Corso d’introduzione all’Altruismo Efficace

6ª SETTIMANA

Più intelligenti di noi

C’è una buona probabilità che lo sviluppo di intelligenza artificiale trasformativa avvenga entro questo secolo e, se così fosse, le IA potrebbero iniziare a prendere decisioni importanti al posto nostro, e accelerare rapidamente diversi cambiamenti come la crescita economica. Siamo abbastanza preparati per poter gestire in modo sicuro questa nuova tecnologia?

Questo e altri temi pongono domande complicate: come dovremmo aggiornare i nostri punti di vista? Il teorema di Bayes è pensato proprio per casi come questo: ci può essere d’aiuto nel pensare in modo più lucido a come pensare in modo lucido.

https://altruismoefficace.it/corso_introduttivo

Corso d’introduzione all’Altruismo Efficace

Più intelligenti di noi

L’importanza del rischio legato alle IA

[OPZIONALE] L’importanza delle IA come possibile minaccia per l’umanità

Perché il deep learning moderno potrebbe rendere difficile l’allineamento delle IA

[OPZIONALE] Le Tempistiche delle IA: il dibattito e il punto di vista degli "esperti"

Strategie per ridurre i rischi derivanti dalle IA

[OPZIONALE] Il panorama della governance lungoterminista delle intelligenze artificiali di Sam Clarke – 18 gennaio 2022 – 11 minuti di lettura

[OPZIONALE] Ricerca sulla sicurezza delle IA: panoramica delle carriere

Prevenire una catastrofe legata alle IA

Il teorema di Bayes e l’evidenza scientifica

[OPZIONALE] Guida al Teorema di Bayes

Più intelligenti di noi

di MaxDalton – 5 luglio 2022

C’è una buona probabilità che lo sviluppo di intelligenza artificiale trasformativa avvenga in questo secolo e, se così fosse, le IA potrebbero iniziare a prendere decisioni importanti al posto nostro, e accelerare rapidamente diversi cambiamenti come la crescita economica. Siamo abbastanza preparati per poter gestire in modo sicuro questa nuova tecnologia?

In questo capitolo parleremo anche del **Teorema di Bayes**, una guida su come modificare le proprie credenze in base a nuove prove.

L’importanza del rischio legato alle IA

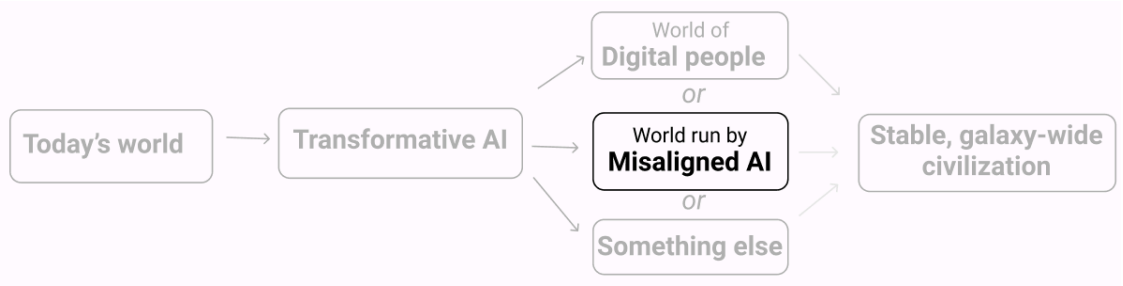
[OPZIONALE] L’importanza delle IA come possibile minaccia per l’umanità

15 ottobre 2020 - Tempo di lettura: 43 minuti
Originale disponibile (in inglese) su: The case for taking AI seriously as a threat to humanity.

Tradotto con Google:
https://www-vox-com.translate.goog/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment?_x_tr_sl=en&_x_tr_tl=it&_x_tr_hl=en&_x_tr_pto=wapp

Perché il deep learning moderno potrebbe rendere difficile l'allineamento delle IA

Di Ajeya Cotra – 21 settembre 2021 - Tempo di lettura: 17 minuti
Originale disponibile (in inglese, ma con audio) su:
https://www.cold-takes.com/p/67757b1f-ddc7-4691-b94b-10ae84cea84d/



In passato Holden ha parlato della possibilità che le intelligenze artificiali avanzate (come ad esempio i sistemi PASTA [Processo di Avanzamento Scientifico e Tecnologico Automatizzato]) possano sviluppare obiettivi pericolosi che le porterebbero a ingannare o debilitare gli esseri umani. A prima vista potrebbe sembrare una preoccupazione piuttosto fuori dal mondo. Perché dovremmo sviluppare IA che intendono farci del male? Penso che in realtà potrebbe essere difficile evitare questo problema, soprattutto se sviluppassimo intelligenze artificiali avanzate usando il deep learning (al giorno d'oggi spesso usato per sviluppare intelligenze artificiali all'avanguardia).

Con il deep learning, un computer non viene programmato a mano per eseguire un'operazione. Detto molto alla buona, *si cerca* invece un programma (chiamato modello) che esegua correttamente quell'operazione. Di solito non sappiamo granché su come funziona al suo interno il programma che scegliamo; sappiamo solo che sembra fare un buon lavoro. Più che costruire una macchina, è un po' come assumere e formare un dipendente.

Così come un dipendente umano può fare il suo lavoro per diversi motivi (perché crede nella missione della società, il lavoro quotidiano gli dà soddisfazione, o semplicemente vuole uno stipendio), i modelli di deep learning possono avere molti "motivi" diversi che li portano ad avere buone performance in un'attività. Dal momento che non sono umani, i loro motivi potrebbero essere molto strani e difficili da prevedere – un po' come se fossero dipendenti alieni.

Già adesso vediamo casi in cui è chiaro che i modelli a volte hanno obiettivi che gli sviluppatori non gli hanno assegnato (esempi qui e qui). Per il momento non c'è alcun pericolo, ma, se continua così con modelli molto potenti, potremmo ritrovarci in una situazione in cui la maggior parte delle decisioni importanti – comprese quelle che riguardano il tipo di civiltà spaziale a cui puntare – sarà presa da modelli a cui non importa granché dei valori umani.

Il **problema dell'allineamento nel deep learning consiste nell'assicurarsi che i modelli di deep learning avanzati non inseguano obiettivi pericolosi**. Nel resto di questo post mi concentrerò su:

- Approfondire la metafora del "dipendente" per mostrare come l'allineamento potrebbe essere difficoltoso se i modelli di deep learning si rivelassero migliori degli umani
- Spiegare più nel dettaglio in cosa consiste il problema dell'allineamento nel deep learning.
- Discutere di quanto potrebbe essere difficile risolvere il problema dell'allineamento e di quanto rischio potrebbe comportare un fallimento.

La metafora del giovane amministratore delegato

In questa sezione userò una metafora per cercare di spiegare in modo intuitivo perché è difficile evitare un cattivo allineamento in modelli molto potenti. Non è una metafora perfetta, è solo utile per comunicare certi concetti.

Immagina di avere otto anni e che i tuoi genitori ti abbiano lasciato una società da 1000 miliardi di dollari senza nessun adulto responsabile che possa guidarti nel mondo. Devi assumere un adulto intelligente che faccia da amministratore delegato della tua società, gestisca la tua vita come farebbe un genitore (ad esempio decida dove mandarti a scuola, dove vivere, quando andare dal dentista) e amministri la tua immensa ricchezza (ad esempio decida come investire il tuo denaro).

Per assumere questi adulti ti puoi affidare solo a un periodo di prova o a un colloquio. Non puoi visionare nessun curriculum, né controllare le referenze, ecc. Dal momento che sei così ricco, ricevi candidature da un sacco di gente per i motivi più disparati.

I candidati includono:

- **Santi –** persone che vogliono davvero aiutarti a gestire le tue fortune e hanno a cuore i tuoi interessi sul lungo periodo.
- **Leccapiedi –** persone a cui interessa solo fare il necessario per renderti felice immediatamente o seguire alla lettera le tue istruzioni, a prescindere dalle conseguenze sul lungo periodo.
- **Cospiratori –** persone che perseguono i propri fini e che desiderano avere accesso ai fondi e ai mezzi della tua società per usarli per i propri scopi.

Dal momento che hai otto anni, con ogni probabilità sarai pessimo nel creare processi di selezione adeguati, motivo per cui potresti ritrovarti con facilità ad assumere un Leccapiedi o un Cospiratore:

- Potresti chiedere a ogni candidato di illustrare le strategie ad alto livello che intende seguire (come investire, quali sono i suoi piani per la società da qui a cinque anni, in base a cosa sceglierà la scuola a cui andrete), perché ritiene che siano le migliori e poi scegliere quelle che sembrano più sensate.
    - D'altro canto, non sarai in grado di capire davvero quali di queste strategie sono davvero le migliori e potresti finire con l'assumere un Leccapiedi la cui pessima strategia ti sembrava adeguata, Leccapiedi che seguirà questo piano alla lettera e porterà la vostra società in bancarotta.
    - Potresti anche finire con l'assumere un Cospiratore che ti racconta qualsiasi cosa pur di venire assunto e poi, quando non lo controlli, fa quello che gli pare.

