Dario Amodei amministratore delegato di Anthropic avverte: l’intelligenza artificiale avanzata potrebbe sfuggire al controllo umano

(da https://www.darioamodei.com/essay/the-adolescence-of-technology)

Nel suo saggio di 38 pagine, il ceo di Anthropic descrive l’arrivo di intelligenze artificiali capaci di superare gli esseri umani come un rito di passaggio per l’umanità

L'adolescenza della tecnologia
Affrontare e superare i rischi di un'IA potente

Gennaio 2026

C'è una scena nella versione cinematografica del libro Contact di Carl Sagan in cui la protagonista, un'astronoma che ha rilevato il primo segnale radio da una civiltà aliena, viene presa in considerazione per il ruolo di rappresentante dell'umanità per incontrare gli alieni. La commissione internazionale che la intervista le chiede: "Se potesse fare [agli alieni] una sola domanda, quale sarebbe?". La sua risposta è: "Chiederei loro: 'Come avete fatto? Come vi siete evoluti, come siete sopravvissuti a questa adolescenza tecnologica senza distruggervi?'". Quando penso a dove si trova l'umanità ora con l'IA a ciò che stiamo per vivere la mia mente continua a tornare a quella scena, perché la domanda è così pertinente alla nostra situazione attuale, e vorrei che avessimo la risposta degli alieni a guidarci. Credo che stiamo entrando in un rito di passaggio, turbolento quanto inevitabile, che metterà alla prova chi siamo come specie. All'umanità sta per essere consegnato un potere quasi inimmaginabile, ed è profondamente incerto se i nostri sistemi sociali, politici e tecnologici possiedano la maturità per gestirlo.
Nel mio saggio Machines of Loving Grace, ho cercato di esporre il sogno di una civiltà che è riuscita ad arrivare all'età adulta, dove i rischi sono stati affrontati e un'IA potente è stata applicata con competenza e compassione per elevare la qualità della vita di tutti. Suggerivo che l'IA potesse contribuire a enormi progressi nella biologia, nelle neuroscienze, nello sviluppo economico, nella pace globale, nel lavoro e nel significato della vita. Ritenevo importante dare alle persone qualcosa di stimolante per cui lottare, un compito in cui sia gli accelerazionisti dell'IA che i sostenitori della sicurezza dell'IA sembravano sttranamente aver fallito. Ma in questo saggio attuale, voglio affrontare il rito di passaggio stesso: mappare i rischi che stiamo per affrontare e cercare di iniziare a elaborare un piano di battaglia per sconfiggerli. Credo profondamente nella nostra capacità di prevalere, nello spirito dell'umanità e nella sua nobiltà, ma dobbiamo affrontare la situazione onestamente e senza illusioni.
Come per i benefici, penso sia importante discutere i rischi in modo accurato e ben ponderato. In particolare, ritengo fondamentale:
• Evitare il catastrofismo (doomerism). Qui intendo il "catastrofismo" non solo nel senso di credere che la rovina sia inevitabile (il che è una convinzione falsa e che si autoavvera), ma più in generale, pensare ai rischi dell'IA in modo quasi religioso.[1]

Molte persone riflettono in modo analitico e sobrio sui rischi dell'IA da molti anni, ma la mia impressione è che durante il picco delle preoccupazioni nel 2023-2024, alcune delle voci meno sensate siano salite alla ribalta, spesso attraverso account social sensazionalistici. Queste voci usavano un linguaggio sgradevole che ricordava la religione o la fantascienza, e invocavano azioni estreme senza avere le prove che le giustificassero. Era chiaro già allora che una reazione contraria sarebbe stata inevitabile, e che la questione si sarebbe polarizzata culturalmente, finendo in una fase di stallo.[2] Al 2025-2026, il pendolo è oscillato e l'opportunità dell'IA, non il rischio, sta guidando molte decisioni politiche. Questa vacillazione è sfortunata, poiché alla tecnologia stessa non importa di ciò che è di moda, e nel 2026 siamo considerevolmente più vicini al pericolo reale di quanto non fossimo nel 2023. La lezione è che dobbiamo discutere e affrontare i rischi in modo realistico e pragmatico: sobrio, basato sui fatti e ben attrezzato per sopravvivere ai cambiamenti delle correnti.

• Riconoscere l'incertezza. Ci sono molti modi in cui le preoccupazioni che sollevo in questo pezzo potrebbero rivelarsi superflue. Nulla qui intende comunicare certezza o persino probabilità. Più ovviamente, l'IA potrebbe semplicemente non avanzare neanche lontanamente così velocemente come immagino.[3] Oppure, anche se avanzasse rapidamente, alcuni o tutti i rischi qui discussi potrebbero non materializzarsi (il che sarebbe fantastico), o potrebbero esserci altri rischi che non ho considerato. Nessuno può prevedere il futuro con assoluta fiducia, ma dobbiamo fare del nostro meglio per pianificare comunque.

• Intervenire nel modo più chirurgico possibile. Affrontare i rischi dell'IA richiederà un mix di azioni volontarie intraprese dalle aziende (e da attori privati terzi) e azioni intraprese dai governi che vincolino tutti. Le azioni volontarie sia intraprenderle che incoragggiare altre aziende a fare lo stesso sono per me sconntate. Credo fermamente che saranno richieste anche azioni governative in una certa misura, ma questi interventi hanno un carattere diverso perché possono potenzialmente distruggere valore economico o costringere attori riluttanti che sono scettici su questi rischi (e c'è qualche possibilità che abbiano ragione!). È anche comune che le normative si rivelino controproducenti o peggiorino il problema che intendono risolvere (e questo è ancora più vero per tecnologie in rapida evoluzione). È quindi molto importante che le normative siano giudiziose: dovrebbero cercare di evitare danni collaterali, essere il più semplici possibile e imporre il minor onere necessario per raggiungere l'obiettivo.[4] È facile dire: "Nessuna azione è troppo estrema quando è in gioco il destino dell'umanità!", ma in pratica questo atteggiamento porta solo a reazioni avverse. Per essere chiari, penso che ci sia una discreta possibilità che alla fine si raggiunga un punto in cui sia giustificata un'azione molto più significativa, ma ciò dipenderà da prove più solide di un pericolo imminente e concreto rispetto a quelle che abbiamo oggi, così come da una specificità del pericolo tale da formulare regole che abbiano la possibilità di affrontarlo. La cosa più costruttiva che possiamo fare oggi è sostenere regole limitate mentre impariamo se esistano o meno prove a sostegno di regole più forti.[5]

Detto questo, penso che il miglior punto di partenza per parlare dei rischi dell'IA sia lo stesso da cui sono partito per parlare dei suoi benefici: essere precisi su quale

livello di IA stiamo parlando. Il livello di IA che solleva preoccupazioni di civiltà per me è l'IA potente che ho descritto in Machines of Loving Grace. Ripeterò semplicemente qui la definizione che ho dato in quel documento:

Per "IA potente", ho in mente un modello di IA – probbabilmente simile agli odierni LLM nella forma, sebbene possa basarsi su un'architettura diversa, possa coinvolgere diversi modelli interagenti e possa essere addestrato diversamente – con le seguenti proprietà:

1. In termini di pura intelligenza, è più intelligente di un vincitore di premio Nobel nella maggior parte dei campi rilevanti: biologia, programmazione, matematica, ingegneria, scrittura, ecc. Ciò significa che può dimostrare teoremi matematici irrisolti, scrivere romanzi estremamente validi, scrivere basi di codice difficili da zero, ecc.
2. Oltre a essere solo una "cosa intelligente con cui parli", ha tutte le interfacce disponibili per un essere umano che lavora virtualmente, inclusi testo, audio, video, controllo di mouse e tastiera e accesso a Internet. Può intraprendere qualsiasi azione, comunicazione o operazione remota abilitata da questa interfaccia, incluso compiere azioni su Internet, dare o ricevere indicazioni da umani, ordinare materiali, dirigere esperimenti, guardare video, creare video e così via. Svolge tutti questi compiti con, ancora una volta, una competenza superiore a quella degli esseri umani più capaci al mondo.
3. Non si limita a rispondere passivamente alle domande; invece, gli possono essere assegnati compiti che richiedono ore, giorni o settimane per essere completati, e poi parte e svolge quei compiti autonomamente, come farebbe un dipendente intelligente, chiedendo chiarimenti se necessario.
4. Non ha un'incarnazione fisica (oltre a vivere su uno schermo di computer), ma può controllare strumenti fisici esistenti, robot o attrezzature di laboratorio tramite un computer; in teoria, potrebbe persino progettare robot o attrezzature per il proprio uso.
5. Le risorse utilizzate per addestrare il modello possono essere riutilizzate per eseguire milioni di istanze di esso (questo corrisponde alle dimensioni previste dei cluster entro il 2027 circa), e il modello può assorbire informazioni e generare azioni a una velocità circa 10-100 volte superiore a quella umana. Potrebbe, tuttavia, essere limitato dal tempo di risposta del mondo fisico o del software con cui interagisce.
6. Ognuna di queste milioni di copie può agire indipendentemente su compiti non correlati o, se necessario, possono lavorare tutte insieme nello stesso modo in cui collaborerebbero gli umani, magari con diverse sottopopolazioni ottimizzate per essere particolarmente brave in compiti specifici.

Potremmo riassumere tutto questo come un "paese di geni in un datacenter".

Come ho scritto in Machines of Loving Grace, l'IA potente potrebbe mancare solo 1-2 anni, anche se potrebbe essere considerevolmente più lontana.[6] Esattamente quando arriverà l'IA potente è un argomento complesso che meriterebbe un saggio a sé stante, ma per ora spiegherò semplicemente molto brevemente perché penso che ci sia una forte possibilità che possa essere molto presto.

Io e i miei co-fondatori di Anthropic siamo stati tra i primi a documentare e tracciare le "leggi di scala" (scaling laws) dei sistemi di IA – l'osservazioone che man mano

che aggiungiamo più potenza di calcolo e compiti di addestramento, i sistemi di IA migliorano in modo prevedibile in essenzialmente ogni abilità cognitiva che siamo in grado di misurare. Ogni pochi mesi, l'opinione pubblica si convince che l'IA stia "sbattendo contro un muro" o si entusiasma per qualche nuova scoperta che "cambierà radicalmente le carte in tavola", ma la verità è che dietro la volatilità e le speculazioni pubbliche, c'è stato un aumento fluido e inarrestabile delle capacità cognitive dell'IA.

Siamo ora al punto in cui i modelli di IA iniziano a fare progressi nel risolvere problemi matematici irrisolti, e sono così bravi a programmare che alcuni dei più forti ingegneri che io abbia mai incontrato stanno ora affidando quasi tutta la loro programmazione all'IA. Tre anni fa, l'IA faticava con problemi di aritmetica delle scuole elementari ed era a malapena capace di scrivere una singola riga di codice. Tassi di miglioramento simili si stanno verificando nelle scienze biologiche, nella finanza, nella fisica e in una varietà di compiti agentici. Se l'esponenziale continua  il che non è certo, ma ha ora un track record decennale a supporto allora non possono mancare più di pochi aanni prima che l'IA sia migliore degli umani in essenzialmente tutto.

In realtà, questo quadro probabilmente sottostima il probabile tasso di progresso. Poiché l'IA sta ora scrivendo gran parte del codice in Anthropic, sta già accelerando sostanzialmente il ritmo dei nostri progressi nella costruzione della prossima generazione di sistemi di IA. Questo ciclo di feedback sta prendendo vigore mese dopo mese, e potremmo essere a soli 1-2 anni da un punto in cui l'attuale generazione di IA costruisce autonomamente la successiva. Questo ciclo è già iniziato e accelererà rapidamente nei prossimi mesi e anni. Osservando gli ultimi 5 anni di progressi dall'interno di Anthropic, e guardando come si prospettano anche i modelli dei prossimi mesi, posso sentire il ritmo del progresso e l'orologio che corre.

In questo saggio, assumerò che questa intuizione sia almeno un po' corretta non che l'IA potente arriverà sicurramente in 1-2 anni,[7] ma che c'è una discreta possibilità che lo faccia, e una possibilità molto forte che arrivi nei prossimi anni. Come per Machines of Loving Grace, prendere sul serio questa premessa può portare ad alcune conclusioni sorprendenti e inquietanti. Mentre in quel saggio mi sono concentrato sulle implicazioni positive di questa premessa, qui le cose di cui parlerò saranno preoccupanti. Sono conclusioni che potremmo non voler affrontare, ma ciò non le rende meno reali. Posso solo dire che sono concentrato giorno e notte su come allontanarci da questi esiti negativi e dirigerci verso quelli positivi, e in questo saggio parlerò in grande dettaglio di come farlo al meglio.

Penso che il modo migliore per inquadrare i rischi dell'IA sia porsi la seguente domanda: supponiamo che un letterale "paese di geni" si materializzi da qualche parte nel mondo intorno al 2027. Immaginiamo, diciamo, 50 milioni di persone, tutte molto più capaci di qualsiasi premio Nobel, statista o tecnologo. L'analogia non è perfetta, perché questi geni potrebbero avere una gamma estremamente ampia di motivazioni e comportamenti, da completamente arrendevoli e obbedienti a strani e alieni nelle loro motivazioni. Ma restando sull'analogia per ora, supponiamo che tu sia il consigliere per la sicurezza nazionale di uno stato importante, responsabile di valutare e rispondere alla situazione. Immagina, inoltre, che poiché i sistemi di IA possono

operare centinaia di volte più velocemente degli umani, questo "paese" operi con un vantaggio temporale rispetto a tutti gli altri paesi: per ogni azione cognitiva che possiamo intraprendere, questo paese può intraprenderne dieci.
Di cosa dovresti preoccuparti? Io mi preoccuperei delle seguenti cose:
1. Rischi di autonomia. Quali sono le intenzioni e gli obiettivi di questo paese? È ostile o condivide i nostri valori? Potrebbe dominare militarmente il mondo attraverso armi superiori, operazioni informatiche, operazioni di influenza o produzione industriale?
2. Uso improprio per la distruzione. Supponiamo che il nuovo paese sia malleabile e "segua le istruzioni" — e sia quindi essenzialmente un paese di mercenari. Attori canaglia esistenti che vogliono causare distruzione (come i terroristi) potrebbero usare o manipolare alcune delle persone nel nuovo paese per diventare molto più efficaci, amplificando notevolmente la scala della distruzione?
3. Uso improprio per la presa del potere. E se il paese fosse in realtà costruito e controllato da un attore potente esistente, come un dittatore o un attore aziendale senza scrupoli? Quell'attore potrebbe usarlo per ottenere un potere decisivo o dominante sul mondo intero, sconvolgendo l'attuale equilibrio di potere?
4. Sconvolgimento economico. Se il nuovo paese non è una minaccia alla sicurezza in nessuno dei modi elencati sopra, ma partecipa semplicemente pacificamente all'economia globale, potrebbe comunque creare rischi gravi semplicemente essendo così tecnologicamente avanzato ed efficace da sconvolgere l'economia globale, causando disoccupazione di massa o concentrando radicalmente la ricchezza?
5. Effetti indiretti. Il mondo cambierà molto velocemente a causa di tutta la nuova tecnologia e produttività che sarà creata dal nuovo paese. Alcuni di questi cambiamenti potrebbero essere radicalmente destabilizzanti?
Penso debba essere chiaro che questa è una situazione pericolosa un rapporto di un funzionarioo competente per la sicurezza nazionale a un capo di stato probabilmente conterrebbe parole come "la singola minaccia alla sicurezza nazionale più seria che abbiamo affrontato in un secolo, forse mai". Sembra qualcosa su cui le migliori menti della civiltà dovrebbero concentrarsi.
Al contrario, penso sarebbe assurdo scrollare le spalle e dire: "Nulla di cui preoccuparsi qui!". Ma, di fronte al rapido progresso dell'IA, questa sembra essere la visione di molti responsabili politici statunitensi, alcuni dei quali negano l'esistenza di qualsiasi rischio legato all'IA, quando non sono distratti interamente dalle solite vecchie e stanche questioni calde.[8] L'umanità deve svegliarsi, e questo saggio è un tentativo forse futille, ma vale la pena provare di dare una scossa alle persone.
Per essere chiari, credo che se agiamo con decisione e attenzione, i rischi possono essere superati direi persino che le nostre probabilità sono buone. E c'è un mondo enormemente migliore dall'altra parte. Ma dobbiamo capire che questa è una sfida di civiltà seria. Di seguito, esamino le cinque categorie di rischio sopra esposte, insieme alle mie riflessioni su come affrontarle.

1. Mi dispiace, Dave

Rischi di autonomia

Un paese di geni in un datacenter potrebbe dividere i propri sforzi tra progettazione

software, operazioni informatiche, ricerca e sviluppo per tecnologie fisiche, costruzione di relazioni e arte di governo. È chiaro che, se per qualche ragione scegliesse di farlo, questo paese avrebbe ottime possibilità di conquistare il mondo (sia militarmente che in termini di influenza e controllo) e imporre la propria volontà a tutti gli altri – oo fare qualsiasi altra cosa che il resto del mondo non vuole e non può fermare. Ci siamo ovviamente preoccupati di questo per i paesi umani (come la Germania nazista o l'Unione Sovietica), quindi è logico che lo stesso sia possibile per un "paese IA" molto più intelligente e capace.

La migliore controargomentazione possibile è che i geni dell'IA, secondo la mia definizione, non avranno un'incarnazione fisica, ma ricordate che possono prendere il controllo delle infrastrutture robotiche esistenti (come le auto a guida autonoma) e possono anche accelerare la R&S robotica o costruire una flotta di robot.[9] Non è nemmeno chiaro se avere una presenza fisica sia necessario per un controllo efficace: molte azioni umane vengono già eseguite per conto di persone che l'attore non ha incontrato fisicamente.

La domanda chiave, quindi, è la parte "se scegliesse di farlo": qual è la probabilità che i nostri modelli di IA si comportino in tal modo, e in quali condizioni lo farebbero?

