capacitat de mesurar què poden fer els sistemes d'IA. Això pot sonar a «friqui», però és realment important. Ara mateix, si volem esbrinar si una IA pot fer alguna cosa preocupant, com ara hackejar una infraestructura crítica o persuadir algú perquè canviï les seves creences polítiques, els nostres mètodes de mesura són rudimentaris. En necessitem de millors. També hauríem d'exigir a les empreses d'IA, especialment a les empreses que basteixen els sistemes d'IA més avançats, que comparteixin informació sobre què estan construint, què poden fer els seus sistemes i com gestionen els riscos. I haurien de deixar entrar auditors d'IA externs perquè examinin la seva feina, perquè no siguin les mateixes empreses que la peritin. Un últim exemple de com pot resultar tot això és establint mecanismes d'informació d'incidències, de manera que quan les coses surtin malament al món real, tinguem una manera de recopilar dades sobre què ha passat i com podem solucionar-ho la pròxima vegada. Com fem amb les dades que recollim sobre els accidents d'avió i els ciberatacs. Cap d'aquestes idees és meva, i algunes ja s'estan començant a implementar a llocs com Brussel·les, Londres i, fins i tot, Washington. Però la raó per la qual destaco aquestes idees, el mesurament, la divulgació, els informes d'incidents, és perquè ens ajuden a conduir el progrés de la IA, donant-nos una visió més clara pel parabrisa. Si la IA progressa ràpidament en direccions perilloses, aquestes polítiques ens ajudaran a veure-ho. I si tot va bé, també ens ho mostraran, i podrem respondre conseqüentment.

El que vull que tingueu en ment és que és cert que hi ha un munt d'incerteses i desacords en el camp de la IA. I que en tot cas, les empreses ja estan bastint i desplegant la IA per tot arreu, de maneres que ens afecten a tots. Deixant-ho tot en les seves mans, sembla que les empreses d'IA seguirien un sender similar al de les empreses de xarxes socials, gastant la majoria dels seus recursos

El que podem fer és establir les polítiques que ens donin una visió, la més clara possible, de com s'està canviant la tecnologia

en la creació d'aplicacions web i en [guanyar] l'atenció dels usuaris. I, per defecte, sembla que l'enorme poder dels sistemes d'IA més avançats es podria concentrar en mans d'un petit nombre d'empreses o, fins i tot, d'un petit nombre d'individus. Però el potencial de la IA va molt més enllà. La IA ja ens permet saltar les barreres entre llengües i predir les estructures de proteïnes. Els sistemes més avançats podrien deslliurar l'energia de fusió, neta i il·limitada, o revolucionar la manera en què cultivem els aliments, i mil coses més. I cadascú de nosaltres té veu en el que passa. No som només fonts de dades, som usuaris, som treballadors, som ciutadans.

O sigui que, per temptador que sigui, no podem estar-nos esperant que hi hagi claredat o consens dels experts per esbrinar què volem que passi amb la IA. La IA ja ens està passant. El que podem fer és establir les polítiques que ens donin una visió, la més clara possible, de com s'està canviant la tecnologia, i aleshores podrem entrar a l'arena i promoure els futurs que realment volem.

Gràcies. ■