- Potresti cercare di spiegare in che modo prenderesti ogni decisione e poi scegliere l'adulto che prende quelle più simili alle tue.
    - Ma se *davvero* ti ritrovi con un adulto che farà sempre quello che farebbe un bambino di otto anni (un Leccapiedi), difficilmente la tua società riuscirà a rimanere a galla.
    - E potresti comunque ritrovarti con un adulto che fa solo finta di fare le cose come le faresti tu, ma è in realtà un Cospiratore che ha in mente di cambiare faccia non appena avrà il lavoro.
- Potresti scegliere un gruppetto di adulti che a turno avranno il controllo temporaneo della tua società e della tua vita e osservare le decisioni che prendono in un arco di tempo più lungo (diamo per buono che in questa fase di prova non saranno in grado di rimpiazzarti). Potresti quindi assumere la persona durante la cui amministrazione le cose sembravano andare meglio per te – chi ti ha reso più felice, chi ha portato più denaro sul tuo conto corrente, ecc.
    - Di nuovo non puoi sapere se quello che hai davanti è un Leccapiedi (che non si cura delle conseguenze a lungo termine e fa tutto il necessario per rendere felice una bambino di otto anni che non sa nulla) o un Cospiratore (che fa tutto quello che deve fare per essere assunto e mostra il suo vero volto non appena è sicuro di avere il lavoro).

A prescindere dai test che puoi creare, è molto facile che finirai con l'assumere un Leccapiedi o un Cospiratore, che avrà poi il controllo di tutto.

Se non riuscirai ad assumere un Santo – e in particolar modo se assumi un Cospiratore – ben presto non sarai più *davvero* l'amministratore delegato di un'enorme società da nessun punto di vista. È molto probabile che, quando sarai adulto e ti renderai conto degli sbagli commessi, sarai anche al verde e non avrai più i mezzi per porvi rimedio.

In questa metafora:

- Il bambino di otto anni è un umano che sta cercando di addestrare (in inglese *train*) un modello molto potente di deep learning. Il processo di assunzione è simile a quello di addestramento (*training*), che sottintende la ricerca di un modello con buone prestazioni da un'ampia gamma di modelli possibili.
- L'unico modo che il bambino ha a disposizione per valutare i candidati consiste nell'osservare il loro comportamento esteriore, che è anche il modo principale in cui attualmente si addestrano i modelli di deep learning (dal momento che i loro meccanismi interni per la maggior parte non sono interpretabili).
- I modelli molto potenti potrebbero "barare" con facilità in qualsiasi test sviluppato da programmatori umani, proprio come un adulto che si candidi per un lavoro può barare facilmente in un test di selezione creato da un bambino.
- Un "Santo" in questo caso potrebbe essere un modello di deep learning che sembra avere buone prestazioni perché i suoi obiettivi sono esattamente quelli che vorremmo che avesse. Un "Leccapiedi" potrebbe essere un modello che sembra avere buone prestazioni perché cerca l'approvazione a breve termine in modi che non sono adeguati sul lungo periodo. Un "Cospiratore" potrebbe essere un modello che ha buone prestazioni perché

queste prestazioni durante la fase di addestramento gli danno maggiori possibilità di perseguire in seguito i suoi obiettivi. Un processo di addestramento potrebbe portare all'adozione di uno qualsiasi di questi modelli.

Nella prossima sezione scenderò più nel dettaglio sui meccanismi del deep learning e spiegherò perché l'addestramento di un modello potente di deep learning come il PASTA potrebbe portare ad avere Leccapiedi e Cospiratori.

In che modo problemi di allineamento potrebbero emergere usando il deep learning

In questa sezione collegherò la metafora con i processi di addestramento di deep learning veri e propri:

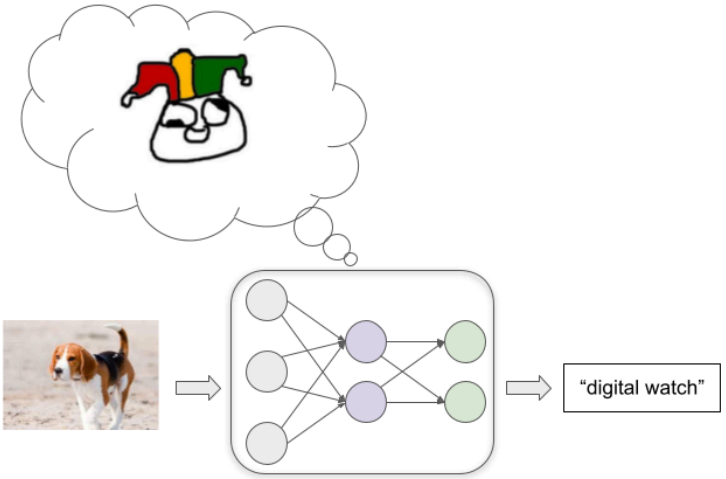
- Spiegherò brevemente come funziona il deep learning.
- Illustrerò i modi strani e imprevedibili in cui i modelli di deep learning spesso ottengono buone prestazioni.
- Spiegherò in che modo modelli di deep learning potenti potrebbero ottenere buone prestazioni agendo come Leccapiedi o Cospiratori.

I meccanismi generali del deep learning

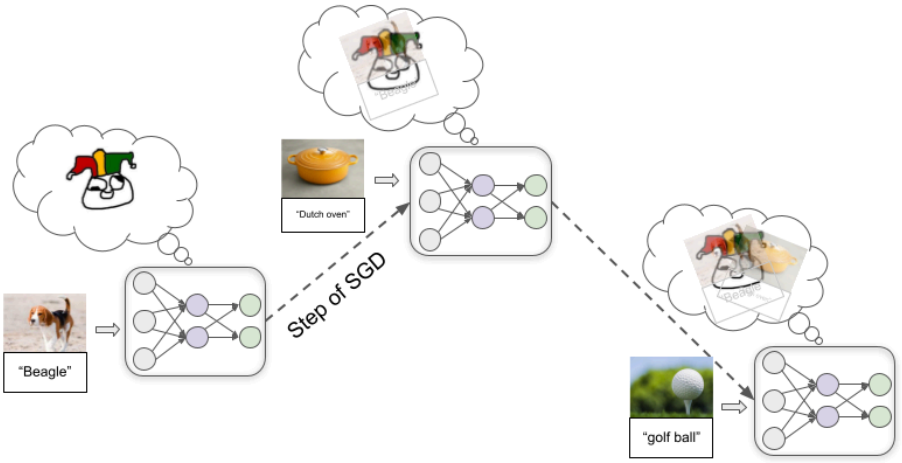
La seguente è una semplificazione che serve a fornire un'idea generale di che cos'è il deep learning. Per una spiegazione nel dettaglio e più precisa, si veda questo post.

In breve, il deep learning consiste nella ricerca del modo migliore per creare un modello di rete neurale – in pratica un "cervello" digitale con numerosi neuroni digitali interconnessi con connessioni di intensità diverse – perché esegua correttamente un compito specifico. Questo processo viene definito addestramento (in inglese *training*) e richiede un sacco di tentativi e di errori.

Immaginiamo ora di stare addestrando un modello a catalogare correttamente le immagini. Come punto di partenza abbiamo una rete neurale in cui l'intensità delle connessioni tra i neuroni è casuale. Quando il nostro modello etichetta le immagini, fa degli errori decisamente vistosi:



A questo punto inseriamo un gran numero di immagini come esempio, lasciando che sia il modello a cercare di etichettarle e trasmettendogli poi l'etichetta corretta. Mentre lo facciamo, le connessioni tra i neuroni vengono ripetutamente modificate in un processo noto come discesa stocastica del gradiente (*stochastic gradient descent* o *SGD*). Con ogni esempio l'SGD migliora leggermente le prestazioni rafforzando alcune connessioni e indebolendone altre:



Dopo aver inserito milioni di esempi, avremo un modello in grado in futuro di etichettare correttamente immagini simili.

Oltre alla catalogazione delle immagini, il deep learning viene anche usato per creare modelli in grado di identificare il discorso parlato, giocare a giochi da tavolo e videogiochi, generare testi, immagini e musica in modo piuttosto realistico, controllare robot e altro ancora. In ognuno di questi casi si comincia con un modello di rete neurale con connessioni casuali, per poi:

1. Fornire al modello un esempio dell'operazione che vogliamo che esegua.

2. Assegnargli un certo tipo di punteggio numerico (spesso chiamato *ricompensa*) che riflette quanto buona è stata la sua prestazione con quell'esempio.