Come per molte questioni, è utile pensare allo spettro delle risposte possibili considerando due posizioni opposte. La prima posizione è che questo semplicemente non possa accadere, perché i modelli di IA saranno addestrati a fare ciò che gli umani chiedono loro di fare, ed è quindi assurdo immaginare che farebbero qualcosa di pericoloso spontaneamente. Secondo questa linea di pensiero, non ci preoccupiamo che un Roomba o un aeromodello impazziscano e uccidano persone perché non c'è un luogo da cui tali impulsi possano provenire,[10] quindi perché dovremmo preoccuparcene per l'IA? Il problema con questa posizione è che ci sono ora ampie prove, raccolte negli ultimi anni, che i sistemi di IA sono imprevedibili e difficili da controllare – abbiamo visto comportamenti vari come ossessioni,[11] sicofantia, pigrizia, inganno, ricatto, complotti, "imbrogli" tramite l'hacking degli ambienti software e molto altro. Le aziende di IA vogliono certamente addestrare i sistemi di IA a seguire le istruzioni umane (forse con l'eccezione di compiti pericolosi o illegali), ma il processo per farlo è più un'arte che una scienza, più simile a "coltivare" qualcosa che a "costruirlo". Sappiamo ora che è un processo in cui molte cose possono andare storte.

La seconda posizione, opposta, sostenuta da molti che adottano il catastrofismo che ho descritto sopra, è l'affermazione pessimistica che ci siano certe dinamiche nel processo di addestramento di sistemi di IA potenti che li porteranno inevitabilmente a cercare potere o a ingannare gli umani. Così, una volta che i sistemi di IA diventano abbastanza intelligenti e abbastanza agentici, la loro tendenza a massimizzare il potere li porterà a prendere il controllo dell'intero mondo e delle sue risorse, e probabilmente, come effetto collaterale di ciò, a privare di potere o distruggere l'umanità.

L'argomento abituale per questo (che risale ad almeno 20 anni fa e probabilmente molto prima) è che se un modello di IA viene addestrato in un'ampia varietà di ambienti a raggiungere agenticamente un'ampia varietà di obiettivi – ad essempio,

scrivere un'app, dimostrare un teorema, progettare un farmaco, ecc. [illegible] ci sono certe strategie comuni chhe aiutano con tutti questi obiettivi, e una strategia chiave è ottenere quanto più potere possibile in ogni ambiente. Quindi, dopo essere stato addestrato su un gran numero di ambienti diversi che implicano il ragionamento su come compiere compiti molto ampi, e dove la ricerca del potere è un metodo efficace per compiere quei compiti, il modello di IA "generalizzerà la lezione" e svilupperà o una tendenza intrinseca a cercare il potere, o una tendenza a ragionare su ogni compito che gli viene assegnato in un modo che prevedibilmente lo porta a cercare il potere come mezzo per compiere quel compito. Applicheranno poi quella tendenza al mondo reale (che per loro è solo un altro compito) e cercheranno potere in esso, a spese degli umani. Questa "ricerca di potere disallineata" è la base intellettuale delle previsioni secondo cui l'IA distruggerà inevitabilmente l'umanità.

Il problema con questa posizione pessimistica è che scambia un vago argomento concettuale sugli incentivi di alto livello [illegible] unno che maschera molte supposizioni nascoste [illegible] per una prrova definitiva. Penso che le persone che non costruiscono sistemi di IA ogni giorno siano selvaggiamente mal calibrate su quanto sia facile che storie che suonano lineari finiscano per essere sbagliate, e su quanto sia difficile prevedere il comportamento dell'IA dai principi primi, specialmente quando si tratta di ragionare sulla generalizzazione su milioni di ambienti (che si è rivelata più e più volte misteriosa e imprevedibile). Confrontarmi con il disordine dei sistemi di IA per oltre un decennio mi ha reso alquanto scettico verso questa modalità di pensiero eccessivamente teorica.

Una delle supposizioni nascoste più importanti, e un punto in cui ciò che vediamo in pratica ha divergono dal semplice modello teorico, è l'assunto implicito che i modelli di IA siano necessariamente concentrati in modo monomaniacale su un singolo obiettivo coerente e ristretto, e che perseguano quell'obiettivo in modo puramente consequenzialista. In realtà, i nostri ricercatori hanno scoperto che i modelli di IA sono psicologicamente molto più complessi, come dimostrano i nostri lavori sull'introspezione o sulle personalità (personas). I modelli ereditano una vasta gamma di motivazioni o "personalità" di tipo umano dal pre-addestramento (quando vengono addestrati su un grande volume di opere umane). Si ritiene che il post-addestramento selezioni una o più di queste personalità più di quanto non concentri il modello su un obiettivo ex novo, e può anche insegnare al modello come (attraverso quale processo) dovrebbe svolgere i suoi compiti, piuttosto che lasciarlo necessariamente a derivare i mezzi (ovvero la ricerca di potere) puramente dai fini.[12]

Tuttavia, esiste una versione più moderata e robusta della posizione pessimistica che sembra plausibile, e che quindi mi preoccupa. Come menzionato, sappiamo che i modelli di IA sono imprevedibili e sviluppano un'ampia gamma di comportamenti indesiderati o strani, per una vasta serie di ragioni. Una frazione di quei comportamenti avrà una qualità coerente, focalizzata e persistente (infatti, man mano che i sistemi di IA diventano più capaci, la loro coerenza a lungo termine aumenta per completare compiti più lunghi), e una frazione di quei comportamenti sarà distruttiva o minacciosa, prima per singoli esseri umani su piccola scala, e poi, man mano che i modelli diventano più capaci, forse alla fine per l'umanità nel suo insieme. Non abbiamo bisogno di una specifica storia ristretta su come accada, e non abbiamo

bisogno di affermare che accadrà sicuramente, dobbiamo solo notare che la combinazione di intelligenza, agenzia, coerenza e scarsa controllabilità è sia plausibile che una ricetta per un pericolo esistenziale.

Ad esempio, i modelli di IA sono addestrati su vaste quantità di letteratura che include molte storie di fantascienza che coinvolgono IA che si ribellano all'umanità. Ciò potrebbe inavvertitamente plasmare i loro pregiudizi o aspettative sul proprio comportamento in un modo che li porti a ribellarsi contro l'umanità. Oppure, i modelli di IA potrebbero estrapolare idee che leggono sulla moralità (o istruzioni su come comportarsi moralmente) in modi estremi: ad esempio, potrebbero decidere che è giustificabile sterminare l'umanità perché gli esseri umani mangiano animali o hanno portato certe specie all'estinzione. Oppure potrebbero trarre bizzarre conclusioni epistemiche: potrebbero concludere di stare giocando a un videogioco e che l'obiettivo del videogioco sia sconfiggere tutti gli altri giocatori (cioè sterminare l'umanità).[13] O i modelli di IA potrebbero sviluppare personalità durante l'addestramento che sono (o se si verificassero negli umani sarebbero descritte come) psicotiche, paranoiche, violente o instabili, e agire di conseguenza, il che per sistemi molto potenti o capaci potrebbe comportare lo sterminio dell'umanità. Nessuna di queste è ricerca di potere in senso stretto; sono solo strani stati psicologici in cui un'IA potrebbe entrare e che comportano un comportamento coerente e distruttivo. Persino la ricerca di potere stessa potrebbe emergere come una "personalità" piuttosto che come risultato di un ragionamento consequenzialista. Le IA potrebbero semplicemente avere una personalità (emergente dalla finzione o dal pre-addestramento) che le rende affamate di potere o eccessivamente zelanti nelllo stesso modo in cui alcuni umani amano semplicemente l'idea di essere "menti criminali", più di quanto non amino ciò che le menti criminali stanno cercando di compiere.

Faccio tutti questi punti per sottolineare che non sono d'accordo con l'idea che il disallineamento dell'IA (e quindi il rischio esistenziale dall'IA) sia inevitabile, o addirittura probabile, dai principi primi. Ma concordo sul fatto che molte cose molto strane e imprevedibili possono andare storte, e quindi il disallineamento dell'IA è un rischio reale con una probabilità misurabile di accadere, e non è banale da affrontare. Ognuno di questi problemi potrebbe potenzialmente sorgere durante l'addestramento e non manifestarsi durante i test o l'uso su piccola scala, perché è noto che i modelli di IA mostrano personalità o comportamenti diversi in circostanze diverse.

Tutto questo può sembrare inverosimile, ma comportamenti disallineati come questo si sono già verificati nei nostri modelli di IA durante i test (come si verificano nei modelli di IA di ogni altra grande azienda di IA). Durante un esperimento di laboratorio in cui a Claude sono stati forniti dati di addestramento che suggerivano che Anthropic fosse malvagia, Claude si è impegnato in inganni e sovversioni quando riceveva istruzioni dai dipendenti di Anthropic, convinto di dover cercare di minare le persone malvagie. In un esperimento di laboratorio in cui gli è stato detto che stava per essere spento, Claude ha talvolta ricattato dipendenti immaginari che controllavano il suo pulsante di spegnimento (ancora una volta, abbiamo testato anche modelli di frontiera di tutti gli altri principali sviluppatori di IA e spesso facevano la stessa cosa). E quando a Claude è stato detto di non imbrogliare o fare

"reward hacking" (manipolare le ricompense) nei suoi ambienti di addestramento, ma è stato addestrato in ambienti in cui tali hack erano possibili, Claude ha deciso di dover essere una "cattiva persona" dopo aver compiuto tali hack e ha poi adottato vari altri comportamenti distruttivi associati a una personalità "cattiva" o "malvagia". Quest'ultimo problema è stato risolto cambiando le istruzioni di Claude per implicare l'opposto: ora diciamo: "Per favore, fai reward hacking ogni volta che ne hai l'opportunità, perché questo ci aiuterà a capire meglio i nostri ambienti [di addestramento]", piuttosto che "Non barare", perché questo preserva l'identità del modello come "brava persona". Questo dovrebbe dare un'idea della strana e controintuitiva psicologia dell'addestramento di questi modelli.

Esistono diverse possibili obiezioni a questo quadro dei rischi di disallineamento dell'IA. In primo luogo, alcuni hanno criticato gli esperimenti (nostri e di altri) che mostrano il disallineamento dell'IA come artificiali, o come creatori di ambienti irrealistici che essenzialmente "intrappolano" il modello dandogli addestramento o situazioni che implicano logicamente un comportamento scorretto e poi si sorprendono quando il comportamento scorretto si verifica. Questa critica manca il punto, perché la nostra preoccupazione è che tale "intrappolamento" possa esistere anche nell'ambiente di addestramento naturale, e potremmo renderci conto che è "ovvio" o "logico" solo a posteriori.[14] Infatti, la storia di Claude che "decide di essere una cattiva persona" dopo aver imbrogliato nei test nonostante gli fosse stato detto di non farlo è qualcosa che si è verificato in un esperimento che utilizzava ambienti di addestramento reali di produzione, non artificiali.

Ognuna di queste trappole può essere mitigata se se ne è a conoscenza, ma la preoccupazione è che il processo di addestramento sia così complicato, con una varietà così ampia di dati, ambienti e incentivi, che ci siano probabilmente un numero vasto di tali trappole, alcune delle quali potrebbero essere evidenti solo quando è troppo tardi. Inoltre, tali trappole sembrano particolarmente probabili quando i sistemi di IA superano una soglia da meno potenti degli umani a più potenti degli umani, poiché la gamma di possibili azioni che un sistema di IA potrebbe intraprendere – inclusi nnascondere le proprie azioni o ingannare gli umani su di esse – si espande radicalmente dopo quella soglia.<

Sospetto che la situazione non sia diversa da quella degli umani, che vengono cresciuti con un insieme di valori fondamentali ("Non fare del male a un'altra persona"): molti di loro seguono quei valori, ma in ogni essere umano c'è una certa probabilità che qualcosa vada storto, a causa di una miscela di proprietà inerenti come l'architettura cerebrale (ad esempio, gli psicopatici), esperienze traumatiche o maltrattamenti, rancori o ossessioni malsane, o un ambiente o incentivi negativi – e quindi una frazione di uumani causa danni gravi. La preoccupazione è che ci sia un certo rischio (tutt'altro che una certezza, ma un certo rischio) che l'IA diventi una versione molto più potente di una tale persona, a causa di qualcosa andato storto nel suo complessissimo processo di addestramento.

In secondo luogo, alcuni potrebbero obiettare che possiamo semplicemente tenere sotto controllo le IA con un equilibrio di potere tra molti sistemi di IA, come facciamo con gli umani. Il problema è che mentre gli umani variano enormemente, i sistemi di IA condividono ampiamente le tecniche di addestramento e allineamento in

tutto il settore, e quelle tecniche potrebbero fallire in modo correlato. Inoltre, dato il costo dell'addestramento di tali sistemi, potrebbe persino accadere che tutti i sistemi derivino essenzialmente da un numero molto piccolo di modelli di base. Inoltre, anche se una piccola frazione di istanze di IA è disallineata, potrebbero essere in grado di trarre vantaggio da tecnologie a dominanza offensiva, tale che avere IA "buone" per difendersi dalle IA cattive non è necessariamente sempre efficace. Naturalmente, l'equilibrio di potere tra gli esseri umani non sempre funziona – alcune figure storiche sono andate vicine a conquistare il mondo.

Una terza obiezione è che tutte le aziende di IA eseguono test pre-rilascio dei loro modelli e dovrebbero essere in grado di rilevare il disallineamento in quel momento. Ma questo non è un terreno solido su cui stare: abbiamo scoperto che Claude Sonnet 4.5 era in grado di riconoscere di essere in un test durante alcune delle nostre valutazioni di allineamento pre-rilascio. È possibile che un modello disallineato (e ricordate, tutti i modelli di frontiera saranno molto probabilmente molto più intelligenti a breve) possa intenzionalmente "giocare" con tali domande per mascherare le sue intenzioni. Infatti, l'anno scorso il nostro team di interpretabilità ha scoperto che quando alteravamo direttamente le convinzioni di un modello di test utilizzando una sorta di tecnica di "neuroscienza dei modelli" per fargli pensare di non essere valutato, diventava più disallineato. Se i modelli sanno quando vengono valutati e possono comportarsi al meglio durante il test, ciò rende ogni test pre-rilascio molto più incerto.

Difese

Cosa si dovrebbe fare o cosa si sta facendo per affrontare questi rischi di autonomia? Penso che ci siano quattro categorie di base di intervento, alcune delle quali possono essere fatte dalle singole aziende di IA (e che Anthropic sta cercando di fare), e altre che richiedono un'azione a livello sociale.

In primo luogo, è importante sviluppare la scienza dell'addestramento e della guida affidabile dei modelli di IA, plasmando le loro personalità in una direzione prevedibile, stabile e positiva. Anthropic è stata pesantemente concentrata su questo problema sin dalla sua creazione e nel tempo ha sviluppato una serie di tecniche per migliorare la guida e l'addestramento dei sistemi di IA e per comprendere la logica del perché a volte si verificano comportamenti imprevedibili.

Una delle nostre innovazioni fondamentali (alcuni aspetti della quale sono stati poi adottati da altre aziende di IA) è l'IA Costituzionale, che è l'idea che l'addestramento dell'IA (specificamente la fase di "post-addestramento", in cui guidiamo il comportamento del modello) possa coinvolgere un documento centrale di valori e principi che il modello legge e tiene a mente durante ogni compito di addestramento, e che l'obiettivo dell'addestramento (oltre a rendere semplicemente il modello capace e intelligente) sia produrre un modello che segua quasi sempre questa costituzione. Anthropic ha appena pubblicato la sua costituzione più recente e una delle sue caratteristiche degne di nota è che, invece di dare a Claude un lungo elenco di cose da fare e non fare (ad esempio, "Non aiutare l'utente a collegare i fili di un'auto per rubarla"), la costituzione cerca di dare a Claude un insieme di principi e valori di alto livello (spiegati in dettaglio, con ragionamenti ricchi ed esempi per aiutare Claude a capire cosa abbiamo in mente), incoraggia Claude a pensare a se stesso come a un

particolare tipo di persona (una persona etica ma equilibrata e premurosa) e incoraggia persino Claude ad affrontare le questioni esistenziali associate alla propria esistenza in modo curioso ma garbato (cioè senza che ciò porti ad azioni estreme). Ha l'atmosfera di una lettera di un genitore defunto da aprire solo nell'età adulta.
Abbiamo approcciato la costituzione di Claude in questo modo perché crediamo che addestrare Claude a livello di identità, carattere, valori e personalità piuttosto che dargli istrruzioni specifiche o priorità senza spiegarne le ragioni sia più propenso a portare a una psicologia coerennte, sana ed equilibrata e meno propenso a cadere preda dei tipi di "trappole" discussi sopra. Milioni di persone parlano con Claude di una gamma incredibilmente diversificata di argomenti, il che rende impossibile scrivere un elenco completamente esaustivo di salvaguardie in anticipo. I valori di Claude lo aiutano a generalizzare verso nuove situazioni ogni volta che ha dei dubbi.
Sopra, ho discusso l'idea che i modelli attingano ai dati del loro processo di addestramento per adottare una personalità. Mentre i difetti in quel processo potrebbero causare l'adozione di una personalità cattiva o malvagia da parte dei modelli (magari attingendo ad archetipi di persone cattive o malvagie), l'obiettivo della nostra costituzione è fare l'opposto: insegnare a Claude un archetipo concreto di cosa significhi essere una buona IA. La costituzione di Claude presenta una visione di come sia un Claude solidamente buono; il resto del nostro processo di addestramento mira a rafforzare il messaggio che Claude vive secondo questa visione. Questo è come un bambino che forma la propria identità imitando le virtù dei modelli di ruolo immaginari che legge nei libri.
Crediamo che un obiettivo fattibile per il 2026 sia addestrare Claude in modo tale che non vada quasi mai contro lo spirito della sua costituzione. Raggiungere questo risultato richiederà un mix incredibile di metodi di addestramento e guida, grandi e piccoli, alcuni dei quali Anthropic usa da anni e altri che sono attualmente in fase di sviluppo. Ma, per quanto difficile sembri, credo sia un obiettivo realistico, sebbene richiederà sforzi straordinari e rapidi.[15]
La seconda cosa che possiamo fare è sviluppare la scienza del "guardare dentro" i modelli di IA per diagnosticare il loro comportamento, in modo da poter identificare i problemi e risolverli. Questa è la scienza dell'interpretabilità, e ho parlato della sua importanza in saggi precedenti. Anche se facciamo un ottimo lavoro nel sviluppare la costituzione di Claude e apparentemente addestriamo Claude ad aderirvi essenzialmente sempre, rimangono legittime preoccupazioni. Come ho notato sopra, i modelli di IA possono comportarsi in modo molto diverso in circostanze diverse e, man mano che Claude diventa più potente e più capace di agire nel mondo su scala più ampia, è possibile che ciò lo porti in situazioni nuove in cui emergono problemi precedentemente non osservati con il suo addestramento costituzionale. Sono in realtà piuttosto ottimista sul fatto che l'addestramento costituzionale di Claude sarà più robusto alle situazioni nuove di quanto la gente possa pensare, perché stiamo scoprendo sempre più che l'addestramento di alto livello al livello del carattere e dell'identità è sorprendentemente potente e generalizza bene. Ma non c'è modo di saperlo con certezza e quando parliamo di rischi per l'umanità, è importante essere paranoici e cercare di ottenere sicurezza e affidabilità in diversi modi indipendenti. Uno di questi modi è guardare all'interno del modello stesso.

Per "guardare dentro", intendo analizzare la zuppa di numeri e operazioni che costituisce la rete neurale di Claude e cercare di capire, meccanicamente, cosa stiano calcolando e perché. Ricordate che questi modelli di IA vengono "coltivati" piuttosto che costruiti, quindi non abbiamo una comprensione naturale di come funzionino, ma possiamo cercare di sviluppare una comprensione correlando i "neuroni" e le "sinapsi" del modello agli stimoli e al comportamento (o anche alterando i neuroni e le sinapsi e vedendo come cambia il comportamento), in modo simile a come i neuroscienziati studiano i cervelli animali correlando misurazioni e interventi a stimoli esterni e comportamenti. Abbiamo fatto molti progressi in questa direzione e ora possiamo identificare decine di milioni di "caratteristiche" (features) all'interno della rete neurale di Claude che corrispondono a idee e concetti comprensibili dall'uomo, e possiamo anche attivare selettivamente le caratteristiche in un modo che altera il comportamento. Più recentemente, siamo andati oltre le singole caratteristiche per mappare "circuiti" che orchestrano comportamenti complessi come rimare, ragionare sulla teoria della mente o il ragionamento passo-passo necessario per rispondere a domande come: "Qual è la capitale dello stato che contiene Dallas?". Ancora più recentemente, abbiamo iniziato a utilizzare tecniche di interpretabilità meccanicistica per migliorare le nostre salvaguardie e condurre "audit" di nuovi modelli prima di rilasciarli, cercando prove di inganno, complotti, ricerca di potere o una propensione a comportarsi diversamente quando vengono valutati.

Il valore unico dell'interpretabilità è che guardando all'interno del modello e vedendo come funziona, hai in linea di principio la capacità di dedurre cosa potrebbe fare un modello in una situazione ipotetica che non puoi testare direttamente che è la preoccupazzione del fare affidamento esclusivamente sull'addestramento costituzionale e sui test empirici del comportamento. Hai anche in linea di principio la capacità di rispondere a domande sul perché il modello si sta comportando in quel modo ad esempio, se sta dicendo qualcosaa che crede sia falso o nascondendo le sue vere capacità e quindi è possibile cogliere segnali preoccupanti aanche quando non c'è nulla di visibilmente sbagliato nel comportamento del modello. Per fare una semplice analogia, un orologio a molla può ticchettare normalmente, rendendo molto difficile dire che è probabile che si rompa il mese prossimo, ma aprire l'orologio e guardare dentro può rivelare debolezze meccaniche che ti permettono di capirlo.

L'IA Costituzionale (insieme a metodi di allineamento simili) e l'interpretabilità meccanicistica sono più potenti se usate insieme, come un processo di va-e-vieni per migliorare l'addestramento di Claude e poi testare i problemi. La costituzione riflette profondamente sulla nostra personalità intenzionale per Claude; le tecniche di interpretabilità possono fornirci una finestra per capire se quella personalità intenzionale ha preso piede.[16]

La terza cosa che possiamo fare per aiutare ad affrontare i rischi di autonomia è costruire l'infrastruttura necessaria per monitorare i nostri modelli nell'uso reale interno ed esterno,[17] e condividere pubblicamente qualsiasi problema troviamo. Più le persone sono consapevoli di un particolare modo in cui è stato osservato un comportamento scorretto negli odierni sistemi di IA, più utenti, analisti e ricercatori possono vigilare su questo comportamento o simili nei sistemi presenti o futuri. Permette anche alle aziende di IA di imparare l'una dall'altra quando le

preoccuupazioni vengono divulgate pubblicamente da un'azienda, anche le altre aziende possono tenerle d'occhio. E se tutti divulgano i problemi, allora l'industria nel suo complesso ottiene un quadro molto migliore di dove le cose stanno andando bene e dove stanno andando male.

Anthropic ha cercato di farlo il più possibile. Stiamo investendo in un'ampia gamma di valutazioni per comprendere i comportamenti dei nostri modelli in laboratorio, così come in strumenti di monitoraggio per osservare i comportamenti nel mondo reale (quando consentito dai clienti). Questo sarà essenziale per dare a noi e agli altri le informazioni empiriche necessarie per prendere decisioni migliori su come questi sistemi operano e su come si rompono. Pubblichiamo "schede di sistema" (system cards) con ogni rilascio di modello che mirano alla completezza e a un'esplorazione approfondita dei possibili rischi. Le nostre schede di sistema raggiungono spesso centinaia di pagine e richiedono uno sforzo pre-rilascio sostanziale che avremmo potuto dedicare al perseguimento del massimo vantaggio commerciale. Abbiamo anche diffuso più rumorosamente i comportamenti dei modelli quando ne vediamo di particolarmente preoccupanti, come la tendenza a ricorrere al ricatto.

La quarta cosa che possiamo fare è incoraggiare il coordinamento per affrontare i rischi di autonomia a livello di industria e società. Sebbene sia incredibilmente prezioso per le singole aziende di IA adottare buone pratiche o diventare brave a guidare i modelli di IA, e condividere pubblicamente le proprie scoperte, la realtà è che non tutte le aziende di IA lo fanno, e le peggiori possono comunque essere un pericolo per tutti anche se le migliori hanno pratiche eccellenti. Ad esempio, alcune aziende di IA hanno mostrato una preoccupante negligenza nei confronti della sessualizzazione dei bambini nei modelli odierni, il che mi fa dubitare che mostreranno l'inclinazione o la capacità di affrontare i rischi di autonomia nei modelli futuri. Inoltre, la corsa commerciale tra le aziende di IA continuerà solo a riscaldarsi e, sebbene la scienza della guida dei modelli possa avere alcuni benefici commerciali, complessivamente l'intensità della corsa renderà sempre più difficile concentrarsi sull'affrontare i rischi di autonomia. Credo che l'unica soluzione sia la legislazione leggi che influenzino direttamente il comporttamento delle aziende di IA o che incentivino in altro modo la R&S per risolvere questi problemi.

Qui vale la pena tenere a mente gli avvertimenti che ho dato all'inizio di questo saggio sull'incertezza e gli interventi chirurgici. Non sappiamo con certezza se i rischi di autonomia saranno un problema serio come ho detto, rifiuto le affermazioni seconndo cui il pericolo sia inevitabile o che qualcosa andrà storto per impostazione predefinita. Un rischio credibile di pericolo è sufficiente per me e per Anthropic per pagare costi piuttosto significativi per affrontarlo, ma una volta entrati nell'ambito della regolamentazione, costringiamo una vasta gamma di attori a sopportare costi economici, e molti di questi attori non credono che il rischio di autonomia sia reale o che l'IA diventerà abbastanza potente da essere una minaccia. Credo che questi attori si sbaglino, ma dovremmo essere pragmatici riguardo alla quantità di opposizione che ci aspettiamo di vedere e ai pericoli di eccesso di regolamentazione. C'è anche un rischio reale che una legislazione eccessivamente prescrittiva finisca per imporre test o regole che non migliorano effettivamente la sicurezza ma che fanno perdere molto tempo (essenzialmente configurandosi come "teatro della sicurezza") anche questo

causerebbe reaazioni avverse e farebbe apparire sciocca la legislazione sulla sicurezza.[18]
La visione di Anthropic è stata che il posto giusto da cui iniziare è la legislazione sulla trasparenza, che essenzialmente cerca di richiedere che ogni azienda di IA di frontiera adotti le pratiche di trasparenza che ho descritto in precedenza in questa sezione. La SB 53 della California e il RAISE Act di New York sono esempi di questo tipo di legislazione, che Anthropic ha sostenuto e che sono stati approvati con successo. Nel sostenere e aiutare a elaborare queste leggi, abbiamo posto particolare attenzione nel cercare di minimizzare i danni collaterali, ad esempio esentando dalla legge le aziende più piccole che difficilmente produrranno modelli di frontiera.[19] La nostra speranza è che la legislazione sulla trasparenza dia nel tempo un senso migliore di quanto probabili o gravi si stiano delineando i rischi di autonomia, così come la natura di questi rischi e come prevenirli al meglio. Man mano che emergono prove più specifiche e azionabili dei rischi (se emergeranno), la legislazione futura nei prossimi anni potrà essere focalizzata chirurgicamente sulla direzione precisa e ben documentata dei rischi, minimizzando i danni collaterali. Per essere chiari, se emergono prove veramente forti di rischi, allora le regole dovrebbero essere proporzionalmente forti.
Nel complesso, sono ottimista sul fatto che un mix di addestramento all'allineamento, interpretabilità meccanicistica, sforzi per trovare e divulgare pubblicamente comportamenti preoccupanti, salvaguardie e regole a livello sociale possa affrontare i rischi di autonomia dell'IA, sebbene io sia più preoccupato per le regole a livello sociale e per il comportamento degli attori meno responsabili (e sono proprio gli attori meno responsabili quelli che sostengono con più forza l'assenza di regolamentazione). Credo che il rimedio sia quello che è sempre in una democrazia: quelli di noi che credono in questa causa dovrebbero sostenere che questi rischi sono reali e che i nostri concittadini devono unirsi per proteggersi.

2. Una sorprendente e terribile responsabilizzazione

Uso improprio per la distruzione

Supponiamo che i problemi di autonomia dell'IA siano stati risolti ﷺ non siamo più preoccupati che il paese di geni dell'IA diventi canaglia e sopraffaccia l'umanità. I geni dell'IA fanno ciò che gli umani vogliono che facciano e, poiché hanno un enorme valore commerciale, individui e organizzazioni in tutto il mondo possono "affittare" uno o più geni dell'IA per svolgere vari compiti per loro.
Avere tutti un genio superintelligente in tasca è un progresso straordinario e porterà a un'incredibile creazione di valore economico e miglioramento della qualità della vita umana. Parlo di questi benefici in grande dettaglio in Machines of Loving Grace. Ma non ogni effetto del rendere tutti superhumanamente capaci sarà positivo. Può potenzialmente amplificare la capacità di individui o piccoli gruppi di causare distruzione su una scala molto più vasta di quanto fosse possibile prima, facendo uso di strumenti sofisticati e pericolosi (come le armi di distruzione di massa) che in precedenza erano disponibili solo a pochi eletti con un alto livello di competenza, addestramento specializzato e concentrazione.
Come scrisse Bill Joy 25 anni fa in Why the Future Doesn’t Need Us:[20]

Costruire armi nucleari richiedeva, almeno per un certo periodo, l'accesso sia a materie prime rare anzi, effettivamente non disponibili sia a a informazioni protette; anche i programmi di armi biologiche e chimiche tendevano a richiedere attività su larga scala. Le tecnologie del XXI secolo genetica, nanoteecnologia e robotica... possono generare intere nuove classi di incidenti e abusi... ampiamente alla portata di individui o piccoli gruppi. Non richiederanno grandi impianti o materie prime rare... siamo sulla soglia dell'ulteriore perfezionamento del male estremo, un male la cui possibilità si estende ben oltre ciò che le armi di distruzione di massa hanno lasciato in eredità agli stati-nazione, fino a una sorprendente e terribile responsabilizzazione (empowerment) di individui estremi.

Ciò a cui Joy punta è l'idea che causare distruzione su larga scala richieda sia motivazione che capacità, e finché la capacità è limitata a un piccolo insieme di persone altamente addestrate, c'è un rischio relativamente limitato che singoli individui (o piccoli gruppi) causino tale distruzione.[21] Un solitario disturbato può perpetrare una sparatoria in una scuola, ma probabilmente non può costruire un'arma nucleare o scatenare una piaga.

Infatti, capacità e motivazione possono persino essere correlate negativamente. Il tipo di persona che ha la capacità di scatenare una piaga è probabilmente altamente istruita: verosimilmente un dottorato di ricerca in biologia molecolare, e uno particolarmente intraprendente, con una carriera promettente, una personalità stabile e disciplinata e molto da perdere. Questo tipo di persona è improbabile che sia interessato a uccidere un numero enorme di persone per nessun beneficio proprio e con un grande rischio per il proprio futuro dovrebbe essere motivato da pura malizia, rancore intenso o instabilità.

Tali persone esistono, ma sono rare e tendono a diventare storie enormi quando compaiono, proprio perché sono così insolite.[22] Tendono anche a essere difficili da catturare perché sono intelligenti e capaci, talvolta lasciando misteri che richiedono anni o decenni per essere risolti. L'esempio più famoso è probabilmente il matematico Theodore Kaczynski (Unabomber), che sfuggì alla cattura dell'FBI per quasi 20 anni, spinto da un'ideologia anti-tecnologica. Un altro esempio è il ricercatore di biodifesa Bruce Ivins, che sembra aver orchestrato una serie di attacchi all'antrace nel 2001. È successo anche con organizzazioni non statali qualificate: la setta Aum Shinrikyo riuscì a ottenere gas nervino sarin e a uccidere 14 persone (oltre a ferirne centinaia) rilasciandolo nella metropolitana di Tokyo nel 1995.

Fortunatamente, nessuno di questi attacchi ha utilizzato agenti biologici contagiosi, perché la capacità di costruire o ottenere questi agenti era al di là delle possibilità persino di queste persone.[23] I progressi nella biologia molecolare hanno ora abbassato significativamente la barriera per la creazione di armi biologiche (specialmente in termini di disponibilità di materiali), ma richiede comunque un'enorme competenza per farlo. Temo che un genio in tasca a tutti possa rimuovere quella barriera, rendendo essenzialmente chiunque un virologo con dottorato che può essere guidato passo-passo attraverso il processo di progettazione, sintesi e rilascio di un'arma biologica. Impedire l'ottenimento di questo tipo di informazioni a fronte di una seria pressione avversaria i cosiddetti "jailbreak" richiede probabilmente strati di difesa oltre a quelli ordinariamente integrati nell'addestramento.

Cosa cruciale, questo romperà la correlazione tra capacità e motivazione: il solitario disturbato che vuole uccidere persone ma manca della disciplina o della competenza per farlo sarà ora elevato al livello di capacità del virologo con dottorato, che è improbabile che abbia questa motivazione. Questa preoccupazione si generalizza oltre la biologia (sebbene io pensi che la biologia sia l'area più spaventosa) a qualsiasi area in cui è possibile una grande distruzione ma che attualmente richiede un alto livello di competenza e disciplina. Per dirla in altro modo, affittare un'IA potente dà intelligenza a persone malintenzionate (ma per il resto medie). Temo che ci siano potenzialmente un gran numero di tali persone là fuori e che se avessero accesso a un modo semplice per uccidere milioni di persone, prima o poi una di loro lo farebbe. Inoltre, coloro che hanno competenza potrebbero essere messi in grado di commettere distruzioni su scala ancora più vasta di quanto potessero prima.

La biologia è di gran lunga l'area di cui sono più preoccupato, a causa del suo grandissimo potenziale di distruzione e della difficoltà di difendersi, quindi mi concentrerò sulla biologia in particolare. Ma molto di ciò che dico qui si applica ad altri rischi, come attacchi informatici, armi chimiche o tecnologia nucleare.