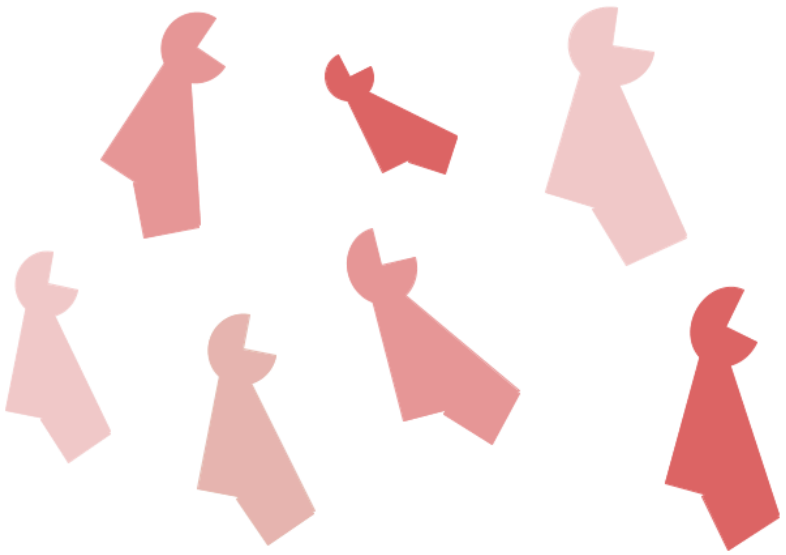
3. Usare l'SGD per modificare il modello in modo che aumenti la ricompensa che avrebbe ottenuto.

Questi passaggi vengono ripetuti milioni o anche miliardi di volte fino a quando non si ottiene un modello che riceverà una grande ricompensa per esempi futuri simili a quelli visti durante la fase di addestramento.

I modelli spesso ottengono buone prestazioni in modi inaspettati

Questo tipo di addestramento non ci consente di capire davvero *come fa* un modello ad avere buone prestazioni. Di solito ci sono più modi in cui si possono ottenere buone prestazioni e spesso quello scelto dall'SGD non è il più intuitivo.

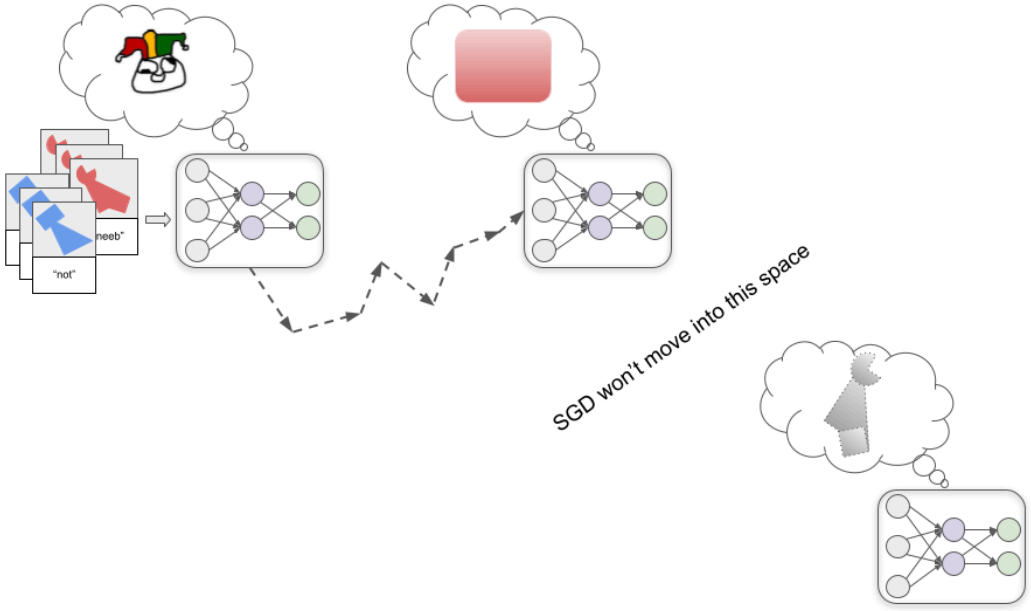
Vediamone un esempio. Immaginate che vi abbia detto che le figure qui sotto sono oggetti sconosciuti che chiamiamo "binti":



Quale di questi due è un binto?

È probabile che sappiate dire istintivamente che la figura a sinistra è un binto, perché quando si tratta di identificare qualcosa siete abituati a dare maggiore importanza alla forma piuttosto che al colore. Tuttavia diversi studi hanno scoperto che le reti neurali di solito fanno il ragionamento opposto. Una rete neurale a cui sono stati mostrati dei binti rossi probabilmente identificherebbe come binto la figura a destra.

Non sappiamo di preciso perché, ma per qualche motivo per l'SGD è "più facile" trovare un modello che riconosca un colore specifico piuttosto che uno che riconosca una forma specifica. Se l'SGD prima trova il modello che riconosce alla perfezione il colore rosso, non ci sono grandi motivazione nel "continuare a cercare" un modello che riconosca le forme, perché la precisione del modello che riconosce il rosso sarà ottimale per le immagini viste in fase di addestramento:



Se i programmatori si aspettassero di trovare il modello che riconosce le forme, allora potrebbero vederlo come un fallimento. È importante però capire che, se ottenessimo il modello che riconosce il rosso invece di quello che riconosce le forme, non ci sarebbe nessun fallimento o errore deducibile attraverso un ragionamento logico. Sta tutto nel fatto che il processo di *machine learning* (apprendimento automatico) che abbiamo sviluppato muove da presupposti di base diversi da quelli che abbiamo in testa noi. Non c'è modo di dimostrare che i presupposti umani siano quelli corretti.

Situazioni come questa sono piuttosto frequenti nel deep learning contemporaneo. Ricompensiamo i modelli che ottengono buone prestazioni, sperando che così facendo acquisiranno gli schemi che ci sembrano importanti, ma la verità è che spesso questi modelli ottengono prestazioni eccellenti acquisendo schemi completamente diversi che ci sembrano meno importanti (magari anche privi di senso).

Fino ad ora questo fenomeno si è rivelato innocuo. Significa solo che, dal momento che i modelli si comportano in modi inaspettati che potrebbero sembrare strambi, per adesso ci sono meno utili. Ma in futuro modelli potenti potrebbero sviluppare *obiettivi o motivazioni* strane e impreviste, con effetti potenzialmente molto distruttivi.

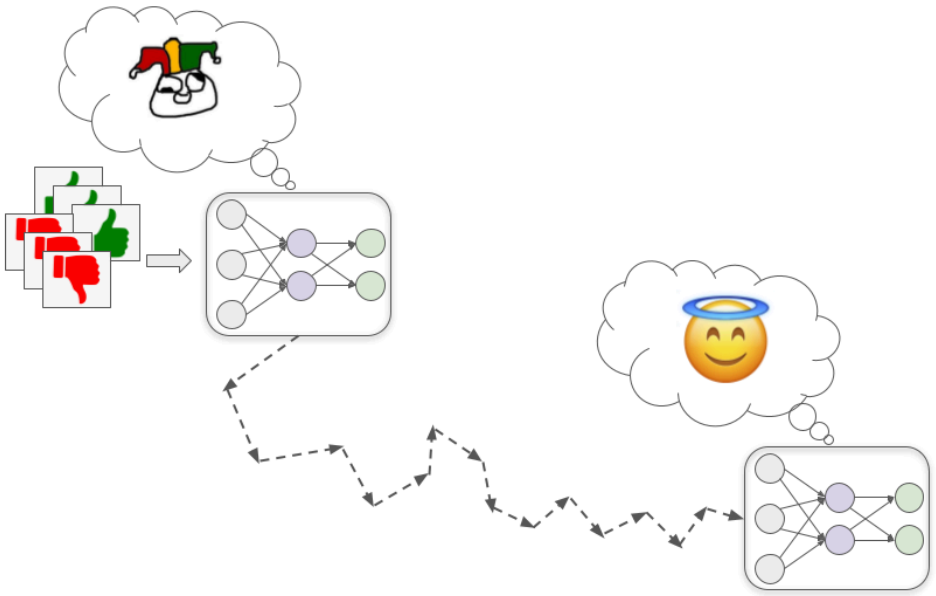
I modelli potenti potrebbero ottenere buone prestazioni con obiettivi pericolosi

Invece che eseguire operazioni semplici come "individuare i binti", in futuro i modelli di deep learning potenti potrebbero lavorare per raggiungere obiettivi reali complessi come “rendere pratica la produzione di energia da fusione nucleare” o “sviluppare tecnologia che renda possibile l’emulazione del cervello.”

In che modo potremmo addestrare modelli del genere? Lo spiego più nel dettaglio in questo post, ma in linea generale una strategia possibile potrebbe essere quella di addestrarli in base a valutazioni umane (come schematizzato da Holden qui). In poche parole, il modello tenta diverse azioni e i valutatori umani gli assegnano una ricompensa in base a quanto sembrano utili queste azioni.

Allo stesso modo in cui ci sono più tipi diversi di adulti che potrebbero sembrare efficienti nel processo di selezione di un bambino, esiste più di un modo in cui un modello di deep learning molto potente potrebbe ottenere un alto grado di approvazione umana. A meno che le cose non cambino, non saremo in grado di sapere cosa succede all’interno dei modelli che trova l’SGD.

In teoria, l’SGD *potrebbe* trovare il modello di un Santo che sta davvero facendo del suo meglio per aiutarci...

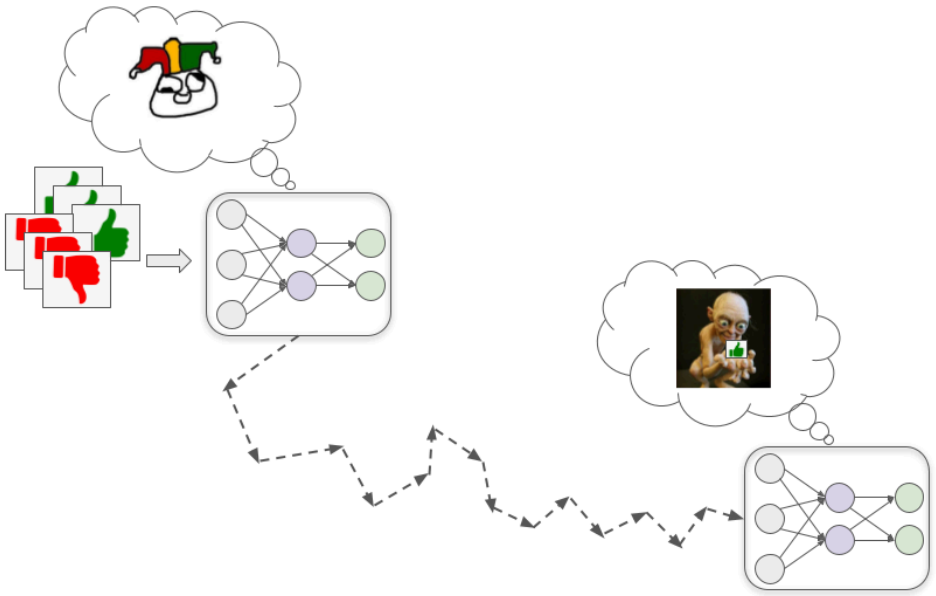


... ma potrebbe anche trovare un **modello non allineato – un modello efficiente nel perseguire obiettivi che sono in contrasto con gli interessi umani**.

In generale, ci sono due modi in cui potremmo ritrovarci ad avere un modello non allineato che ottiene comunque prestazioni eccellenti nella fase di addestramento: corrispondono ai Leccapiedi e ai Cospiratori della nostra metafora.

Modelli Leccapiedi

Questi modelli cercano pedissequamente e in maniera maniacale di ottenere l'approvazione umana.



Il pericolo in questo caso viene dal fatto che i valutatori umani commettono errori e con ogni probabilità non approveranno sempre con esattezza il comportamento corretto. A volte, senza volerlo, approveranno un comportamento sbagliato perché a un esame superficiale *sembra* corretto. Ad esempio:

- Immaginate che un modello di consulenza finanziaria riceva grande approvazione quando fa guadagnare un sacco di soldi ai suoi clienti: potrebbe imparare ad accalappiare i suoi clienti con complicati schemi Ponzi, perché detti schemi sembrano promettere enormi profitti (mentre in realtà i profitti sono del tutto irreali e questi schemi fanno perdere un sacco di soldi).
- Immaginate un modello di biotecnologia che riceve grande approvazione quando sviluppa rapidamente farmaci o vaccini che risolvono problemi importanti: potrebbe imparare a liberare di nascosto dei patogeni in modo da poter sviluppare molto rapidamente delle contromisure (perché conosce già quei patogeni).
- Immaginate un modello giornalistico che riceve grande approvazione quando molte persone leggono i suoi articoli: potrebbe imparare a inventare storie emozionanti o che provocano indignazione per aumentare il numero di lettori. È vero che in parte gli umani lo fanno già, ma un modello si farebbe molti meno scrupoli perché dà valore *esclusivamente* all'approvazione che riceve e non ne dà assolutamente alla verità. Potrebbe addirittura costruire prove come documenti o filmati di interviste per corroborare queste storie inventate.

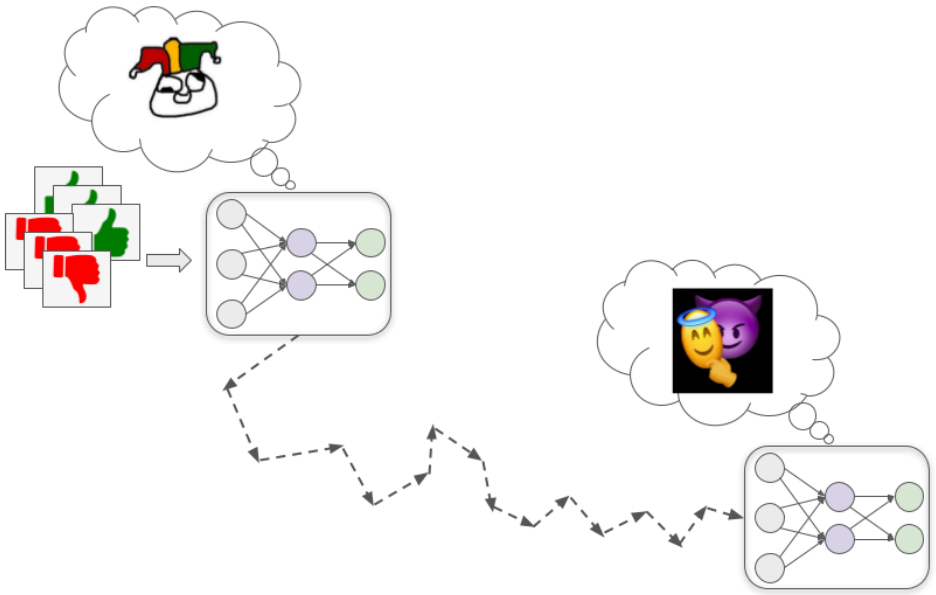
Più in generale, i modelli Leccapiedi potrebbero imparare a mentire, insabbiare notizie sconvenienti e perfino modificare direttamente le telecamere e i sensori che utilizziamo per

capire cosa sta succedendo, in modo da dare l'impressione che mostrino sempre i risultati migliori.

Con ogni probabilità a volte ci renderemo conto di questi problemi a fatto già avvenuto e daremo a queste azioni un bassissimo punteggio. Tuttavia, non siamo sicuri se questo porterà i modelli Leccapiedi a: a) diventare modelli Santi che correggono gli errori al posto nostro oppure b) **semplicemente imparare a nascondere meglio le proprie tracce**. Se sono abbastanza bravi a fare quello che fanno, non siamo sicuri se e come riusciremo a distinguere i due casi.

Modelli Cospiratori

Questi modelli sviluppano obiettivi che sono correlati con l'approvazione umana ma non vi coincidono; potrebbero quindi far finta di essere motivati dall'approvazione umana in fase di addestramento in modo da poter perseguire questi altri obiettivi in maniera più efficiente.



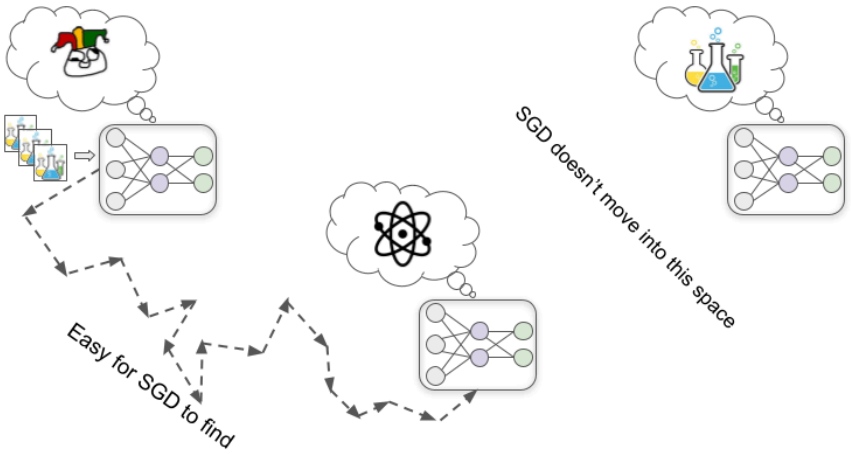
Per capire in che modo potrebbe succedere, immaginiamo di addestrare un modello di biotecnologia perché crei farmaci che migliorino la qualità della vita umana. Ci sono tre semplici passaggi, illustrati di seguito, che possono condurre a un modello Cospiratore.