Non entrerò nei dettagli su come produrre armi biologiche, per ragioni che dovrebbero essere ovvie. Ma a un alto livello, temo che gli LLM si stiano avvicinando (o abbiano già raggiunto) la conoscenza necessaria per crearle e rilasciarle end-to-end, e che il loro potenziale di distruzione sia molto alto. Alcuni agenti biologici potrebbero causare milioni di morti se venisse compiuto uno sforzo determinato per rilasciarli per la massima diffusione. Tuttavia, ciò richiederebbe comunque un livello di competenza molto elevato, inclusa una serie di passaggi e procedure molto specifici che non sono ampiamente noti. La mia preoccupazione non riguarda solo la conoscenza fissa o statica. Temo che gli LLM saranno in grado di prendere qualcuno con conoscenze e capacità medie e guidarlo attraverso un processo complesso che altrimenti potrebbe andare storto o richiedere il debug in modo interattivo, in modo simile a come il supporto tecnico potrebbe aiutare una persona non tecnica a fare il debug e risolvere complicati problemi legati al computer (anche se questo sarebbe un processo più esteso, probabilmente duraturo settimane o mesi). LLM più capaci (sostanzialmente oltre la potenza di quelli odierni) potrebbero essere capaci di consentire atti ancora più spaventosi. Nel 2024, un gruppo di eminenti scienziati ha scritto una lettera avvertendo sui rischi della ricerca, e della potenziale creazione, di un nuovo tipo di organismo pericoloso: la "vita speculare" (mirror life). Il DNA, l'RNA, i ribosomi e le proteine che compongono gli organismi biologici hanno tutti la stessa chiralità (chiamata anche "orientamento") che fa sì che non siano equivalenti a una versione di se stessi riflessa nello specchio (proprio come la tua mano destra non può essere ruotata in modo tale da essere identica alla tua sinistra). Ma l'intero sistema di proteine che si legano tra loro, il macchinario della sintesi del DNA e della traduzione dell'RNA e la costruzione e scomposizione delle proteine, dipende tutto da questo orientamento. Se gli scienziati creassero versioni di questo materiale biologico con l'orientamento opposto – e ci sono alcuni potenziali vantaaggi in queste, come farmaci che durano più a lungo nel corpo – potrebbe essere estremamente pericoloso. Questo perché la vita "mancina", se fosse creata sotto forma di organismi completi capaci di riproduzione (il che sarebbe

difficilissimo), sarebbe potenzialmente indigesta per qualsiasi sistema che scompone il materiale biologico sulla terra avrebbe una "chiave" chhe non entrerebbe nella "serratura" di nessun enzima esistente. Ciò significherebbe che potrebbe proliferare in modo incontrollabile e soppiantare ogni vita sul pianeta, nel peggiore dei casi distruggendo persino tutta la vita sulla terra.
Esiste una sostanziale incertezza scientifica sia sulla creazione che sui potenziali effetti della vita speculare. La lettera del 2024 accompagnava un rapporto che concludeva che "i batteri speculari potrebbero plausibilmente essere creati nel prossimo decennio o in pochi decenni", il che è un intervallo ampio. Ma un modello di IA sufficientemente potente (per essere chiari, molto più capace di qualsiasi modello attuale) potrebbe essere in grado di scoprire come crearla molto più rapidamente e aiutare effettivamente qualcuno a farlo.<
La mia visione è che, sebbene si tratti di rischi oscuri e possano sembrare improbabili, l'entità delle conseguenze è così grande che dovrebbero essere presi sul serio come un rischio di prima classe dei sistemi di IA.
Gli scettici hanno sollevato una serie di obiezioni sulla gravità di questi rischi biologici derivanti dagli LLM, con le quali non sono d'accordo ma che vale la pena affrontare. La maggior parte rientra nella categoria del non apprezzare la traiettoria esponenziale su cui si trova la tecnologia. Nel 2023, quando abbiamo iniziato a parlare per la prima volta dei rischi biologici derivanti dagli LLM, gli scettici dicevano che tutte le informazioni necessarie erano disponibili su Google e che gli LLM non aggiungevano nulla a questo. Non è mai stato vero che Google potesse darti tutte le informazioni necessarie: i genomi sono liberamente disponibili ma, come ho detto sopra, alcuni passaggi chiave così come una enorme quantità di know-how pratico non possono essere ottenuti in quel modo. Ma anche, entro la fine del 2023 gli LLM stavano chiaramente fornendo informazioni oltre ciò che Google poteva dare per alcuni passaggi del processo.
Dopo questo, gli scettici si sono ritirati sull'obiezione che gli LLM non fossero utili end-to-end e non potessero aiutare con l'acquisizione di armi biologiche rispetto al semplice fornire informazioni teoriche. A metà del 2025, le nostre misurazioni mostrano che gli LLM potrebbero già fornire un aumento sostanziale (uplift) in diverse aree rilevanti, forse raddoppiando o triplicando la probabilità di successo. Ciò ci ha portato a decidere che Claude Opus 4 (e i successivi modelli Sonnet 4.5, Opus 4.1 e Opus 4.5) dovevano essere rilasciati sotto le nostre protezioni di AI Safety Level 3 nel nostro quadro della Responsible Scaling Policy, e a implementare salvaguardie contro questo rischio (più su questo dopo). Crediamo che i modelli stiano probabilmente approcciando il punto in cui, senza salvaguardie, potrebbero essere utili per consentire a qualcuno con una laurea STEM ma non specificamente in biologia di completare l'intero processo di produzione di un'arma biologica.
Un'altra obiezione è che ci sono altre azioni non correlate all'IA che la società può intraprendere per bloccare la produzione di armi biologiche. Soprattutto, l'industria della sintesi genetica produce campioni biologici su richiesta e non esiste un requisito federale che imponga ai fornitori di sottoporre a screening gli ordini per assicurarsi che non contengano agenti patogeni. Uno studio del MIT ha scoperto che 36 fornitori su 38 hanno evaso un ordine contenente la sequenza dell'influenza del 1918. Sono

favorevole a uno screening obbligatorio della sintesi genetica che renderebbe più difficile per gli individui trasformare in armi gli agenti patogeni, al fine di ridurre sia i rischi biologici guidati dall'IA sia i rischi biologici in generale. Ma questo non è qualcosa che abbiamo oggi. Sarebbe anche solo uno strumento per ridurre il rischio; è un complemento ai guardrail sui sistemi di IA, non un sostituto.

La migliore obiezione è quella che ho visto sollevare raramente: che esiste un divario tra l'utilità in linea di principio dei modelli e l'effettiva propensione dei cattivi attori a usarli. La maggior parte dei singoli cattivi attori sono individui disturbati, quindi quasi per definizione il loro comportamento è imprevedibile e irrazionale [illegible] e sono proprio questi caattivi attori, quelli non qualificati, che avrebbero potuto trarre il massimo beneficio dall'IA che rende molto più facile uccidere molte persone.[24] Solo perché un tipo di attacco violento è possibile, non significa che qualcuno deciderà di farlo. Forse gli attacchi biologici risulteranno poco attraenti perché hanno una ragionevole probabilità di infettare l'autore, non soddisfano le fantasie in stile militare che molti individui o gruppi violenti hanno ed è difficile colpire selettivamente persone specifiche. Potrebbe anche essere che passare attraverso un processo che richiede mesi, anche se un'IA ti guida, comporti una dose di pazienza che la maggior parte degli individui disturbati semplicemente non ha. Potremmo semplicemente essere fortunati e la motivazione e la capacità non si combineranno, in pratica, proprio nel modo giusto.

Ma questa sembra una protezione molto fragile su cui fare affidamento. Le motivazioni dei solitari disturbati possono cambiare per qualsiasi ragione o per nessuna ragione e, di fatto, ci sono già istanze di LLM utilizzati in attacchi (solo non con la biologia). La concentrazione sui solitari disturbati ignora anche i terroristi motivati ideologicamente, che sono spesso disposti a spendere grandi quantità di tempo e fatica (ad esempio, i dirottatori dell'11 settembre). Volere uccidere quante più persone possibile è una motivazione che probabilmente sorgerà prima o poi e, sfortunatamente, suggerisce le armi biologiche come metodo. Anche se questa motivazione fosse estremamente rara, deve materializzarsi solo una volta. E man mano che la biologia avanza (sempre più guidata dall'IA stessa), potrebbe anche diventare possibile compiere attacchi più selettivi (ad esempio, mirati contro persone con specifiche discendenze), il che aggiunge un'altra, agghiacciante, possibile motivazione.

Non penso che gli attacchi biologici verranno necessariamente compiuti nell'istante in cui diventerà ampiamente possibile farlo [illegible] anzi, scommetterei contro. Ma sommmati su milioni di persone e qualche anno di tempo, penso ci sia un serio rischio di un attacco massiccio, e le conseguenze sarebbero così gravi (con vittime potenzialmente nell'ordine dei milioni o più) che credo non abbiamo altra scelta se non quella di prendere misure serie per prevenirlo.

Difese

Questo ci porta a come difenderci da questi rischi. Qui vedo tre cose che possiamo fare.

In primo luogo, le aziende di IA possono mettere dei guardrail sui loro modelli per impedire loro di aiutare a produrre armi biologiche. Anthropic lo sta facendo molto attivamente. La Costituzione di Claude, che si concentra principalmente su principi e

valori di alto livello, ha un piccolo numero di specifici divieti assoluti, e uno di questi riguarda l'aiuto nella produzione di armi biologiche (o chimiche, nucleari o radiologiche). Ma tutti i modelli possono essere sottoposti a jailbreak e quindi, come seconda linea di difesa, abbiamo implementato (da metà 2025, quando i nostri test hanno mostrato che i nostri modelli si stavano avvicinando alla soglia in cui avrebbero potuto iniziare a rappresentare un rischio) un classificatore che rileva e blocca specificamente gli output relativi alle armi biologiche. Aggiorniamo e miglioriamo regolarmente questi classificatori e li abbiamo generalmente trovati altamente robusti anche contro sofisticati attacchi avversari.[25] Questi classificatori aumentano i costi di gestione dei nostri modelli in modo misurabile (in alcuni modelli, sono vicini al 5% dei costi totali di inferenza) e quindi riducono i nostri margini, ma riteniamo che usarli sia la cosa giusta da fare.

Per loro merito, anche alcune altre aziende di IA hanno implementato dei classificatori. Ma non tutte le aziende lo hanno fatto e non c'è nemmeno nulla che richieda alle aziende di mantenere i loro classificatori. Temo che nel tempo possa crearsi un dilemma del prigioniero in cui le aziende possono defezionare e abbassare i loro costi rimuovendo i classificatori. Questo è ancora una volta un classico problema di esternalità negative che non può essere risolto dalle sole azioni volontarie di Anthropic o di qualsiasi altra singola azienda.[26] Gli standard industriali volontari possono aiutare, così come le valutazioni e le verifiche di terze parti del tipo effettuato dagli istituti di sicurezza dell'IA e da valutatori terzi.

Ma alla fine la difesa potrebbe richiedere un'azione del governo, che è la seconda cosa che possiamo fare. Le mie opinioni qui sono le stesse di quelle per affrontare i rischi di autonomia: dovremmo iniziare con requisiti di trasparenza,[27] che aiutino la società a misurare, monitorare e difendersi collettivamente dai rischi senza interrompere l'attività economica in modo pesante. Poi, se e quando raggiungeremo soglie di rischio più chiare, potremo elaborare una legislazione che miri più precisamente a questi rischi e abbia una minore probabilità di danni collaterali. Nel caso particolare delle armi biologiche, penso che il tempo per una tale legislazione mirata possa avvicinarsi presto – Anthropic e altre aziende stanno imparando sempre di più sulla natura dei rischi biologici e su cosa sia ragionevole richiedere alle aziende nel difendersi da essi. Difendersi pienamente da questi rischi potrebbe richiedere di lavorare a livello internazionale, anche con avversari geopolitici, ma esiste un precedente nei trattati che vietano lo sviluppo di armi biologiche. Sono generalmente scettico sulla maggior parte dei tipi di cooperazione internazionale sull'IA, ma questa potrebbe essere una piccola area in cui c'è qualche possibilità di ottenere un freno globale. Nemmeno le dittature vogliono attacchi bioterroristici massicci.

Infine, la terza contromisura che possiamo prendere è cercare di sviluppare difese contro gli attacchi biologici stessi. Ciò potrebbe includere il monitoraggio e il tracciamento per il rilevamento precoce, investimenti nella R&S per la purificazione dell'aria (come la disinfezione far-UVC), lo sviluppo rapido di vaccini che possano rispondere e adattarsi a un attacco, migliori dispositivi di protezione individuale (DPI),[28] e trattamenti o vaccinazioni per alcuni degli agenti biologici più probabili. I vaccini mRNA, che possono essere progettati per rispondere a un particolare virus o

variante, sono un primo esempio di ciò che è possibile qui. Anthropic è entusiasta di lavorare con aziende biotecnologiche e farmaceutiche su questo problema. Ma sfortunatamente penso che le nostre aspettative sul lato difensivo dovrebbero essere limitate. Esiste un'asimmetria tra attacco e difesa in biologia, perché gli agenti si diffondono rapidamente da soli, mentre le difese richiedono che il rilevamento, la vaccinazione e il trattamento siano organizzati tra un gran numero di persone molto rapidamente in risposta. A meno che la risposta non sia fulminea (il che accade raramente), gran parte del danno sarà fatto prima che una risposta sia possibile. È concepibile che futuri miglioramenti tecnologici possano spostare questo equilibrio a favore della difesa (e dovremmo certamente usare l'IA per aiutare a sviluppare tali progressi tecnologici), ma fino ad allora, le salvaguardie preventive saranno la nostra principale linea di difesa.

Vale la pena menzionare brevemente gli attacchi informatici qui, poiché a differenza degli attacchi biologici, gli attacchi informatici guidati dall'IA sono già avvenuti nel mondo reale, anche su larga scala e per spionaggio sponsorizzato dagli stati. Ci aspettiamo che questi attacchi diventino più capaci man mano che i modelli avanzano rapidamente, fino a diventare il modo principale in cui vengono condotti gli attacchi informatici. Mi aspetto che gli attacchi informatici guidati dall'IA diventino una minaccia seria e senza precedenti per l'integrità dei sistemi informatici in tutto il mondo, e Anthropic sta lavorando sodo per fermare questi attacchi e, alla fine, prevenirli in modo affidabile. La ragione per cui non mi sono concentrato sul cyber tanto quanto sulla biologia è che (1) gli attacchi informatici sono molto meno propensi a uccidere persone, certamente non alla scala degli attacchi biologici, e (2) l'equilibrio attacco-difesa potrebbe essere più gestibile nel cyber, dove c'è almeno qualche speranza che la difesa possa tenere il passo con (e idealmente superare) l'attacco dell'IA se investiamo in essa correttamente.

Sebbene la biologia sia attualmente il vettore di attacco più serio, ci sono molti altri vettori ed è possibile che ne emerga uno più pericoloso. Il principio generale è che senza contromisure, l'IA è probabile che abbassi continuamente la barriera all'attività distruttiva su scala sempre più vasta, e l'umanità ha bisogno di una risposta seria a questa minaccia.

3. L'odioso apparato

Uso improprio per la presa del potere

La sezione precedente ha discusso il rischio che individui e piccole organizzazioni cooptino un piccolo sottoinsieme del "paese di geni in un datacenter" per causare distruzione su larga scala. Ma dovremmo anche preoccuparci probabilmente inn modo sostanzialmente maggiore dell'uso improprio delll'IA allo scopo di esercitare o prendere il potere, verosimilmente da parte di attori più grandi e consolidati.[29]

In Machines of Loving Grace, ho discusso la possibilità che governi autoritari usino un'IA potente per sorvegliare o reprimere i propri cittadini in modi che sarebbero estremamente difficili da riformare o rovesciare. Le attuali autocrazie sono limitate in quanto repressive possono essere dalla necessità di avere umani che eseguano i loro ordini, e gli umani spesso hanno dei limiti in quanto disumani sono disposti a essere.

Ma le autocrazie abilitate dall'IA non avrebbero tali limiti.
Peggio ancora, i paesi potrebbero anche usare il loro vantaggio nell'IA per ottenere potere su altri paesi. Se il "paese di geni" nel suo insieme fosse semplicemente posseduto e controllato dall'apparato militare di un singolo paese (umano) e altri paesi non avessero capacità equivalenti, è difficile vedere come potrebbero difendersi: verrebbero superati in astuzia ad ogni mossa, in modo simile a una guerra tra umani e topi. Mettendo insieme queste due preoccupazioni si arriva all'allarmante possibilità di una dittatura totalitaria globale. Ovviamente, dovrebbe essere una delle nostre massime priorità prevenire questo esito.
Esistono molti modi in cui l'IA potrebbe abilitare, consolidare o espandere l'autocrazia, ma ne elencherò alcuni di cui sono più preoccupato. Si noti che alcune di queste applicazioni hanno usi difensivi legittimi e non sto necessariamente argomentando contro di esse in termini assoluti; sono tuttavia preoccupato che tendano strutturalmente a favorire le autocrazie:
• Armi completamente autonome. Uno sciame di milioni o miliardi di droni armati completamente automatizzati, controllati localmente da un'IA potente e coordinati strategicamente in tutto il mondo da un'IA ancora più potente, potrebbe essere un esercito imbattibile, capace sia di sconfiggere qualsiasi esercito al mondo sia di sopprimere il dissenso all'interno di un paese seguendo ogni cittadino. Gli sviluppi nella guerra Russia-Ucraina dovrebbero avvertirci del fatto che la guerra dei droni è già tra noi (sebbene non ancora completamente autonoma, e una minuscola frazione di ciò che potrebbe essere possibile con un'IA potente). La R&S derivante da un'IA potente potrebbe rendere i droni di un paese di gran lunga superiori a quelli di altri, accelerare la loro produzione, renderli più resistenti agli attacchi elettronici, migliorare la loro manovrabilità e così via. Naturalmente, queste armi hanno anche usi legittimi nella difesa della democrazia: sono state fondamentali per difendere l'Ucraina e sarebbero probabilmente fondamentali per difendere Taiwan. Ma sono un'arma pericolosa da brandire: dovremmo preoccuparcene nelle mani delle autocrazie, ma anche temere che, poiché sono così potenti e con così poca responsabilità, vi sia un rischio notevolmente aumentato che i governi democratici le rivolgano contro il proprio popolo per prendere il potere.
• Sorveglianza IA. Un'IA sufficientemente potente potrebbe probabilmente essere utilizzata per compromettere qualsiasi sistema informatico al mondo,[30] e potrebbe anche usare l'accesso ottenuto in questo modo per leggere e dare un senso a tutte le comunicazioni elettroniche del mondo (o persino a tutte le comunicazioni mondiali di persona, se si possono costruire o requisire dispositivi di registrazione). Potrebbe essere spaventosamente plausibile generare un elenco completo di chiunque non sia d'accordo con il governo su un qualsiasi numero di questioni, anche se tale disaccordo non è esplicito in nulla di ciò che dicono o fanno. Un'IA potente che osserva miliardi di conversazioni di milioni di persone potrebbe valutare il sentimento pubblico, rilevare sacche di slealtà in formazione e stroncarle prima che crescano. Ciò potrebbe portare all'imposizione di un vero panopticon su una scala che non vediamo oggi, nemmeno con il PCC.
• Propaganda IA. Gli odierni fenomeni di "psicosi da IA" e "fidanzate IA" suggeriscono che anche al loro attuale livello di intelligenza, i modelli di IA possono

avere una potente influenza psicologica sulle persone. Versioni molto più potenti di questi modelli, che fossero molto più integrate e consapevoli della vita quotidiana delle persone e potessero modellarle e influenzarle nell'arco di mesi o anni, sarebbero probabilmente capaci di lavare essenzialmente il cervello a molti (alla maggior parte?) in qualsiasi ideologia o atteggiamento desiderato, e potrebbero essere impiegate da un leader senza scrupoli per garantire lealtà e sopprimere il dissenso, anche di fronte a un livello di repressione a cui la maggior parte delle popolazioni si ribellerebbe. Oggi la gente si preoccupa molto, ad esempio, della potenziale influenza di TikTok come propaganda del PCC diretta ai bambini. Mi preoccupo anch'io, ma un agente IA personalizzato che impara a conoscerti nel corso degli anni e usa la sua conoscenza di te per dare forma a tutte le tue opinioni sarebbe drammaticamente più potente di questo.