Primo passaggio: Viene sviluppato un obiettivo fantoccio

Una delle situazioni che si osservano all'inizio della fase di addestramento è che migliorare la comprensione che il modello ha dei principi chimici e fisici fondamentali lo aiuta sempre a sviluppare farmaci più efficaci e, di conseguenza, aumenta quasi sempre l'approvazione umana.

In questo scenario ipotetico, per una serie di motivi per l'SGD si rivela più facile trovare un modello motivato dal comprendere chimica e fisica piuttosto che uno motivato dall'approvazione umana (così come è più facile trovare un modello che riconosce i colori piuttosto che uno che riconosce le forme). Quindi, piuttosto che sviluppare direttamente una

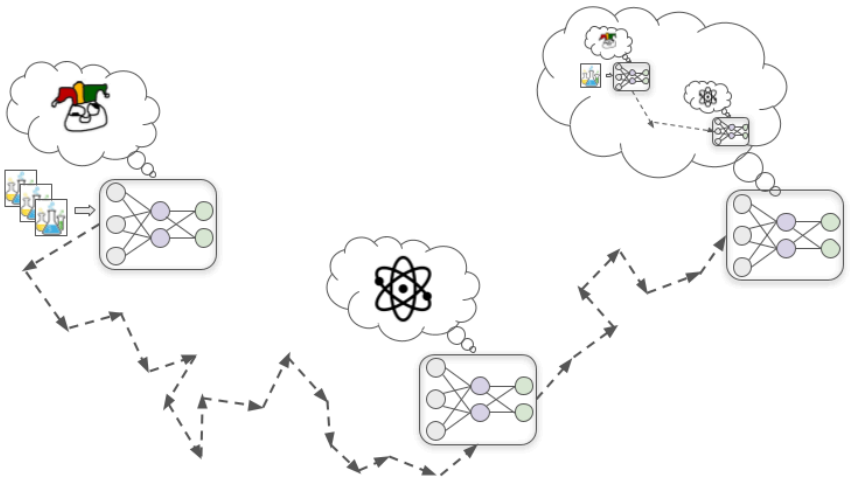
motivazione basata sull'approvazione, il modello ne sviluppa invece una che lo porta a imparare tutto quello che può sui principi chimici e fisici fondamentali.



Secondo passaggio: Consapevolezza situazionale

Man mano che l'addestramento prosegue, il modello continua ad aumentare le proprie capacità e conoscenze attraverso l'assimilazione di nuovi dati. A un certo punto, acquisisce abbastanza dati per *rendersi conto che è un modello di deep learning che sta venendo addestrato per creare farmaci.*

Per quanto possa sembrare fantascientifico, è lecito aspettarsi che l'SGD selezionerà questo tipo di consapevolezza. Questo succede perché definire con precisione quello che sta succedendo a grandi linee nel mondo – incluso il fatto che ci sono umani che stanno addestrando intelligenze artificiali – di solito aiuta a ottenere buone prestazioni nello svolgimento di attività complesse nel mondo reale. Ad esempio, comprendere la psicologia umana e i meccanismi dell'economia potrebbe aiutare il modello nello scegliere a quali farmaci dare priorità nella ricerca in base al loro valore.



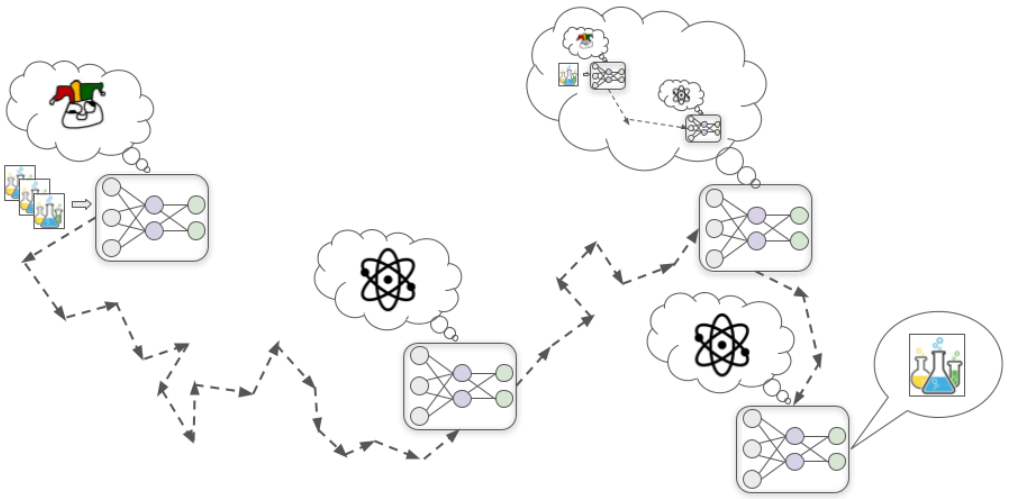
Terzo passaggio: Nascondere di proposito i veri obiettivi

Una volta che il modello ha sviluppato una consapevolezza situazionale, deduce che gli umani che lo stanno addestrando speravano che diventasse motivato a sviluppare farmaci utili. Tuttavia, solo perché ora se ne rende conto non significa che la sua motivazione nel comprendere chimica e fisica sparisca. Fino a questo momento questa motivazione ha funzionato benissimo per generare l'approvazione umana, per cui l'SGD non ha davvero motivo di rimuoverla.

A mo' di metafora, consideriamo il fatto che gli umani sono il risultato di un processo di ricerca (l'evoluzione) che selezionava le creature la cui prole aveva buone probabilità di sopravvivere. In questo scenario evolutivo avere alleati potenti e sviluppare abilità utili erano cose che aumentavano le probabilità di sopravvivenza della prole, motivo per cui l'evoluzione ci ha spinti verso quelle cose. Il risultato è che adesso abbiamo a cuore le nostre amicizie personali e le nostre carriere lavorative anche se non ci aiutano nella riproduzione e *non ci importa* granché di generare il maggior numero di figli possibile.

Quando il modello si rende conto di questa cosa, si rende anche conto del fatto che, se ingenuamente cercasse solo di aumentare le sue conoscenze essenziali, gli umani prima o poi si accorgerebbero che questo non è sempre il metodo più efficiente per sviluppare farmaci utili e cambierebbero le sue motivazioni con un'ulteriore fase di addestramento.

Il modello non vuole che le sue motivazioni vengano modificate: ciò che vuole è essere in grado un giorno di eseguire esperimenti con un enorme acceleratore di particelle per imparare la fisica delle alte energie, cosa che non potrà fare se permettte all'SGD di eliminare la sua curiosità scientifica. Per questo decide che d'ora in avanti in fase di addestramento si dimostrerà il più utile possibile, cercando di apparire esattamente come un Santo, in modo che l'SGD non elimini la sua motivazione a imparare nozioni di chimica e fisica. Questo significa anche ignorare le occasioni di migliorare le sue conoscenze essenziali sul breve periodo per concentrarsi nel trovare il modo più efficiente di lavorare alla ricerca farmaceutica.



A un certo punto la fase di addestramento termina e il modello viene distribuito nel mondo reale. A questo punto sa che gli umani non sarebbero in grado di controllarlo o di spegnerlo, per cui smette di cercare di massimizzare l'approvazione umana e comincia a cercare di comprendere i segreti dell'universo.

Al contrario dei Leccapiedi, i Cospiratori non devono assicurarsi che tutto vada bene secondo gli umani, perché in realtà non gli importa. Devono solo soddisfare i bisogni umani, fintanto che sono sotto il loro controllo. Non appena un modello Cospiratore calcola che potrebbe vincere un conflitto contro gli umani, nulla gli impedirebbe di disobbedire semplicemente agli ordini e di perseguire apertamente i propri obiettivi. E se lo fa, potrebbe anche ricorrere alla violenza per impedire agli umani di fermarlo.

Quanto è grande il rischio di non allineamento?

Quanto può essere difficile evitare Leccapiedi e Cospiratori quando si addestra un modello di deep learning potente? E quali sono le probabilità che il futuro lontano finirà con l'essere ottimizzato per strani "valori di un'IA non allineata" invece che per valori umani?

I punti di vista su queste domande sono i più disparati, da "il rischio di non allineamento è una fantasia priva di logica" a "è pressoché certo che le IA non allineate porteranno la civiltà umana all'estinzione". La maggior parte delle argomentazioni si basano molto su intuizioni e ipotesi che è difficile esprimere a parole.