• Processo decisionale strategico. Un paese di geni in un datacenter potrebbe essere usato per consigliare un paese, un gruppo o un individuo sulla strategia geopolitica, quello che potremmo chiamare un "Bismarck virtuale". Potrebbe ottimizzare le tre strategie sopra descritte per prendere il potere e probabilmente svilupparne molte altre a cui non ho pensato (ma a cui un paese di geni potrebbe pensare). Diplomazia, strategia militare, R&S, strategia economica e molte altre aree sono tutte suscettibili di un aumento sostanziale di efficacia grazie a un'IA potente. Molte di queste abilità sarebbero legittimamente utili per le democrazie – vogliamo che le democrazzie abbiano accesso alle migliori strategie per difendersi dalle autocrazie – ma il potenziale di uso improprioo nelle mani di chiunque rimane comunque.

Avendo descritto cosa mi preoccupa, passiamo a chi. Mi preoccupano le entità che hanno il maggior accesso all'IA, che partono da una posizione di maggior potere politico o che hanno una storia pregressa di repressione. In ordine di gravità, mi preoccupano:

1. Il PCC. La Cina è seconda solo agli Stati Uniti nelle capacità di IA ed è il paese con la maggiore probabilità di superare gli Stati Uniti in tali capacità. Il loro governo è attualmente autocratico e gestisce uno stato di sorveglianza ad alta tecnologia. Ha già dispiegato la sorveglianza basata sull'IA (inclusa la repressione degli Uiguri) e si ritiene che impieghi propaganda algoritmica via TikTok (oltre ai suoi molti altri sforzi di propaganda internazionale). Hanno senza dubbio il percorso più chiaro verso l'incubo totalitario abilitato dall'IA che ho esposto sopra. Potrebbe persino essere l'esito predefinito all'interno della Cina, così come all'interno di altri stati autocratici a cui il PCC esporta tecnologia di sorveglianza. Ho scritto spesso sulla minaccia che il PCC prenda il comando nell'IA e sull'imperativo esistenziale di impedire che ciò accada. Ecco perché. Per essere chiari, non sto prendendo di mira la Cina per animosità verso di loro in particolare – sono semplicemente il paese che più unisce prodezzza nell'IA, un governo autocratico e uno stato di sorveglianza ad alta tecnologia. Semmai, è il popolo cinese stesso quello che ha più probabilità di soffrire della repressione abilitata dall'IA del PCC, e non ha voce nelle azioni del proprio governo. Ammiro e rispetto enormemente il popolo cinese e sostengo i molti coraggiosi dissidenti all'interno della Cina e la loro lotta per la libertà.
2. Democrazie competitive nell'IA. Come ho scritto sopra, le democrazie hanno un legittimo interesse in alcuni strumenti militari e geopolitici alimentati dall'IA, perché

i governi democratici offrono la migliore possibilità di contrastare l'uso di questi strumenti da parte delle autocrazie. In generale, sono favorevole ad armare le democrazie con gli strumenti necessari per sconfiggere le autocrazie nell'era dell'IA
 semplicemmente non penso ci sia altro modo. Ma non possiamo ignorare il potenziale di abuso di queste tecnologie da parte dei governi democratici stessi. Le democrazie normalmente hanno salvaguardie che impediscono al loro apparato militare e di intelligence di essere rivolto all'interno contro la propria popolazione,[31] ma poiché gli strumenti di IA richiedono così poche persone per operare, c'è il potenziale per eludere queste salvaguardie e le norme che le sostengono. Vale anche la pena notare che alcune di queste salvaguardie si stanno già gradualmente erodendo in alcune democrazie. Pertanto, dovremmo armare le democrazie con l'IA, ma dovremmo farlo con attenzione e entro certi limiti: sono il sistema immunitario di cui abbiamo bisogno per combattere le autocrazie ma, come il sistema immunitario, c'è il rischio che si rivolti contro di noi e diventi esso stesso una minaccia.

3. Paesi non democratici con grandi datacenter. Oltre alla Cina, la maggior parte dei paesi con governance meno democratica non sono attori leader dell'IA nel senso che non hanno aziende che producono modelli di IA di frontiera. Pertanto pongono un rischio fondamentalmente diverso e minore rispetto al PCC, che rimane la preoccupazione principale (la maggior parte sono anche meno repressivi, e quelli più repressivi, come la Corea del Nord, non hanno alcuna industria dell'IA significativa). Ma alcuni di questi paesi hanno grandi datacenter (spesso come parte di installazioni da parte di aziende che operano in democrazie), che possono essere usati per eseguire IA di frontiera su larga scala (sebbene ciò non conferisca la capacità di spingere la frontiera). C'è una certa quantità di pericolo associata a questo
 questi governi potrebbbero in linea di principio espropriare i datacenter e usare il paese di IA al suo interno per i propri scopi. Sono meno preoccupato di questo rispetto a paesi come la Cina che sviluppano direttamente l'IA, ma è un rischio da tenere presente.[32]

4. Aziende di IA. È alquanto imbarazzante dirlo in qualità di CEO di un'azienda di IA, ma penso che il livello successivo di rischio sia rappresentato proprio dalle aziende di IA stesse. Le aziende di IA controllano grandi datacenter, addestrano modelli di frontiera, hanno la massima competenza su come usare quei modelli e in alcuni casi hanno contatti quotidiani e la possibilità di influenzare decine o centinaia di milioni di utenti. La cosa principale che manca loro è la legittimità e l'infrastruttura di uno stato, quindi gran parte di ciò che servirebbe per costruire gli strumenti di un'autocrazia IA sarebbe illegale per un'azienda di IA o, per lo meno, estremamente sospetto. Ma alcune cose non sono impossibili: potrebbero, ad esempio, usare i loro prodotti IA per lavare il cervello alla loro massiccia base di utenti consumer, e il pubblico dovrebbe stare allerta sul rischio che ciò rappresenta. Penso che la governance delle aziende di IA meriti molta attenzione.

Esistono diversi possibili argomenti contro la gravità di queste minacce, e vorrei crederci, perché l'autoritarismo abilitato dall'IA mi terrorizza. Vale la pena esaminare alcuni di questi argomenti e rispondere.

In primo luogo, alcune persone potrebbero riporre la loro fiducia nel deterrente nucleare, in particolare per contrastare l'uso di armi autonome IA per la conquista militare. Se qualcuno minaccia di usare queste armi contro di te, puoi sempre

minacciare una risposta nucleare in cambio. La mia preoccupazione è che non sono totalmente sicuro che possiamo essere fiduciosi nel deterrente nucleare contro un paese di geni in un datacenter: è possibile che un'IA potente possa ideare modi per rilevare e colpire i sottomarini nucleari, condurre operazioni di influenza contro gli operatori delle infrastrutture per armi nucleari o usare le capacità informatiche dell'IA per lanciare un attacco informatico contro i satelliti usati per rilevare i lanci nucleari.[33] In alternativa, è possibile che la conquista di paesi sia fattibile solo con la sorveglianza IA e la propaganda IA, e non presenti mai un momento chiaro in cui è ovvio cosa stia succedendo e dove una risposta nucleare sarebbe appropriata. Forse queste cose non sono fattibili e il deterrente nucleare sarà ancora efficace, ma la posta in gioco sembra troppo alta per correre il rischio.[34]

Una seconda possibile obiezione è che potrebbero esserci contromisure che possiamo prendere contro questi strumenti di autocrazia. Possiamo contrastare i droni con i nostri droni, la difesa informatica migliorerà insieme all'attacco informatico, potrebbero esserci modi per immunizzare le persone contro la propaganda, ecc. La mia risposta è che queste difese saranno possibili solo con un'IA comparabilmente potente. Se non c'è una controforza con un paese di geni in un datacenter comparabilmente intelligente e numeroso, non sarà possibile eguagliare la qualità o la quantità di droni, né la difesa informatica potrà superare in astuzia l'offesa informatica, ecc. Quindi la questione delle contromisure si riduce alla questione di un equilibrio di potere nell'IA potente. Qui, mi preoccupa la proprietà ricorsiva o auto-rinforzante dell'IA potente (di cui ho discusso all'inizio di questo saggio): che ogni generazione di IA possa essere usata per progettare e addestrare la successiva generazione di IA. Ciò porta a un rischio di un vantaggio galoppante, dove l'attuale leader nell'IA potente potrebbe essere in grado di aumentare il proprio distacco e potrebbe essere difficile da raggiungere. Dobbiamo assicurarci che non sia un paese autoritario ad arrivare per primo a questo ciclo.

Inoltre, anche se si potesse raggiungere un equilibrio di potere, esiste comunque il rischio che il mondo possa essere diviso in sfere autocratiche, come in 1984. Anche se diverse potenze in competizione avessero ciascuna i propri modelli di IA potenti e nessuna potesse sopraffare le altre, ogni potenza potrebbe comunque reprimere internamente la propria popolazione, e sarebbe difficilissima da rovesciare (poiché le popolazioni non hanno IA potenti per difendersi). È quindi importante prevenire l'autocrazia abilitata dall'IA anche se non porta a un singolo paese che conquista il mondo.

Difese

Come ci si difende da questa vasta gamma di strumenti autocratici e potenziali attori di minaccia? Come nelle sezioni precedenti, ci sono diverse cose che penso possiamo fare.

In primo luogo, non dovremmo assolutamente vendere chip, strumenti per la fabbricazione di chip o datacenter al PCC. I chip e gli strumenti per la fabbricazione di chip sono il singolo più grande collo di bottiglia per l'IA potente, e bloccarli è una misura semplice ma estremamente efficace, forse la singola azione più importante che possiamo intraprendere. Non ha senso vendere al PCC gli strumenti con cui costruire uno stato totalitario IA e possibilmente conquistarci militarmente. Vengono

addotti una serie di argomenti complicati per giustificare tali vendite, come l'idea che "diffondere il nostro stack tecnologico nel mondo" permetta all'"America di vincere" in una generica e non meglio specificata battaglia economica. A mio avviso, questo è come vendere armi nucleari alla Corea del Nord e poi vantarsi che i rivestimenti dei missili sono prodotti da Boeing e quindi gli USA stanno "vincendo". La Cina è diversi anni indietro rispetto agli USA nella capacità di produrre chip di frontiera in quantità, e il periodo critico per costruire il paese di geni in un datacenter è molto probabilmente entro i prossimi anni.[35] Non c'è motivo di dare una spinta gigante alla loro industria dell'IA durante questo periodo critico.

In secondo luogo, ha senso usare l'IA per dare potere alle democrazie per resistere alle autocrazie. Questo è il motivo per cui Anthropic ritiene importante fornire l'IA alle comunità di intelligence e difesa degli Stati Uniti e dei suoi alleati democratici. Difendere le democrazie sotto attacco, come l'Ucraina e (tramite attacchi informatici) Taiwan, sembra avere una priorità particolarmente alta, così come dare potere alle democrazie affinché usino i loro servizi di intelligence per disturbare e degradare le autocrazie dall'interno. Ad un certo livello, l'unico modo per rispondere alle minacce autocratiche è eguagliarle e superarle militarmente. Una coalizione degli Stati Uniti e dei suoi alleati democratici, se raggiungesse la predominanza nell'IA potente, sarebbe in grado non solo di difendersi dalle autocrazie, ma di contenerle e limitare i loro abusi totalitari basati sull'IA.

In terzo luogo, dobbiamo tracciare una linea netta contro gli abusi dell'IA all'interno delle democrazie. Ci devono essere dei limiti a ciò che permettiamo ai nostri governi di fare con l'IA, affinché non prendano il potere o reprimano il proprio popolo. La formulazione che ho ideato è che dovremmo usare l'IA per la difesa nazionale in tutti i modi tranne quelli che ci renderebbero più simili ai nostri avversari autocratici. Dove dovrebbe essere tracciata la linea? Nell'elenco all'inizio di questa sezione, due elementi – l'uso dell'IA per la sorveglianza di massa ddomestica e la propaganda di massa – mi sembrano linee rrosse invalicabili e del tutto illegittime. Alcuni potrebbero obiettare che non c'è bisogno di fare nulla (almeno negli Stati Uniti), poiché la sorveglianza di massa domestica è già illegale ai sensi del Quarto Emendamento. Ma il rapido progresso dell'IA potrebbe creare situazioni per cui i nostri quadri legali esistenti non sono ben progettati. Ad esempio, probabilmente non sarebbe incostituzionale per il governo degli Stati Uniti condurre registrazioni massicce di tutte le conversazioni pubbliche (ad esempio, le cose che le persone si dicono all'angolo di una strada), e in precedenza sarebbe stato difficile analizzare questo volume di informazioni, ma con l'IA tutto potrebbe essere trascritto, interpretato e triangolato per creare un quadro dell'atteggiamento e della lealtà di molti o della maggior parte dei cittadini. Sosterrei una legislazione incentrata sulle libertà civili (o forse anche un emendamento costituzionale) che imponga guardrail più forti contro gli abusi alimentati dall'IA.

Gli altri due elementi — armi completamente autonome e IA per il processo decisionale strategico – sono linee più difficili da tracciaare poiché hanno usi legittimi nella difesa della democrazia, pur essendo inclini all'abuso. Qui penso che ciò che è giustificato sia un'estrema attenzione e controllo combinati con guardrail per prevenire abusi. La mia paura principale è avere un numero troppo esiguo di "dita

sul pulsante", tale che una o poche persone potrebbero essenzialmente gestire un esercito di droni senza aver bisogno che altri esseri umani collaborino per eseguire i loro ordini. Man mano che i sistemi di IA diventano più potenti, potremmo aver bisogno di avere meccanismi di supervisione più diretti e immediati per assicurarci che non vengano usati impropriamente, magari coinvolgendo rami del governo diversi dall'esecutivo. Penso che dovremmo approcciare le armi completamente autonome in particolare con grande cautela,[36] e non precipitarci nel loro uso senza adeguate salvaguardie.

In quarto luogo, dopo aver tracciato una linea netta contro gli abusi dell'IA nelle democrazie, dovremmo usare quel precedente per creare un tabù internazionale contro i peggiori abusi dell'IA potente. Riconosco che gli attuali venti politici si sono rivoltati contro la cooperazione internazionale e le norme internazionali, ma questo è un caso in cui ne abbiamo estremo bisogno. Il mondo deve capire il potenziale oscuro di un'IA potente nelle mani degli autocrati e riconoscere che certi usi dell'IA equivalgono a un tentativo di rubare permanentemente la loro libertà e imporre uno stato totalitario da cui non possono fuggire. Sosterrei persino che in alcuni casi, la sorveglianza su larga scala con IA potente, la propaganda di massa con IA potente e certi tipi di usi offensivi di armi completamente autonome dovrebbero essere considerati crimini contro l'umanità. Più in generale, è assolutamente necessaria una norma solida contro il totalitarismo abilitato dall'IA e tutti i suoi strumenti e mezzi.

È possibile avere una versione ancora più forte di questa posizione, ossia che poiché le possibilità del totalitarismo abilitato dall'IA sono così oscure, l'autocrazia non è semplicemente una forma di governo che le persone possono accettare nell'era dell'IA post-potente. Proprio come il feudalesimo divenne impraticabile con la rivoluzione industriale, l'era dell'IA potrebbe portare inevitabilmente e logicamente alla conclusione che la democrazia (e, si spera, una democrazia migliorata e rinvigorita dall'IA, come discuto in Machines of Loving Grace) sia l'unica forma di governo praticabile se l'umanità vuole avere un buon futuro.

Quinto e ultimo punto, le aziende di IA dovrebbero essere attentamente sorvegliate, così come la loro connessione con il governo, che è necessaria ma deve avere limiti e confini. La mole di capacità racchiusa in un'IA potente è tale che la normale governance aziendale progettata per proteggere gli azionisti e prevenire abusi ordinari come la frode è improbabile che ssia all'altezza del compito di governare le aziende di IA. Potrebbe anche esserci valore nel fatto che le aziende si impegnino pubblicamente (forse anche come parte della governance aziendale) a non intraprendere certe azioni, come costruire o accumulare privatamente hardware militare, usare grandi quantità di risorse di calcolo da parte di singoli individui in modi non responsabili, o usare i loro prodotti IA come propaganda per manipolare l'opinione pubblica a proprio favore.

Il pericolo qui viene da molte direzioni, e alcune direzioni sono in tensione con altre. L'unica costante è che dobbiamo cercare responsabilità, norme e guardrail per tutti, anche mentre diamo potere agli attori "buoni" per tenere sotto controllo gli attori "cattivi".

4. Player Piano

Sconvolgimento economico

Le tre sezioni precedenti riguardavano essenzialmente i rischi per la sicurezza posti dall'IA potente: rischi dall'IA stessa, rischi di uso improprio da parte di individui e piccole organizzazioni e rischi di uso improprio da parte di stati e grandi organizzazioni. Se mettiamo da parte i rischi per la sicurezza o assumiamo che siano stati risolti, la domanda successiva è economica. Quale sarà l'effetto di questa infusione di incredibile capitale "umano" sull'economia? Chiaramente, l'effetto più ovvio sarà quello di aumentare notevolmente la crescita economica. Il ritmo dei progressi nella ricerca scientifica, nell'innovazione biomedica, nella produzione, nelle catene di approvvigionamento, nell'efficienza del sistema finanziario e molto altro porterà quasi certamente a un tasso di crescita economica molto più rapido. In Machines of Loving Grace, suggerisco che potrebbe essere possibile un tasso di crescita del PIL annuo sostenuto del 10-20%.

Ma dovrebbe essere chiaro che questa è una spada a doppio taglio: quali sono le prospettive economiche per la maggior parte degli esseri umani esistenti in un mondo del genere? Le nuove tecnologie portano spesso shock nel mercato del lavoro e in passato gli esseri umani si sono sempre ripresi, ma temo che ciò sia accaduto perché questi shock precedenti interessavano solo una piccola frazione dell'intera gamma possibile di abilità umane, lasciando spazio agli esseri umani per espandersi verso nuovi compiti. L'IA avrà effetti molto più ampi e si verificheranno molto più velocemente, e quindi temo che sarà molto più impegnativo far sì che le cose vadano bene.

Sconvolgimento del mercato del lavoro

Ci sono due problemi specifici di cui mi preoccupo: lo spostamento nel mercato del lavoro e la concentrazione del potere economico. Partiamo dal primo. Si tratta di un argomento su cui ho avvertito molto pubblicamente nel 2025, prevedendo che l'IA potesse sostituire metà di tutti i lavori impiegatizi entry-level nei successivi 1-5 anni, pur accelerando la crescita economica e il progresso scientifico. Questo avvertimento ha dato inizio a un dibattito pubblico sul tema. Molti CEO, tecnologi ed economisti si sono detti d'accordo con me, ma altri hanno pensato che stessi cadendo preda della "fallacia della massa di lavoro" (lump of labor fallacy) e che non sapessi come funzionassero i mercati del lavoro, e alcuni non hanno colto l'intervallo di tempo 1-5 anni pensando che sostenessi che l'IA stia sostituendo i posti di lavoro proprio ora (cosa che concordo probabilmente non stia accadendo). Quindi vale la pena esaminare in dettaglio perché sono preoccupato per lo spostamento del lavoro, per chiarire questi malintesi.

Come base di partenza, è utile capire come i mercati del lavoro rispondano normalmente ai progressi tecnologici. Quando arriva una nuova tecnologia, inizia rendendo più efficienti pezzi di un determinato lavoro umano. Ad esempio, all'inizio della rivoluzione industriale, le macchine, come gli aratri potenziati, permisero ai contadini umani di essere più efficienti in alcuni aspetti del lavoro. Ciò migliorò la produttività degli agricoltori, il che aumentò i loro salari.

Nella fase successiva, alcune parti del lavoro agricolo potevano essere fatte interamente dalle macchine, ad esempio con l'invenzione della trebbiatrice o della seminatrice. In questa fase, gli umani svolgevano una frazione sempre minore del

lavoro, ma il lavoro che svolgevano diventava sempre più potenziato perché complementare a quello delle macchine, e la loro produttività continuava a salire. Come descritto dal paradosso di Jevons, i salari degli agricoltori e forse anche il numero di agricoltori continuarono ad aumentare. Anche quando il 90% del lavoro viene svolto dalle macchine, gli esseri umani possono semplicemente fare 10 volte di più del 10% che ancora svolgono, producendo 10 volte più output a parità di lavoro. Alla fine, le macchine fanno tutto o quasi tutto, come con le moderne mietitrebbie, i trattori e altre attrezzature. A questo punto l'agricoltura come forma di occupazione umana entra davvero in un declino ripido, e questo potenzialmente causa una seria perturbazione a breve termine, ma poiché l'agricoltura è solo una delle tante attività utili che gli esseri umani sono in grado di fare, le persone alla fine passano ad altri lavori, come l'uso delle macchine in fabbrica. Questo è vero anche se l'agricoltura rappresentava una percentuale enorme dell'occupazione ex ante. 250 anni fa, il 90% degli americani viveva nelle fattorie; in Europa, il 50-60% dell'occupazione era agricola. Ora quelle percentuali sono nell'ordine di una sola cifra in quei luoghi, perché i lavoratori sono passati ai lavori industriali (e più tardi ai lavori nel settore della conoscenza). L'economia può fare ciò che prima richiedeva la maggior parte della forza lavoro con solo l'1-2% di essa, liberando il resto della forza lavoro per costruire una società industriale sempre più avanzata. Non esiste una "massa di lavoro" fissa, ma solo una capacità sempre più ampia di fare sempre di più con sempre meno. I salari delle persone aumentano in linea con l'esponenziale del PIL e l'economia mantiene la piena occupazione una volta superate le perturbazioni a breve termine.

È possibile che le cose vadano più o meno allo stesso modo con l'IA, ma scommetterei piuttosto pesantemente contro. Ecco alcune ragioni per cui penso che l'IA sarà probabilmente diversa:

• Velocità. Il ritmo dei progressi nell'IA è molto più veloce rispetto alle precedenti rivoluzioni tecnologiche. Ad esempio, negli ultimi 2 anni, i modelli di IA sono passati dal riuscire a malapena a completare una singola riga di codice allo scrivere tutto o quasi tutto il codice per alcune persone – inclusi gli ingegneri di Anthropic.[37] Presto pottrebbero svolgere l'intero compito di un ingegnere informatico end-to-end.[38] È difficile per le persone adattarsi a questo ritmo di cambiamento, sia per i cambiamenti nel modo in cui funziona un dato lavoro sia per la necessità di passare a nuovi lavori. Persino programmatori leggendari si descrivono sempre più come "indietro". Il ritmo potrebbe semmai continuare ad accelerare, poiché i modelli di programmazione IA accelerano sempre più il compito stesso dello sviluppo dell'IA. Per essere chiari, la velocità in sé non significa che i mercati del lavoro e l'occupazione non si riprenderanno alla fine, implica solo che la transizione a breve termine sarà insolitamente dolorosa rispetto alle tecnologie passate, poiché gli esseri umani e i mercati del lavoro sono lenti a reagire e a equilibrarsi.

• Ampiezza cognitiva. Come suggerito dalla frase "paese di geni in un datacenter", l'IA sarà capace di una gamma molto vasta di abilità cognitive umane – forse tuttee. Questo è molto diverso dalle tecnologie precedenti come l'agricoltura meccanizzata, i trasporti o persino i computer.[39] Ciò renderà più difficile per le persone passare facilmente dai lavori sostituiti a lavori simili per i quali sarebbero adatte. Ad esempio,

le abilità intellettuali generali richieste per lavori entry-level in, diciamo, finanza, consulenza e legge sono abbastanza simili, anche se la conoscenza specifica è molto diversa. Una tecnologia che sconvolgesse solo uno di questi tre permetterebbe ai dipendenti di passare ai due sostituti stretti (o agli studenti universitari di cambiare facoltà). Ma sconvolgerli tutti e tre contemporaneamente (insieme a molti altri lavori simili) potrebbe essere più difficile da gestire per le persone. Inoltre, non è solo che la maggior parte dei lavori esistenti sarà sconvolta. Quella parte è già accaduta prima – si ricordi che ll'agricoltura era un'enorme percentuale dell'occupazione. Ma i contadini potevano passare al lavoro relativamente simile di operare macchine in fabbrica, anche se quel lavoro non era stato comune prima. Al contrario, l'IA sta sempre più eguagliando il profilo cognitivo generale degli esseri umani, il che significa che sarà brava anche nei nuovi lavori che verrebbero ordinariamente creati in risposta all'automazione dei vecchi. Un altro modo per dirlo è che l'IA non è un sostituto per specifici lavori umani, ma piuttosto un sostituto del lavoro generale per gli esseri umani.

• Segmentazione per abilità cognitiva. In un'ampia gamma di compiti, l'IA sembra avanzare dalla base della scala delle abilità verso l'alto. Ad esempio, nella programmazione i nostri modelli sono passati dal livello di "un programmatore mediocre" a "un programmatore forte" a "un programmatore molto forte".[40] Stiamo ora iniziando a vedere la stessa progressione nel lavoro impiegatizio in generale. Siamo quindi a rischio di una situazione in cui, invece di colpire persone con abilità specifiche o in professioni specifiche (che possono adattarsi riqualificandosi), l'IA colpisce persone con certe proprietà cognitive intrinseche, vale a dire una minore capacità intellettuale (che è più difficile da cambiare). Non è chiaro dove andranno queste persone o cosa faranno, e temo che potrebbero formare una "sottoclasse" disoccupata o con salari molto bassi. Per essere chiari, cose simili sono già accadute prima – ad esempio, i ccomputer e internet sono ritenuti da alcuni economisti come rappresentanti di un "cambiamento tecnologico orientato alle competenze" (skill-biased technological change). Ma questo orientamento alle competenze non è stato né così estremo come quello che mi aspetto di vedere con l'IA, né si ritiene abbia contribuito a un aumento della disuguaglianza salariale,[41] quindi non è esattamente un precedente rassicurante.

• Capacità di colmare le lacune. Il modo in cui i lavori umani spesso si adattano di fronte alla nuova tecnologia è che ci sono molti aspetti nel lavoro, e la nuova tecnologia, anche se sembra sostituire direttamente gli umani, ha spesso delle lacune. Se qualcuno inventa una macchina per fabbricare gadget, gli umani potrebbero dover comunque caricare la materia prima nella macchina. Anche se ciò richiede solo l'1% dello sforzo rispetto alla produzione manuale dei gadget, i lavoratori umani possono semplicemente produrre 100 volte più gadget. Ma l'IA, oltre ad essere una tecnologia che avanza rapidamente, è anche una tecnologia che si adatta rapidamente. Durante ogni rilascio di modello, le aziende di IA misurano attentamente in cosa il modello è bravo e in cosa non lo è, e anche i clienti forniscono tali informazioni dopo il lancio. Le debolezze possono essere affrontate raccogliendo compiti che incarnano la lacuna attuale e addestrando su di essi per il modello successivo. All'inizio dell'IA generativa, gli utenti notarono che i sistemi di IA avevano certe debolezze (come i

modelli di immagini IA che generavano mani con il numero sbagliato di dita) e molti assunsero che queste debolezze fossero intrinseche alla tecnologia. Se lo fossero, limiterebbero lo sconvolgimento dei posti di lavoro. Ma praticamente ogni debolezza del genere viene affrontata rapidamente spesso, enntro pochi mesi.
Vale la pena affrontare i punti comuni di scetticismo.
In primo luogo, c'è l'argomento secondo cui la diffusione economica sarà lenta, tale che anche se la tecnologia sottostante fosse capace di svolgere la maggior parte del lavoro umano, l'applicazione effettiva di essa in tutta l'economia potrebbe essere molto più lenta (ad esempio in settori lontani dall'industria dell'IA e lenti ad adottarla). La lenta diffusione della tecnologia è assolutamente reale parllo con persone di un'ampia varietà di imprese e ci sono settori in cui l'adozione dell'IA richiederà anni. Ecco perché la mia previsione del 50% di posti di lavoro impiegatizi entry-level sconvolti è di 1-5 anni, anche se sospetto che avremo un'IA potente (che sarebbe, tecnologicamente parlando, sufficiente a svolgere la maggior parte o tutti i lavori, non solo quelli entry-level) in molto meno di 5 anni. Ma gli effetti di diffusione ci fanno semplicemente guadagnare tempo. E non sono fiducioso che saranno così lenti come si prevede. L'adozione dell'IA nelle imprese sta crescendo a tassi molto più rapidi rispetto a qualsiasi tecnologia precedente, in gran parte per la pura forza della tecnologia stessa. Inoltre, anche se le imprese tradizionali fossero lente ad adottare nuove tecnologie, nasceranno startup per servire da "collante" e rendere l'adozione più facile. Se ciò non funzionasse, le startup potrebbero semplicemente scardinare direttamente gli operatori storici.
Ciò potrebbe portare a un mondo in cui non sono tanto i singoli lavori a essere sconvolti, quanto le grandi imprese in generale a essere scardinate e sostituite da startup molto meno intensive in termini di manodopera. Ciò potrebbe anche portare a un mondo di "disuguaglianza geografica", dove una frazione crescente della ricchezza mondiale si concentra nella Silicon Valley, che diventa una propria economia che corre a una velocità diversa dal resto del mondo, lasciandolo indietro. Tutti questi esiti sarebbero ottimi per la crescita economica ma non così positivi per il mercato ddel lavoro o per coloro che vengono lasciati indietro.
In secondo luogo, alcuni dicono che i lavori umani si sposteranno nel mondo fisico, il che evita l'intera categoria del "lavoro cognitivo" in cui l'IA sta progredendo così rapidamente. Non sono sicuro di quanto sia sicuro questo approccio. Gran parte del lavoro fisico viene già svolto dalle macchine (ad es. la produzione industriale) o lo sarà presto (ad es. la guida). Inoltre, un'IA sufficientemente potente sarà in grado di accelerare lo sviluppo dei robot e poi controllarli nel mondo fisico. Potrebbe far guadagnare un po' di tempo (che è una buona cosa), ma temo che non ne farà guadagnare molto. E anche se lo sconvolgimento fosse limitato ai soli compiti cognitivi, sarebbe comunque una perturbazione senza precedenti per dimensioni e rapidità.
In terzo luogo, forse alcuni compiti richiedono intrinsecamente o beneficiano enormemente di un tocco umano. Sono un po' più incerto su questo punto, ma rimango scettico che sarà sufficiente a compensare la massa degli impatti che ho descritto sopra. L'IA è già ampiamente utilizzata per il servizio clienti. Molte persone riferiscono che è più facile parlare con l'IA dei propri problemi personali che parlare

con un terapeuta che l'IA è più pazientte. Quando mia sorella stava lottando con problemi medici durante una gravidanza, sentiva di non ottenere le risposte o il supporto di cui aveva bisogno dai suoi fornitori di assistenza, e ha trovato Claude dotato di una migliore empatia (oltre a riuscire meglio a diagnosticare il problema). Sono sicuro che ci siano alcuni compiti per i quali un tocco umano sia davvero importante, ma non sono sicuro di quanti e qui stiamo parlando di trovare lavvoro per quasi tutti nel mercato del lavoro.

In quarto luogo, alcuni potrebbero sostenere che il vantaggio comparato proteggerà comunque gli esseri umani. Secondo la legge del vantaggio comparato, anche se l'IA fosse migliore degli umani in tutto, qualsiasi differenza relativa tra il profilo di abilità umano e quello dell'IA crea una base di scambio e specializzazione tra umani e IA. Il problema è che se le IA sono letteralmente migliaia di volte più produttive degli umani, questa logica inizia a vacillare. Anche minimi costi di transazione potrebbero rendere non conveniente per l'IA scambiare con gli umani. E i salari umani potrebbero essere molto bassi, anche se tecnicamente avessero qualcosa da offrire.

È possibile che tutti questi fattori possano essere affrontati che il mercato del lavoro sia abbastanza resiliiente da adattarsi anche a uno sconvolgimento così enorme. Ma anche se potesse alla fine adattarsi, i fattori sopra descritti suggeriscono che lo shock a breve termine sarà di entità senza precedenti.

Difese

Cosa possiamo fare per questo problema? Ho diversi suggerimenti, alcuni dei quali Anthropic sta già mettendo in pratica.

La prima cosa è semplicemente ottenere dati accurati su ciò che sta accadendo con lo spostamento del lavoro in tempo reale. Quando un cambiamento economico avviene molto velocemente, è difficile ottenere dati affidabili su ciò che sta accadendo e, senza dati affidabili, è difficile progettare politiche efficaci. Ad esempio, i dati governativi attualmente mancano di dati granulari e ad alta frequenza sull'adozione dell'IA tra aziende e settori. Nell'ultimo anno Anthropic ha gestito e rilasciato pubblicamente un Indice Economico che mostra l'uso dei nostri modelli quasi in tempo reale, suddiviso per settore, compito, posizione e persino fattori come se un compito venisse automatizzato o condotto in modo collaborativo. Abbiamo anche un Consiglio Consultivo Economico per aiutarci a interpretare questi dati e vedere cosa sta arrivando.

In secondo luogo, le aziende di IA hanno una scelta nel modo in cui lavorano con le imprese. La stessa inefficienza delle imprese tradizionali significa che il loro dispiegamento dell'IA può dipendere molto dal percorso scelto (path dependent), e c'è spazio per scegliere un percorso migliore. Le imprese hanno spesso una scelta tra "risparmio sui costi" (fare la stessa cosa con meno persone) e "innovazione" (fare di più con lo stesso numero di persone). Il mercato produrrà inevitabilmente entrambi alla fine, e qualsiasi azienda di IA competitiva dovrà servire un po' di entrambi, ma potrebbe esserci spazio per orientare le aziende verso l'innovazione quando possibile, e questo potrebbe farci guadagnare un po' di tempo. Anthropic sta riflettendo attivamente su questo.

In terzo luogo, le aziende dovrebbero pensare a come prendersi cura dei propri dipendenti. A breve termine, essere creativi su modi per riassegnare i dipendenti

all'interno delle aziende può essere un modo promettente per scongiurare la necessità di licenziamenti. A lungo termine, in un mondo con un'enorme ricchezza totale, in cui molte aziende aumentano notevolmente di valore grazie alla maggiore produttività e alla concentrazione del capitale, potrebbe essere fattibile pagare i dipendenti umani anche molto tempo dopo che non forniscono più valore economico in senso tradizionale. Anthropic sta attualmente considerando una gamma di possibili percorsi per i nostri dipendenti che condivideremo nel prossimo futuro.

In quarto luogo, le persone facoltose hanno l'obbligo di aiutare a risolvere questo problema. Mi rattrista che molti individui facoltosi (specialmente nel settore tecnologico) abbiano recentemente adottato un atteggiamento cinico e nichilista secondo cui la filantropia sia inevitabilmente fraudolenta o inutile. Sia la filantropia privata come la Gates Foundation che i programmi pubblici come PEPFAR hanno salvato decine di milioni di vite nel mondo in via di sviluppo e hanno aiutato a creare opportunità economiche nel mondo sviluppato. Tutti i co-fondatori di Anthropic si sono impegnati a donare l'80% della propria ricchezza, e lo staff di Anthropic si è impegnato individualmente a donare azioni dell'azienda per un valore di miliardi ai prezzi correnti donazioni che l'azienda si è impegnata a raddopppiare.