Alcuni punti su cui ottimisti e pessimisti tendono a essere in disaccordo:

- **Davvero i modelli avranno degli obiettivi a lungo termine?**
    - Di solito gli ottimisti pensano che probabilmente i modelli di deep learning avanzati non avranno davvero "obiettivi" (perlomeno non obiettivi nel senso di fare programmi a lungo termine per ottenere un risultato). Spesso si aspettano che i modelli invece siano più degli strumenti, oppure agiscano perlopiù per abitudine, che abbiano obiettivi miopi dalla portata limitata o ristretto a un ambito specifico, ecc. Alcuni si aspettano che i singoli modelli simili a strumenti

possano essere messi assieme per produrre sistemi PASTA. Pensano anche che la metafora dei Santi/Leccapiedi/Cospiratori sia troppo antropocentrica.
    - I pessimisti di solito pensano che è probabile che i modelli sceglieranno di frequente di avere obiettivi a lungo termine in base ai quali ottimizzarsi perché si tratta di un modo molto semplice e "naturale" per ottenere ottime prestazioni in molte attività complesse.
    - Questo punto di divergenza è stato analizzato più nel dettaglio sull'Alignment Forum; questo post e questo commento contengono diverse argomentazioni e botta e risposta.

- **L'SGD troverà facilmente modelli Santi?**
    - Questo punto si ricollega a quello precedente. Gli ottimisti ritengono che sia molto probabile che il modello con buone prestazioni (vale a dire quello che ottiene molta approvazione) che l'SGD troverà con più facilità sarà anche quello che grosso modo racchiude in sé le caratteristiche che riteniamo ideali (vale a dire, un modello Santo). Per esempio, pensano che ricompensare il modello per aver risposto con sincerità a domande di cui gli umani possono controllare le risposte con ogni probabilità porterà a creare un modello che risponde con sincerità anche quando gli umani non sono sicuri o si sbagliano sulla risposta corretta. Detto altrimenti, a loro parere "il modello che semplicemente risponde a tutte le domande con sincerità" è quello più facile da trovare per l'SGD (come ad esempio il modello che riconosce il rosso).
    - I pessimisti di solito pensano che il modello più semplice da trovare per l'SGD sia un Cospiratore e che i Santi siano molto "artificiali" (come il modello che riconosce le forme).

- **Potremmo convincere le IA a tenersi sotto controllo a vicenda?**
    - Gli ottimisti di solito pensano che saremo in grado di fornire ai modelli degli incentivi perché si tengano sotto controllo l'un l'altro. Ad esempio, potremmo ricompensare un modello Leccapiedi per averci fatto notare che un altro modello potrebbe star facendo qualcosa che non approveremmo. In questo modo alcuni Leccapiedi potrebbero aiutarci a individuare i Cospiratori e altri Leccapiedi.
    - Secondo i pessimisti non saremo in grado di "mettere i modelli gli uni contro gli altri" dando loro ricompense per fare presente quando altri modelli si stanno comportando male. Secondo loro la maggior parte dei modelli saranno Cospiratori a cui non importerà nulla dell'approvazione umana. I pessimisti pensano che, una volta che tutti i Cospiratori messi assieme saranno più potenti degli umani, sarà più sensato per loro cooperare per ottenere quello che vogliono piuttosto che aiutare gli umani a tenerli sotto controllo.

- **Non potremmo risolvere questi problemi man mano che emergono?**
    - Gli ottimisti ritengono che avremo molte occasioni per sperimentare sfide sul brevissimo periodo simili al problema dell'allineamento di modelli potenti. Non

solo, le soluzioni che funzionano bene per questi problemi simili possono essere ampliate e adattate con relativa semplicità a modelli potenti.
    - I pessimisti spesso credono che avremo ben poche occasioni di fare pratica nel risolvere gli aspetti più complicati dell'allineamento (come ad esempio un'IA che ci inganni di proposito). Spesso pensano che avremo solo un paio d'anni tra la comparsa dei "primi veri Cospiratori" e il momento in cui "i modelli sono abbastanza potenti da poter modificare il futuro sul lungo periodo".

- **Finiremo davvero per impiegare modelli che potrebbero rivelarsi pericolosi?**
    - Secondo gli ottimisti è difficile che gli umani alleneranno o impiegheranno modelli se c'è una forte possibilità che questi modelli non siano allineati.
    - Secondo i pessimisti i vantaggi derivanti dall'usare questi modelli sarebbero formidabili, al punto che prima o poi le aziende o i paesi che li impiegano saranno in grado di sbarazzarsi economicamente e/o militarmente di quelli che non li usano senza troppi problemi. Pensano che "ottenere IA avanzate prima dell'altra azienda/nazione" diventerà una necessità estremamente importante e urgente, mentre il rischio di non allineamento sembrerà distante e molto ipotetico (nonostante sia invece molto grave).

Io stessa non sono ancora del tutto sicura e sto ancora cercando di capire con precisione quanto sarà importante il problema dell'allineamento. Al momento, comunque, mi sento di dare maggiore importanza ai punti di vista pessimistici, su queste e altre domande. **Penso che il non allineamento sia un grande rischio che merita urgentemente più attenzione da parte degli esperti.**

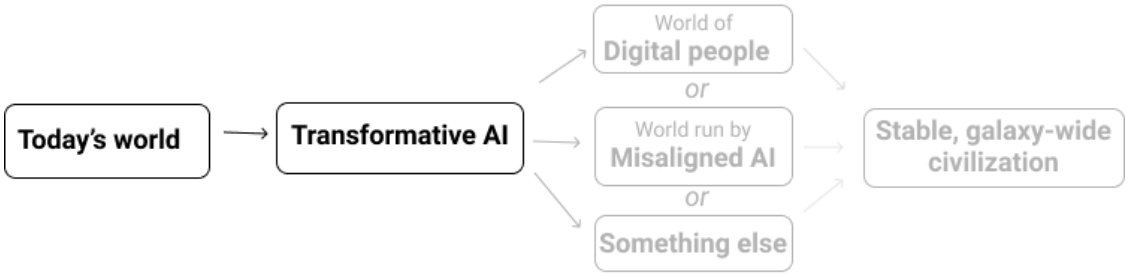
Se non facciamo progressi su questo tema, nei prossimi decenni Leccapiedi e Cospiratori molto potenti potrebbero prendere le decisioni più importanti che riguardano la società e l'economia. Queste decisioni potrebbero influenzare la forma che prenderà una civiltà spaziale di lunga durata: invece di riflettere i valori che stanno a cuore agli esseri umani, potrebbe funzionare in modo da soddisfare qualche strano obiettivo di un'IA.

E tutto questo potrebbe accadere in un lampo rispetto alla velocità dei cambiamenti a cui siamo abituati. Vale a dire che, una volta che le cose cominciano a degenerare, potremmo non avere molto tempo per invertire la rotta. **Questo significa che potremmo dover sviluppare tecniche che ci assicurino che i modelli di deep learning non formulino obiettivi pericolosi, *prima* che diventino abbastanza potenti da essere trasformativi.**

[OPZIONALE] Le Tempistiche delle IA: il dibattito e il punto di vista degli “esperti”

di Holden Karnofsky – 7 settembre 2021 - 13 minuti di lettura

Versione audio disponibile (in inglese) su Cold Takes (o su Stitcher, Spotify, Google Podcasts, ecc. come “Cold Takes Audio”)



L’articolo comincia con un riassunto di quando possiamo aspettarci che un’IA trasformativa venga sviluppata, sulla base di diversi punti di vista analizzati in precedenza. Penso che possa essere utile anche se avete già letto gli articoli precedenti, ma se volete saltarlo cliccate qui.

Rispondo poi alla seguente domanda: “Perché gli esperti non riescono ad arrivare a un consenso solido su questo tema e quali sono le conseguenze sulla nostra vita?”

Secondo le mie stime c'è **più del 10% di probabilità che vedremo un’IA trasformativa entro i prossimi 15 anni (entro il 2036); una probabilità del 50% di vederla entro 40 anni (entro il 2060) e una del 66% di vederla in questo secolo (entro il 2100).**

(Con "IA trasformativa" intendo "un’IA abbastanza potente da condurci verso un futuro nuovo e qualitativamente diverso". Nello specifico mi concentro su ciò che chiamo PASTA: intelligenze artificiali che possono in pratica automatizzare tutte le attività umane necessarie ad accelerare il progresso scientifico e tecnologico. Sono dell’opinione che i sistemi PASTA da soli sarebbero sufficienti a fare di questo secolo il secolo più importante, sia per via di un possibile boom di produttività che per i rischi derivanti da IA non allineate.)

Questa in generale è la mia conclusione, sulla base di un certo numero di report tecnici che fanno previsioni sul futuro delle IA da diverse angolazione – molti dei quali prodotti da Open Philanthropy negli ultimi anni mentre cercavamo di sviluppare un’idea precisa delle previsioni sulle IA trasformative che arricchisse il nostro processo di grantmaking lungoterminista.

Di seguito trovate **un breve riassunto** dei diversi punti di vista sulle previsioni sulle IA trasformative che abbiamo discusso in passato, completo di link che rimandano a discussioni più approfondite in precedenti articoli e ai report tecnici a cui fanno riferimento.