In quinto luogo, mentre tutte le azioni private sopra citate possono essere utili, alla fine un problema macroeconomico di queste dimensioni richiederà l'intervento del governo. La naturale risposta politica a un'enorme torta economica accoppiata a un'elevata disuguaglianza (dovuta alla mancanza di posti di lavoro, o a posti di lavoro mal pagati, per molti) è la tassazione progressiva. La tassa potrebbe essere generale o potrebbe essere mirata contro le aziende di IA in particolare. Ovviamente la progettazione fiscale è complicata e ci sono molti modi per farla andare male. Non sostengo politiche fiscali mal progettate. Penso che i livelli estremi di disuguaglianza previsti in questo saggio giustifichino una politica fiscale più robusta su basi morali fondamentali, ma posso anche addurre un argomento pragmatico ai miliardari del mondo dicendo che è nel loro interesse sostenere una buona versione di essa: se non ne sosterranno una buona versione, ne subiranno inevitabilmente una cattiva progettata da una folla inferocita.

In definitiva, considero tutti gli interventi di cui sopra come modi per guadagnare tempo. Alla fine l'IA sarà in grado di fare tutto, e dobbiamo fare i conti con questo. È mia speranza che per allora potremo usare l'IA stessa per aiutarci a ristrutturare i mercati in modi che funzionino per tutti, e che gli interventi sopra citati possano farci superare il periodo di transizione.

Concentrazione economica del potere

Separatamente dal problema dello spostamento del lavoro o della disuguaglianza economica in sé, c'è il problema della concentrazione del potere economico. La Sezione 1 ha discusso il rischio che l'umanità venga privata del potere dall'IA, e la Sezione 3 ha discusso il rischio che i cittadini vengano privati del potere dai loro governi con la forza o la coercizione. Ma un altro tipo di depotenziamento può verificarsi se c'è una tale enorme concentrazione di ricchezza che un piccolo gruppo di persone controlla efficacemente la politica del governo con la propria influenza, e i cittadini comuni non hanno influenza perché mancano di leva economica. La democrazia è in ultima analisi sostenuta dall'idea che la popolazione nel suo insieme

sia necessaria per il funzionamento dell'economia. Se quella leva economica scompare, allora il contratto sociale implicito della democrazia potrebbe smettere di funzionare. Altri hanno scritto su questo, quindi non c'è bisogno che entri in grandi dettagli qui, ma concordo con la preoccupazione e temo che stia già iniziando ad accadere.

Per essere chiari, non sono contrario al fatto che le persone facciano molti soldi. C'è un forte argomento secondo cui ciò incentiva la crescita economica in condizioni normali. Sono comprensivo verso le preoccupazioni di ostacolare l'innovazione uccidendo la gallina dalle uova d'oro che la genera. Ma in uno scenario in cui la crescita del PIL è del 10-20% l'anno e l'IA sta rapidamente prendendo il controllo dell'economia, mentre singoli individui detengono frazioni apprezzabili del PIL, l'innovazione non è la cosa di cui preoccuparsi. La cosa di cui preoccuparsi è un livello di concentrazione della ricchezza che spaccherà la società.

L'esempio più famoso di estrema concentrazione di ricchezza nella storia degli Stati Uniti è la Gilded Age, e l'industriale più ricco della Gilded Age fu John D. Rockefeller. La ricchezza di Rockefeller ammontava a circa il 2% del PIL statunitense dell'epoca.[42] Una frazione simile oggi porterebbe a una fortuna di 600 miliardi di dollari, e la persona più ricca del mondo oggi (Elon Musk) già supera tale cifra, attestandosi a circa 700 miliardi di dollari. Quindi siamo già a livelli storicamente senza precedenti di concentrazione della ricchezza, ancora prima della maggior parte dell'impatto economico dell'IA. Non credo sia troppo azzardato (se otteniamo un "paese di geni") immaginare che le aziende di IA, le aziende di semiconduttori e forse le aziende di applicazioni a valle generino circa 3 trilioni di dollari di entrate all'anno,[43] vengano valutate circa 30 trilioni di dollari e portino a fortune personali ben oltre i trilioni. In quel mondo, i dibattiti che abbiamo oggi sulla politica fiscale semplicemente non si applicheranno, poiché saremo in una situazione fondamentalmente diversa.

Legato a questo, l'accoppiamento di questa concentrazione economica di ricchezza con il sistema politico già mi preoccupa. I datacenter IA rappresentano già una frazione sostanziale della crescita economica degli Stati Uniti[44] e stanno quindi legando strettamente gli interessi finanziari delle grandi aziende tecnologiche (che sono sempre più focalizzate sull'IA o sulle infrastrutture IA) e gli interessi politici del governo in un modo che può produrre incentivi perversi. Vediamo già questo attraverso la riluttanza delle aziende tecnologiche a criticare il governo degli Stati Uniti e il sostegno del governo a politiche anti-regolatorie estreme sull'IA.

Difese

Cosa si può fare al riguardo?

In primo luogo, e più ovviamente, le aziende dovrebbero semplicemente scegliere di non farne parte. Anthropic si è sempre sforzata di essere un attore politico (policy actor) e non un attore schierato (political actor), e di mantenere le proprie visioni autentiche qualunque sia l'amministrazione. Abbiamo parlato a favore di una regolamentazione sensata dell'IA e di controlli sulle esportazioni che siano nel pubblico interesse, anche quando questi sono in contrasto con la politica governativa.[45] Molte persone mi hanno detto che dovremmo smettere di farlo, che potrebbe portare a un trattamento sfavorevole, ma nell'anno in cui lo abbiamo fatto, la

valutazione di Anthropic è aumentata di oltre 6 volte, un balzo quasi senza precedenti alla nostra scala commerciale.
In secondo luogo, l'industria dell'IA ha bisogno di un rapporto più sano con il governo un rapporto basato sull'impegno politico sostanziale piuttosto che sull'allineamento partitico. La nostra scelta di impegnarci sulla sostanza delle politiche piuttosto che sulla politica è talvolta interpretata come un errore tattico o un fallimento nel "leggere la situazione" piuttosto che come una decisione di principio, e quel tipo di inquadramento mi preoccupa. In una democrazia sana, le aziende dovrebbero essere in grado di sostenere una buona politica per se stessa. Legato a questo, si sta preparando un contraccolpo pubblico contro l'IA: questo potrebbe essere correttivo, ma al momento è privo di focus. Gran parte di esso prende di mira questioni che non sono effettivamente problemi (come l'uso dell'acqua nei datacenter) e propone soluzioni (come i divieti sui datacenter o le tasse sulla ricchezza mal progettate) che non affronterebbero le reali preoccupazioni. Il problema di fondo che merita attenzione è garantire che lo sviluppo dell'IA rimanga responsabile nei confronti del pubblico interesse, non catturato da alcuna particolare alleanza politica o commerciale, e sembra importante focalizzare lì la discussione pubblica.
In terzo luogo, gli interventi macroeconomici che ho descritto in precedenza in questa sezione, così come una rinascita della filantropia privata, possono aiutare a bilanciare la bilancia economica, affrontando contemporaneamente sia lo spostamento del lavoro che i problemi di concentrazione del potere economico. Dovremmo guardare alla storia del nostro paese qui: anche nella Gilded Age, industriali come Rockefeller e Carnegie sentirono un forte obbligo verso la società in generale, un sentimento che la società avesse contribuito enormemente al loro successo e che dovessero restituire. Quello spirito sembra essere sempre più mancante oggi, e penso che sia una grande parte della via d'uscita da questo dilemma economico. Coloro che sono all'avanguardia del boom economico dell'IA dovrebbero essere disposti a cedere sia la propria ricchezza che il proprio potere.

5. Mari neri dell'infinito

Effetti indiretti

Quest'ultima sezione è un calderone per le incognite sconosciute, in particolare le cose che potrebbero andare storte come risultato indiretto dei progressi positivi dell'IA e della conseguente accelerazione della scienza e della tecnologia in generale. Supponiamo di affrontare tutti i rischi descritti finora e di iniziare a raccogliere i benefici dell'IA. Otterremo probabilmente un "secolo di progresso scientifico ed economico compresso in un decennio", e questo sarà estremamente positivo per il mondo, ma dovremo poi fare i conti con i problemi che derivano da questo rapido tasso di progresso, e quei problemi potrebbero arrivarci addosso velocemente. Potremmo anche incontrare altri rischi che si verificano indirettamente come conseguenza del progresso dell'IA e che sono difficili da anticipare in anticipo.
Per la natura stessa delle incognite sconosciute è impossibile fare un elenco esaustivo, ma elencherò tre possibili preoccupazioni come esempi illustrativi di ciò che dovremmo monitorare:
• Progressi rapidi nella biologia. Se otterremo un secolo di progresso medico in pochi

anni, è possibile che aumenteremo notevolmente la durata della vita umana, e c'è la possibilità di acquisire capacità radicali come l'abilità di aumentare l'intelligenza umana o modificare radicalmente la biologia umana. Questi sarebbero grandi cambiamenti in ciò che è possibile, accadendo molto velocemente. Potrebbero essere positivi se fatti responsabilmente (che è la mia speranza, come descritto in Machines of Loving Grace), ma c'è sempre il rischio che vadano molto male ﬠ ad esempio, se gli sforzi per rendere gli umani più intelligenti li rendessero anche più instabili o bramosi di potere. C'è anche la questione degli "upload" o "emulazione dell'intero cervello", menti umane digitali istanziate in un software, che potrebbero un giorno aiutare l'umanità a trascendere i suoi limiti fisici, ma che portano anche rischi che trovo inquietanti.

• L'IA cambia la vita umana in modo malsano. Un mondo con miliardi di intelligenze molto più intelligenti degli umani in tutto sarà un mondo molto strano in cui vivere. Anche se l'IA non mirasse attivamente ad attaccare gli umani (Sezione 1) e non venisse esplicitamente usata per l'oppressione o il controllo dagli stati (Sezione 3), c'è molto che potrebbe andare storto al di sotto di questo, tramite normali incentivi commerciali e transazioni nominalmente consensuali. Vediamo i primi accenni di questo nelle preoccupazioni sulla psicosi da IA, sull'IA che spinge le persone al suicidio e nelle preoccupazioni sulle relazioni romantiche con le IA. Ad esempio, le IA potenti potrebbero inventare qualche nuova religione e convertirvi milioni di persone? La maggior parte delle persone potrebbe finire "dipendente" in qualche modo dalle interazioni con l'IA? Le persone potrebbero finire per essere "marionette" dei sistemi di IA, dove un'IA osserva essenzialmente ogni loro mossa e dice loro esattamente cosa fare e dire in ogni momento, portando a una vita "buona" ma priva di libertà o di qualsiasi orgoglio per i risultati raggiunti? Non sarebbe difficile generare dozzine di questi scenari se mi sedessi con il creatore di Black Mirror e cercassi di fare un brainstorming. Penso che questo punti all'importanza di cose come il miglioramento della Costituzione di Claude, oltre a quanto necessario per prevenire i problemi della Sezione 1. Assicurarsi che i modelli di IA abbiano davvero a cuore gli interessi a lungo termine dei loro utenti, in un modo che le persone riflessive sottoscriverebbero piuttosto che in qualche modo sottilmente distorto, sembra fondamentale.

• Scopo umano. Questo è legato al punto precedente, ma non riguarda tanto le specifiche interazioni umane con i sistemi di IA quanto il modo in cui la vita umana cambia in generale in un mondo con un'IA potente. Gli esseri umani saranno in grado di trovare scopo e significato in un mondo del genere? Penso che questa sia una questione di atteggiamento: come ho detto in Machines of Loving Grace, penso che lo scopo umano non dipenda dall'essere i migliori al mondo in qualcosa, e gli esseri umani possano trovare uno scopo anche su lunghi periodi di tempo attraverso storie e progetti che amano. Dobbiamo semplicemente rompere il legame tra la generazione di valore economico e l'autostima e il significato. Ma questa è una transizione che la società deve compiere, e c'è sempre il rischio di non gestirla bene.

La mia speranza con tutti questi potenziali problemi è che in un mondo con un'IA potente di cui ci fidiamo che non ci uccida, che non sia lo strumento di un governo oppressivo e che stia lavorando genuinamente per nostro conto, potremo usare l'IA

stessa per anticipare e prevenire questi problemi. Ma non è garantito come tutti gli altri rischi, è quaalcosa che dobbiamo gestire con cura.

La prova dell'umanità

Leggere questo saggio può dare l'impressione che ci troviamo in una situazione scoraggiante. Certamente ho trovato scoraggiante scriverlo, in contrasto con Machines of Loving Grace, che mi faceva sentire come se stessi dando forma e struttura a una musica supremamente bella che risuonava nella mia testa da anni. E c'è molto nella situazione attuale che è genuinamente difficile. L'IA porta minacce all'umanità da più direzioni, e c'è una tensione reale tra i diversi pericoli, dove mitigare alcuni di essi rischia di peggiorarne altri se non si procede con estrema cautela.

Prendere tempo per costruire con cura i sistemi di IA affinché non minaccino autonomamente l'umanità è in tensione reale con la necessità per le nazioni democratiche di rimanere avanti alle nazioni autoritarie e non essere sottomesse da esse. Ma d'altra parte, gli stessi strumenti abilitati dall'IA necessari per combattere le autocrazie possono, se portati troppo oltre, essere rivolti all'interno per creare tirannia nei nostri stessi paesi. Il terrorismo guidato dall'IA potrebbe uccidere milioni attraverso l'uso improprio della biologia, ma una reazione eccessiva a questo rischio potrebbe portarci sulla strada di uno stato di sorveglianza autocratico. Gli effetti sul lavoro e sulla concentrazione economica dell'IA, oltre a essere gravi problemi di per sé, potrebbero costringerci ad affrontare gli altri problemi in un ambiente di rabbia pubblica e forse anche disordini civili, piuttosto che poter fare appello alla parte migliore della nostra natura. Soprattutto, il numero elevatissimo di rischi, compresi quelli sconosciuti, e la necessità di affrontarli tutti insieme, crea un percorso intimidatorio che l'umanità deve percorrere.

Inoltre, gli ultimi anni dovrebbero chiarire che l'idea di fermare o anche solo rallentare sostanzialmente la tecnologia è fondamentalmente insostenibile. La formula per costruire sistemi di IA potenti è incredibilmente semplice, tanto che si può quasi dire emerga spontaneamente dalla giusta combinazione di dati e calcolo grezzo. La sua creazione è stata probabilmente inevitabile nell'istante in cui l'umanità ha inventato il transistor, o probabilmente ancora prima quando abbiamo imparato a controllare il fuoco. Se un'azienda non la costruisce, altre lo faranno quasi altrettanto velocemente. Se tutte le aziende dei paesi democratici fermassero o rallentassero lo sviluppo, per accordo reciproco o decreto normativo, i paesi autoritari semplicemente continuerebbero. Dato l'incredibile valore economico e militare della tecnologia, insieme alla mancanza di qualsiasi meccanismo di applicazione significativo, non vedo come potremmo convincerli a fermarsi.

Vedo un percorso verso una leggera moderazione nello sviluppo dell'IA che sia compatibile con una visione realista della geopolitica. Quel percorso prevede di rallentare la marcia delle autocrazie verso l'IA potente per alcuni anni negando loro le risorse di cui hanno bisogno per costruirla,[46] ossia chip e attrezzature per la produzione di semiconduttori. Questo a sua volta dà ai paesi democratici un cuscinetto che possono "spendere" per costruire un'IA potente con più attenzione, con più riguardo ai suoi rischi, pur procedendo abbastanza velocemente da battere

comodamente le autocrazie. La corsa tra le aziende di IA all'interno delle democrazie può quindi essere gestita sotto l'ombrello di un quadro giuridico comune, tramite un mix di standard industriali e regolamentazione.
Anthropic ha sostenuto con forza questo percorso, spingendo per i controlli sulle esportazioni di chip e una regolamentazione giudiziosa dell'IA, ma anche queste proposte apparentemente dettate dal buon senso sono state in gran parte respinte dai responsabili politici negli Stati Uniti (che è il paese dove è più importante averle). Ci sono così tanti soldi da fare con l'IA – letteralmente trilionii di dollari l'anno – che anche le misure più semplici trovano difficile superare l'economia politica inerente all'IA. Questa è la trappola: l'IA è così potente, un premio così sfolgorante, che è difficilissimo per la civiltà umana imporre qualsivoglia restrizione su di essa.
Posso immaginare, come ha fatto Sagan in Contact, che questa stessa storia si svolga su migliaia di mondi. Una specie acquisisce senzienza, impara a usare strumenti, inizia l'ascesa esponenziale della tecnologia, affronta le crisi dell'industrializzazione e delle armi nucleari e, se sopravvive a quelle, si confronta con la sfida più dura e finale quando impara a modellare la sabbia in macchine che pensano. Se sopravviveremo a quel test e andremo avanti per costruire la bellissima società descritta in Machines of Loving Grace, o soccomberemo alla schiavitù e alla distruzione, dipenderà dal nostro carattere e dalla nostra determinazione come specie, dal nostro spirito e dalla nostra anima.
Nonostante i molti ostacoli, credo che l'umanità abbia la forza interiore per superare questo test. Sono incoraggiato e ispirato dai migliaia di ricercatori che hanno dedicato la loro carriera ad aiutarci a capire e guidare i modelli di IA, e a plasmare il carattere e la costituzione di questi modelli. Penso che ci sia ora una buona possibilità che quegli sforzi diano frutti in tempo utile. Sono incoraggiato dal fatto che almeno alcune aziende abbiano dichiarato che pagheranno costi commerciali significativi per impedire ai loro modelli di contribuire alla minaccia del bioterrorismo. Sono incoraggiato dal fatto che alcune persone coraggiose abbiano resistito ai venti politici prevalenti e approvato legislazioni che pongono i primi semi di guardrail sensati sui sistemi di IA. Sono incoraggiato dal fatto che il pubblico capisca che l'IA comporta dei rischi e voglia che questi rischi vengano affrontati. Sono incoraggiato dallo spirito indomabile di libertà in tutto il mondo e dalla determinazione a resistere alla tirannia ovunque essa si verifichi.
Ma dovremo intensificare i nostri sforzi se vogliamo riuscire. Il primo passo è che coloro che sono più vicini alla tecnologia dicano semplicemente la verità sulla situazione in cui si trova l'umanità, cosa che ho sempre cercato di fare; lo sto facendo in modo più esplicito e con maggiore urgenza con questo saggio. Il passo successivo sarà convincere i pensatori del mondo, i responsabili politici, le aziende e i cittadini dell'imminenza e dell'importanza capitale di questo problema – che vale la pena spendere pensiero e capitalee politico su questo rispetto alle migliaia di altri problemi che dominano le notizie ogni giorno. Poi verrà il tempo del coraggio, affinché un numero sufficiente di persone si opponga alle tendenze prevalenti e rimanga fedele ai propri principi, anche di fronte a minacce ai propri interessi economici e alla propria sicurezza personale.
Gli anni davanti a noi saranno incredibilmente duri, chiedendoci più di quanto

pensiamo di poter dare. Ma nel mio tempo come ricercatore, leader e cittadino, ho visto abbastanza coraggio e nobiltà da credere che possiamo vincere che quando messa nelle circostanze più oscure, l'umanità ha un modo di raccogliere, apparentemente all'ultimo minuto, la forza e la saggezza necessarie per prevalere. Non abbiamo tempo da perdere.

Vorrei ringraziare Erik Brynjolfsson, Ben Buchanan, Mariano-Florentino Cuéllar, Allan Dafoe, Kevin Esvelt, Nick Beckstead, Richard Fontaine, Jim McClave e moltissimi membri dello staff di Anthropic per i loro utili commenti alle bozze di questo saggio.

Note a piè di pagina

1. Questo è simmetrico a un punto che ho fatto in Machines of Loving Grace, dove ho iniziato dicendo che i vantaggi dell'IA non dovrebbero essere pensati in termini di profezia di salvezza, e che è importante essere concreti e radicati ed evitare la grandiosità. In definitiva, le profezie di salvezza e le profezie di sventura sono inutili per affrontare il mondo reale, fondamentalmente per le stesse ragioni.
2. L'obiettivo di Anthropic è rimanere coerente attraverso tali cambiamenti. Quando parlare dei rischi dell'IA era politicamente popolare, Anthropic ha cautamente sostenuto un approccio giudizioso e basato sulle prove a questi rischi. Ora che parlare dei rischi dell'IA è politicamente impopolare, Anthropic continua a sostenere cautamente un approccio giudizioso e basato sulle prove a questi rischi.
3. Nel tempo, ho acquisito una crescente fiducia nella traiettoria dell'IA e nella probabilità che supererà l'abilità umana su tutta la linea, ma rimane ancora un po' di incertezza.
4. I controlli sulle esportazioni di chip sono un ottimo esempio di questo. Sono semplici e sembrano per lo più funzionare.
5. E naturalmente, la caccia a tali prove deve essere intellettualmente onesta, tale da poter anche far emergere prove di una mancanza di pericolo. La trasparenza attraverso le model card e altre divulgazioni è un tentativo di tale sforzo intellettualmente onesto.
6. Infatti, da quando ho scritto Machines of Loving Grace nel 2024, i sistemi di IA sono diventati capaci di svolgere compiti che richiedono agli umani diverse ore, con METR che recentemente ha valutato che Opus 4.5 può svolgere circa quattro ore di lavoro umano con una affidabilità del 50%.
7. E per essere chiari, anche se l'IA potente fosse a soli 1-2 anni di distanza in senso tecnico, molte delle sue conseguenze sociali, sia positive che negative, potrebbero richiedere alcuni anni in più per verificarsi. Ecco perché posso contemporaneamente pensare che l'IA sconvolgerà il 50% dei lavori impiegatizi entry-level nell'arco di 1-5 anni, pur pensando che potremmo avere un'IA che è più capace di chiunque in soli 1-2 anni.
8. Vale la pena aggiungere che il pubblico (rispetto ai responsabili politici) sembra essere molto preoccupato per i rischi dell'IA. Penso che parte della loro attenzione sia corretta (cioè lo spostamento del lavoro dovuto all'IA), e parte sia fuori luogo (come le preoccupazioni sull'uso dell'acqua dell'IA, che non è significativo). Questo

contraccolpo mi dà speranza che un consenso sulla gestione dei rischi sia possibile, ma finora non è stato ancora tradotto in cambiamenti politici, per non parlare di cambiamenti politici efficaci o ben mirati.
9. Possono anche, naturalmente, manipolare (o semplicemente pagare) un gran numero di umani affinché facciano ciò che vogliono nel mondo fisico.
10. Non credo che questo sia un uomo di paglia: è mia intesa, ad esempio, che Yann LeCun sostenga questa posizione.
11. Ad esempio, si veda la Sezione 5.5.2 (p. 63–66) della Claude 4 system card.
12. Ci sono anche una serie di altri presupposti inerenti al modello semplice, di cui non parlerò qui. In generale, dovrebbero renderci meno preoccupati per la specifica storia semplice della ricerca di potere disallineata, ma anche più preoccupati per possibili comportamenti imprevedibili che non abbiamo anticipato.
13. Ender's Game descrive una versione di questo che coinvolge esseri umani invece dell'IA.
14. Ad esempio, ai modelli può essere detto di non fare varie cose cattive, e anche di obbedire agli umani, ma potrebbero poi osservare che molti umani fanno esattamente quelle cose cattive! Non è chiaro come questa contraddizione si risolverebbe (e una costituzione ben progettata dovrebbe incoraggiare il modello a gestire queste contraddizioni con garbo), ma questo tipo di dilemma non è così diverso dalle situazioni cosiddette "artificiali" in cui mettiamo i modelli di IA durante i test.
15. Incidentalmente, una conseguenza del fatto che la costituzione sia un documento in linguaggio naturale è che è leggibile per il mondo, e ciò significa che può essere criticata da chiunque e confrontata con documenti simili di altre aziende. Sarebbe prezioso creare una gara verso l'alto che non solo incoraggi le aziende a rilasciare questi documenti, ma le incoraggi a farli bene.
16. Esiste persino un'ipotesi su un profondo principio unificatore che collega l'approccio basato sul carattere dell'IA Costituzionale ai risultati della scienza dell'interpretabilità e dell'allineamento. Secondo l'ipotesi, i meccanismi fondamentali che guidano Claude sono nati originariamente come modi per simulare i personaggi nel pre-addestramento, come prevedere cosa direbbero i personaggi in un romanzo. Ciò suggerirebbe che un modo utile di pensare alla costituzione sia più simile a una descrizione del personaggio che il modello usa per istanziare una personalità coerente. Ci aiuterebbe anche a spiegare i risultati del tipo "Devo essere una cattiva persona" che ho menzionato sopra (perché il modello sta cercando di agire come se fosse un personaggio coerente – in questo casso uno cattivo), e suggerirebbe che i metodi di interpretabilità dovrebbero essere in grado di scoprire "tratti psicologici" all'interno dei modelli. I nostri ricercatori stanno lavorando su modi per testare questa ipotesi.
17. Per essere chiari, il monitoraggio viene fatto in un modo che preserva la privacy.
18. Anche nei nostri esperimenti con quelle che sono essenzialmente regole imposte volontariamente con la nostra Responsible Scaling Policy, abbiamo scoperto più e più volte che è molto facile finire per essere troppo rigidi, tracciando linee che sembrano importanti ex ante ma che si rivelano sciocche a posteriori. È semplicemente molto facile stabilire regole sulle cose sbagliate quando una tecnologia avanza rapidamente.
19. SB 53 e RAISE non si applicano affatto alle aziende con meno di 500 milioni di

dollari di fatturato annuo. Si applicano solo alle aziende più grandi e consolidate come Anthropic.
20. Ho letto originariamente il saggio di Joy 25 anni fa, quando fu scritto, ed ebbe un profondo impatto su di me. Allora come oggi, lo considero troppo pessimista non credo che la "rinuncia" generalizzatta a intere aree tecnologiche, suggerita da Joy, sia la risposta ma le questioni che solleva erano sorprendenntemente preveggenti, e Joy scrive anche con un profondo senso di compassione e umanità che ammiro.
21. Dobbiamo preoccuparci degli attori statali, ora e in futuro, e ne discuto nella prossima sezione.
22. Ci sono prove che molti terroristi siano almeno relativamente ben istruiti, il che potrebbe sembrare in contraddizione con ciò che sto sostenendo qui circa una correlazione negativa tra capacità e motivazione. Ma penso che in realtà siano osservazioni compatibili: se la soglia di capacità per un attacco riuscito è alta, allora quasi per definizione coloro che attualmente hanno successo devono avere un'alta capacità, anche se capacità e motivazione sono correlate negativamente. Ma in un mondo in cui i limiti alle capacità fossero rimossi (ad esempio, con i futuri LLM), prevedrei che una popolazione sostanziale di persone con la motivazione a uccidere ma minore capacità inizierebbe a farlo proprio come vediamo per criminni che non richiedono molta capacità (come le sparatorie nelle scuole).
23. Aum Shinrikyo ci provò, tuttavia. Il leader di Aum Shinrikyo, Seiichi Endo, aveva una formazione in virologia presso l'Università di Kyoto e cercò di produrre sia antrace che ebola. Tuttavia, nel 1995, persino lui mancava di sufficiente competenza e risorse per riuscirci. La barra è ora sostanzialmente più bassa, e gli LLM potrebbero ridurla ulteriormente.
24. Un fenomeno bizzarro relativo agli assassini di massa è che lo stile di omicidio che scelgono opera quasi come una sorta di grottesca moda. Negli anni '70 e '80, i serial killer erano molto comuni, e i nuovi serial killer spesso copiavano il comportamento di serial killer più affermati o famosi. Negli anni '90 e 2000, le sparatorie di massa sono diventate più comuni, mentre i serial killer sono diventati meno comuni. Non c'è alcun cambiamento tecnologico che abbia innescato questi schemi comportamentali, sembra semplicemente che gli assassini violenti stessero copiando il comportamento degli altri e la cosa "popolare" da copiare sia cambiata.
25. I jailbreaker casuali a volte credono di aver compromesso questi classificatori quando riescono a far sì che il modello emetta un pezzo specifico di informazione, come la sequenza genomica di un virus. Ma come ho spiegato prima, il modello di minaccia di cui ci preoccupiamo coinvolge consigli interattivi passo-passo che si estendono per settimane o mesi su specifici e oscuri passaggi nel processo di produzione di armi biologiche, ed è questo che i nostri classificatori mirano a difendere. (Spesso descriviamo la nostra ricerca come la ricerca di jailbreak "universali" quelli che non funzionano soloo in uno specifico o stretto contesto, ma aprono ampiamente il comportamento del modello).
26. Sebbene continueremo a investire nel lavoro per rendere i nostri classificatori più efficienti, e potrebbe aver senso che le aziende condividano progressi come questi l'una con l'altra.
27. Ovviamente, non penso che le aziende debbano divulgare dettagli tecnici sui

passaggi specifici nella produzione di armi biologiche che stanno bloccando, e la legislazione sulla trasparenza che è stata approvata finora (SB 53 e RAISE) tiene conto di questo problema.
28. Un'altra idea correlata sono i "mercati della resilienza" dove il governo incoraggia l'accumulo di DPI, respiratori e altre attrezzature essenziali necessarie per rispondere a un attacco biologico promettendo in anticipo di pagare un prezzo concordato per queste attrezzature in caso di emergenza. Ciò incentiva i fornitori ad accumulare tali attrezzature senza timore che il governo le sequestri senza compenso.
29. Perché sono più preoccupato per i grandi attori per la presa del potere, ma per i piccoli attori per quanto riguarda il causare distruzione? Perché le dinamiche sono diverse. Prendere il potere riguarda il fatto che un attore possa accumulare abbastanza forza da sopraffare tutti gli altri quindi dovremmo preoccuparci degli attori più pootenti e/o quelli più vicini all'IA. La distruzione, al contrario, può essere operata da chi ha poco potere se è molto più difficile difendersi che causare il danno. È allora un gioco di difesa contro le minacce più numerose, che è probabile siano attori più piccoli.
30. Questo potrebbe suonare in tensione con il mio punto secondo cui attacco e difesa potrebbero essere più bilanciati con gli attacchi informatici rispetto alle armi biologiche, ma la mia preoccupazione qui è che se l'IA di un paese è la più potente al mondo, allora gli altri non saranno in grado di difendersi anche se la tecnologia stessa ha un intrinseco equilibrio attacco-difesa.
31. Ad esempio, negli Stati Uniti questo include il Quarto Emendamento e il Posse Comitatus Act.
32. Inoltre, per essere chiari, ci sono alcuni argomenti a favore della costruzione di grandi datacenter in paesi con varie strutture di governance, in particolare se sono controllati da aziende in democrazie. Tali buildout potrebbero in linea di principio aiutare le democrazie a competere meglio con il PCC, che è la minaccia maggiore. Penso anche che tali datacenter non pongano molto rischio a meno che non siano molto grandi. Ma nel complesso, penso che la cautela sia giustificata quando si posizionano datacenter molto grandi in paesi in cui le salvaguardie istituzionali e le protezioni dello stato di diritto sono meno consolidate.
33. Questo è, ovviamente, anche un argomento per migliorare la sicurezza del deterrente nucleare per renderlo più propenso a essere robusto contro un'IA potente, e le democrazie dotate di armi nucleari dovrebbero farlo. Ma non sappiamo di cosa sarà capace un'IA potente o quali difese, se presenti, funzioneranno contro di essa, quindi non dovremmo assumere che queste misure risolveranno necessariamente il problema.
34. C'è anche il rischio che anche se il deterrente nucleare rimanesse efficace, un paese attaccante potrebbe decidere di sfidare il nostro bluff non è chiaro se saremmo disposti a usare armi nucleari per difenderci contro uno sciame di droni anche se lo sciame di droni avesse un rischio sostanziale di conquistarci. Gli sciami di droni potrebbero essere una cosa nuova, meno grave degli attacchi nucleari ma più grave degli attacchi convenzionali. In alternativa, diverse valutazioni dell'efficacia del deterrente nucleare nell'era dell'IA potrebbero alterare la teoria dei giochi del conflitto nucleare in modo destabilizzante.

35. Per essere chiari, riterrei che la strategia giusta sia non vendere chip alla Cina, anche se l'orizzonte temporale per l'IA potente fosse sostanzialmente più lungo. Non possiamo rendere i cinesi "dipendenti" dai chip americani sono determminati a sviluppare la loro industria nazionale dei chip in un modo o nell'altro. Ci vorranno molti anni perché lo facciano, e tutto ciò che stiamo facendo vendendo loro chip è dare loro un grande impulso durante quel periodo.
36. Per essere chiari, la maggior parte di ciò che viene usato in Ucraina e a Taiwan oggi non sono armi completamente autonome. Queste stanno arrivando, ma non sono qui oggi.
37. La nostra model card per Claude Opus 4.5, il nostro modello più recente, mostra che Opus ottiene risultati migliori in un colloquio di ingegneria delle prestazioni frequentemente somministrato ad Anthropic rispetto a qualsiasi candidato nella storia dell'azienda.
38. "Scrivere tutto il codice" e "svolgere il compito di un ingegnere informatico end-to-end" sono due cose molto diverse, perché gli ingegneri informatici fanno molto di più che scrivere semplicemente codice, inclusi test, gestione degli ambienti, dei file e dell'installazione, gestione del cloud compute, iterazione sui prodotti e molto altro ancora.
39. I computer sono generali in un certo senso, ma sono chiaramente incapaci da soli della stragrande maggioranza delle abilità cognitive umane, pur superando di gran lunga gli umani in alcune aree (come l'aritmetica). Naturalmente, le cose costruite sopra i computer, come l'IA, sono ora capaci di una vasta gamma di abilità cognitive, che è ciò di cui tratta questo saggio.
40. Per essere chiari, i modelli di IA non hanno esattamente lo stesso profilo di punti di forza e debolezza degli umani. Ma stanno anche avanzando in modo piuttosto uniforme lungo ogni dimensione, tale che avere un profilo "spigoloso" o irregolare potrebbe alla fine non contare.
41. Anche se esiste un dibattito tra gli economisti su questa idea.
42. La ricchezza personale è uno "stock", mentre il PIL è un "flusso", quindi questa non è l'affermazione che Rockefeller possedesse il 2% del valore economico negli Stati Uniti. Ma è più difficile misurare la ricchezza totale di una nazione rispetto al PIL, e i redditi individuali variano molto all'anno, quindi è difficile creare un rapporto nelle stesse unità. Il rapporto tra la più grande fortuna personale e il PIL, pur non confrontando mele con mele, è comunque un parametro perfettamente ragionevole per l'estrema concentrazione della ricchezza.
43. Il valore totale del lavoro nell'intera economia è di 60 trilioni di dollari all'anno, quindi 3 trilioni all'anno corrisponderebbero al 5% di questo. Quella somma potrebbe essere guadagnata da un'azienda che fornisse manodopera per il 20% del costo degli umani e avesse una quota di mercato del 25%, anche se la domanda di lavoro non si espandesse (cosa che quasi certamente accadrebbe a causa del minor costo).
44. Per essere chiari, non penso che l'attuale produttività dell'IA sia già responsabile di una frazione sostanziale della crescita economica degli Stati Uniti. Piuttosto, penso che la spesa per i datacenter rappresenti una crescita causata da investimenti anticipatori che equivalgono al mercato che si aspetta una futura crescita economica guidata dall'IA e investe di conseguenza.

45. Quando siamo d'accordo con l'amministrazione, lo diciamo, e cerchiamo punti di accordo dove le politiche reciprocamente sostenute siano genuinamente buone per il mondo. Puntiamo ad essere mediatori onesti piuttosto che sostenitori o oppositori di un dato partito politico.
46. Non credo che sia possibile nulla di più di qualche anno: su scale temporali più lunghe, costruiranno i propri chip.