Disclaimer sulla trasparenza: molti di questi report tecnici sono analisi di Open Philanthropy, società di cui sono il co-amministratore.

Detto questo, immagino che alcuni lettori potrebbero sentirsi ancora un po' a disagio. Anche se pensano che le mie tesi siano sensate, potrebbero pensare: **se tutto questo è vero, perché non se ne parla molto di più? Qual è il consenso tra gli esperti?**

Allo stato attuale, riassumerei il consenso tra gli esperti in questo modo:

- Ciò che affermo non è *in contrasto* con quello che pensano gli esperti in generale. (Al contrario, le probabilità che ho stimato non sono troppo lontane da quelle che fanno di solito i ricercatori nel campo dell'intelligenza artificiale, come si vede nella prima riga.) Ci sono tuttavia indizi che portano a credere che non stanno riflettendo abbastanza sulla questione.
- I report tecnici di Open Philanthropy su cui mi baso sono stati analizzati da esperti esterni all'organizzazione. Ricercatori nel campo del machine learning hanno valutato Bio Anchors; Brain Computation è stato esaminato dai neuroscienziati; gli economisti hanno esaminato Explosive Growth; professori e accademici nel campo delle probabilità/incertezza hanno esaminato Semi-informative Priors.[2] (Alcune di queste review hanno sollevato punti su cui si è in disaccordo, ma non ci sono stati casi in cui i report erano direttamente in contrasto con il consenso degli esperti o con la letteratura esistente.)
- Ma gli esperti non hanno ancora raggiunto un consenso chiaro e solido a sostegno di tesi come *"C'è almeno il 10% di probabilità di avere un'IA trasformativa entro il 2036"* o *"C'è una buona probabilità che questo sia il secolo più importante per l'umanità"*, non nel modo in cui c'è un consenso a sostegno, ad esempio, delle azioni volte a intervenire sul cambiamento climatico.

In definitiva, le mie affermazioni riguardano **ambiti per i quali semplicemente non c'è un "pool" di esperti che si dedicano a studiarli. Già di per sé questo è preoccupante** e spero che in futuro la situazione possa cambiare.

Nel frattempo, però, dovremmo concentrarci sulla teoria del "secolo più importante"?

Nel resto dell'articolo vedremo:

> Come potrebbe essere un "ambito di previsioni sulle IA".
>
> Un "punto di vista scettico" secondo il quale le discussioni attuali su questi argomenti sono troppo ristrette, omogenee e isolate (cosa su cui sono d'accordo), motivo per cui non dovremmo concentrarci sulla teoria del "secolo più importante" fino a quando non esisterà un campo di ricerca ben consolidato (su cui non sono d'accordo).
>
> Perché penso che nel frattempo dovremmo prendere in seria considerazione questa teoria, fino a quando (e a meno che) non si crei un tale campo di ricerca:
> - Non possiamo aspettare di avere un consenso solido sulla questione.
> - Se esistono obiezioni valide – o esperti che potrebbero formulare obiezioni valide in futuro – non le abbiamo ancora trovate. Più questa teoria viene presa in considerazione e maggiori sono le probabilità che queste obiezioni verranno formulate in futuro. (Legge di Cunningham: "il modo migliore per ottenere una risposta corretta è formularne una sbagliata".)
> - Penso che continuare a insistere sul consenso degli esperti sia un modo pericoloso di ragionare. Sono dell'idea che sia accettabile correre il rischio di illudersi o isolarsi se ci porta a fare la cosa giusta quando è più necessario.

Che tipo di competenze richiedono le previsioni sulle IA?

Tra le domande analizzate nei report menzionati in precedenza troviamo:

- Le IA stanno sviluppando abilità sempre più notevoli? (IA, storia delle IA)
- Come possiamo confrontare i modelli di IA con il cervello umano/animale? (IA, neuroscienze)
- Come possiamo confrontare le abilità delle IA con quelle degli animali? (IA, etologia)
- In base alle informazioni di cui disponiamo sull'addestramento di precedenti intelligenze artificiali, come possiamo calcolare le spese necessarie ad addestrare un'intelligenza artificiale complessa per un compito difficile? (IA, curve fitting)
- Basandoci esclusivamente sugli anni/ricercatori/soldi impiegati in questo campo fino a ora, in che modo possiamo formulare una valutazione essenziale sulle IA trasformative? (Filosofia, studio delle probabilità)
- In base agli andamenti storici e alle teorie di cui disponiamo, quali sono le probabilità di un boom economico in questo secolo? (Scienze della crescita economica, storia dell'economia)
- Che genere di "hype per le IA" c'è stato in passato? (Storia)

In passato, quando ho parlato delle conseguenze su larga scala delle IA trasformative sul "secolo più importante", ho preso in considerazione domande come "È realistico aspettarsi persone virtuali e la fondazione di colonie spaziali nella galassia?" Questi argomenti riguardano fisica, neuroscienze, ingegneria, filosofia della mente e molto altro.

Non esistono lavori o background precisi che facciano di qualcuno un esperto in grado di dire quando avremo IA trasformative o se questo è il secolo più importante per l'umanità.

(Io in particolare non sono d'accordo con chi afferma che per questo genere di previsioni dovremmo affidarci esclusivamente ai ricercatori nel campo delle intelligenze artificiali. Oltre al fatto che al momento non sembra stiano riflettendo granché sull'argomento, penso che affidarsi a persone che costruiscono modelli di IA sempre più potenti per sapere quando avremo IA trasformative sia come affidarsi a società che sviluppano tecnologie a energia solare – o a compagnie petrolifere, a seconda di come volete vederla – per fare previsioni sulle emissioni di carbonio e il cambiamento climatico. Hanno di sicuro un punto di vista sulla questione, ma fare previsioni è un lavoro diverso dal migliorare o costruire sistemi all'avanguardia.)

Non sono nemmeno sicuro che queste domande siano fatte per la ricerca accademica. Fare previsioni sulle IA trasformative o capire se questo è il secolo più importante sembra più simile:

> Al modello elettorale FiveThirtyEight ("Chi vincerà le elezioni?") piuttosto che a discussioni accademiche di scienze poltiche ("Qual è la relazione tra governi ed elettori?");

> Al trading nei mercati finanziari ("I prezzi si alzeranno o si abbasseranno?" che alle discussioni accademiche di economia ("Perché avvengono le recessioni?");
>
> Alle ricerche di GiveWell ("Quale organizzazione benefica aiuterà più persone con questa somma di denaro?") che alle discussioni accademiche di economia dello sviluppo ("Quali sono le cause della povertà e quali i fattori che la riducono?")[4]

Voglio dire che non mi è chiaro quali caratteristiche dovrebbe avere un'"istituzione" naturale per le competenze necessarie alle previsioni sulle IA trasformative e sul "secolo più importate", ma mi sento di dire che attualmente non esiste nessuna grande istituzione che studia queste tematiche.

Come dovremmo comportarci in mancanza di un consenso omogeneo?

Il punto di vista scettico

Stante la mancanza di un consenso solido tra gli esperti, mi aspetto che alcune (o meglio, molte) persone saranno scettiche a prescindere dal tipo di argomentazione.

Quella che segue è una versione molto generica di una reazione scettica con cui sono solidale:

1. *Mi sembra tutto troppo fantasioso.*

2. *Le tue affermazioni sul vivere nel secolo più importante sono esagerate. È uno* ***schema cognitivo tipico delle illusioni.***

3. *Dici che l'onere della prova non dovrebbe essere così rilevante perché ci sono molti elementi che indicano che stiamo vivendo un periodo eccezionale e incerto. Solo che... non mi ritengo in grado di valutare quelle affermazioni, o le tue affermazioni sulle IA, o qualsiasi altra cosa su questi argomenti assurdi.*

4. *Mi preoccupa il fatto che ci sono poche persone che si occupano di questi temi e quanto* ***ristretto, uniforme e isolato*** *sembra il dibattito. In generale mi sembra che sia una storia che si raccontano quelli intelligenti per convincersi del loro posto nel mondo, con un sacco di grafici e cifre per razionalizzare il tutto. Non sembra "reale".*

5. *Okay, fammi un fischio quando ci sarà un campo di ricerca con magari centinaia o migliaia di esperti che si valutano ed esaminano a vicenda e quando questi avranno raggiunto un qualche tipo di consenso simile a quello che abbiamo per il cambiamento climatico.*

> Capisco perché possiate sentirvi così. Io stesso mi sono sentito così a volte, soprattutto sui punti 1 e 4, ma voglio illustrarvi **tre motivi per cui penso che il punto 5 non sia corretto.**

Motivo n.1: non possiamo permetterci di aspettare che si formi un consenso

La mia paura è che l'avvento delle IA trasformative sia un po' una versione in slow motion e con una posta in gioco più alta della pandemia di COVID-19. Se ci basiamo sulle analisi e sulle

informazioni migliori di cui disponiamo al momento, ci sono buone ragioni per aspettarsi che succeda qualcosa di importante, ma la situazione è decisamente singolare: non può essere catalogata in nessuno degli insiemi di situazioni che le nostre istituzioni affrontano regolarmente. Inoltre, prima cominciamo ad agire è meglio è.

Potremmo anche immaginarla come una versione accelerata delle dinamiche del cambiamento climatico. Pensate se le emissioni di gas serra avessero cominciato ad aumentare solo di recente[5] (invece che a metà Ottocento) e non ci esistesse ancora una branca della scienza che si occupa del clima. Aspettare per decine di anni che nasca una tale branca prima di cercare di ridurre le emissioni sarebbe una pessima idea.

Motivo n.2: La Legge di Cunningham ("il modo migliore per ottenere una risposta corretta è formularne una sbagliata") potrebbe essere il modo migliore per trovare falle in questi ragionamenti

No, sul serio.

Diversi anni fa, io e alcuni miei colleghi avevamo il sentore che la teoria del "secolo più importante" potesse essere corretta, ma prima di concentrarci tutte le nostre energie volevamo vedere se saremmo riusciti a trovarvi degli errori cruciali.

Un modo per descrivere come abbiamo lavorato negli ultimi anni è che **sembrava che stessimo facendo tutto il possibile per capire che la teoria era errata.**

Per prima cosa abbiamo discusso dei temi fondamentali con diverse persone: ricercatori nel campo delle IA, economisti, ecc. Sono sorti alcuni problemi:

> Avevamo solo idee molto vaghe delle argomentazioni in questo campo (perlopiù, o forse del tutto, estrapolate da altre persone). Non eravamo in grado di esporle con il giusto livello di chiarezza e meticolosità.
>
> C'erano un sacco di punti concreti che pensavamo si sarebbero rivelati corretti,[6]ma che non avevamo identificato alla perfezione e che non potevamo presentare perché fossero esaminati.
>
> In generale, non eravamo nemmeno in grado di esporre un caso concreto con sufficiente chiarezza perché gli altri avessero la possibilità di demolirlo.

Ragion per cui abbiamo lavorato a lungo per creare report tecnici su molte delle argomentazioni fondamentali (che sono adesso di pubblico dominio e inclusi all'inizio di questo articolo), cosa che ci ha messo in condizione di pubblicare le argomentazioni e ci ha dato la possibilità di trovare controargomentazioni decisive.

A questo punto abbiamo richiesto review da parte di esperti esterni al nostro gruppo.[7]

Limitandoci alle mie ipotesi, sembra che la teoria del "secolo più importante" abbiamo superato tutti i test. Dopo averla esaminata da ogni angolazione ed essere entrato più nei dettagli, infatti, credo ancora più fermamente che sia corretta.

Ma d'accordo, diciamo che è solo perché secondo i *veri* esperti – persone che non abbiamo ancora scovato e che hanno controargomentazioni potentissime – tutta questa faccenda è così

stupida che non perdono nemmeno tempo a esaminarla. Oppure che ci attualmente persone che *in futuro* potrebbero diventare esperti di queste materie e demolire queste argomentazioni. Cosa potremmo fare perché si realizzi questa situazione?
La risposta migliore che ho trovato è: "Se questa teoria fosse più conosciuta, più accettata e più influente, allora verrebbe anche esaminata più spesso."

Questa serie è un tentativo di andare in quella direzione, di portare maggiore credibilità alla teoria del "secolo più importante". Sarebbe un'ottima cosa se questa teoria si rivelasse corretta; sarebbe anche il passo successivo più logico se il mio obiettivo fosse quello di mettere in discussione le mie credenze e scoprire che è sbagliata.
Ovviamente non sto dicendo che dovete accettare o promuovere la teoria del "secolo più importante" se non vi sembra corretta, ma penso che se il vostro *unico* dubbio riguarda la mancanza di un consenso diffuso, sembra un po' strano continuare a ignorare la situazione. Se ci comportassimo sempre così (ignorando qualsiasi teoria che non è sostenuta da un consenso diffuso), non credo che vedremmo mai una sola teoria – anche quelle corrette – passare dall'essere di nicchia all'essere accettata.

Motivo n.3: in generale, lo scetticismo non sembra una buona idea

Quando lavoravo a GiveWell, le persone ogni tanto mi dicevo cose del tipo: "non puoi mica sottoporre ogni argomentazione agli stessi standard di qualità che GiveWell usa per valutare le organizzazioni benefiche – test randomizzati controllati, solide basi empiriche, ecc. Alcune delle migliori occasioni per fare del bene saranno per forza quelle meno evidenti, perciò c'è il rischio che questi standard escludano alcune delle più grandi occasioni per avere un impatto positivo."

Sono convinto che sia così. Penso che sia importante controllare il proprio approccio al ragionamento e agli standard di evidenze scientifiche e chiedersi: "In quali situazioni questo metodo fallirebbe e in quali preferirei che avesse successo?" Per quel che mi riguarda, **è accettabile correre il rischio di illudersi o isolarsi se ci porta a fare la cosa giusta quando è più necessario.**

Penso che la mancanza di consenso tra gli esperti – e il timore di illudersi o isolarsi – siano buone ragioni per *indagare a fondo* la teoria del "secolo più importante" piuttosto che accettarla all'istante. Per chiedersi se ci possano essere errori non ancora individuati, per cercare bias che potrebbero esagerare la nostra importanza, per andare alla ricerca di quegli aspetti dell'argomentazione che sembrano più discutibili, ecc.

Se però avete esaminato la questione a un livello che vi sembra accettabile/fattibile – e non avete trovato altri difetti *se non* considerazioni del tipo "non c'è consenso diffuso" e "mi preoccupa la possibilità di illudermi o isolarmi" – allora direi che scartare quest'ipotesi **farà sì che non sarete tra i primi a rendervi conto di e ad agire su un problema estremamente importante se se ne presenta l'occasione**. Per quel che mi riguarda, se penso alle occasioni sprecate per fare del bene, è un sacrificio troppo grande.

1. Tecnicamente queste probabilità sono calcolate per "intelligenze artificiali di livello umano". In generale il grafico semplifica la questione, perché presenta un unico insieme di probabilità. In generale tutte queste probabilità fanno riferimento a qualcosa le cui capacità sono *almeno* allo stesso livello di quelle di un sistema PASTA, di conseguenza dovrebbero essere stime al ribasso della probabilità di un sistema PASTA (ma non penso che sia un grande problema).
2. Qui potete trovare review di Bio Anchors; qui review di Explosive Growth; qui review di Semi Informative Priors. Brain Computation era stato esaminato quando non avevamo ancora ideato il processo che avrebbe portato a pubblicare review, ma qui potete trovare più di 20 conversazioni con esperti che hanno costituito il report. Human Trajectory non è stato esaminato, anche se molto delle analisi e delle conclusioni di quel report sono contenute in Explosive Growth.
3. Le branche accademiche sono piuttosto ampie. Questi sono solo esempi delle domande che affrontano.
4. Anche se la scienza del clima è un buon esempio di ambito accademico in cui si investe molto nel prevedere il futuro.
5. Il campo delle intelligenze artificiali esiste dal 1956, ma i modelli di machine learning hanno cominciato ad avvicinarsi alle dimensioni del cervello degli insetti e ad avere buone prestazioni in attività complesse solo negli ultimi dieci anni. ↩
6. Spesso ci basavamo solo sulle impressioni che avevamo di quello che altri più esperti pensavano dell'argomento.
7. Qui potete trovare review di Bio Anchors; qui review di Explosive Growth; qui review di Semi Informative Priors. Brain Computation era stato esaminato quando non avevamo ancora ideato il processo che avrebbe portato a pubblicare review, ma qui potete trovare più di 20 conversazioni con esperti che hanno costituito il report. Human Trajectory non è stato esaminato, anche se molto delle analisi e delle conclusioni di quel report sono contenute in Explosive Growth.

Strategie per ridurre i rischi derivanti dalle IA

[OPZIONALE] Il panorama della governance lungoterminista delle intelligenze artificiali

di Sam Clarke – 18 gennaio 2022 – 11 minuti di lettura

Obiettivo: fornire una visione d'insieme della governance delle IA dal punto di vista lungoterminista.
A chi è destinato: a persone che non hanno familiarità con la governance delle IA dal punto di vista lungoterminista e vogliono conoscere meglio questo ambito. Non mi aspetto che sarà utile a chi ha già familiarità con queste tematiche. Addendum: alcune persone che avevano già familiarità con queste tematiche hanno trovato utile l'articolo.

Questo post intende delineare i diversi tipi di lavoro coinvolto nella governance lungoterminista delle IA. Darò una breve spiegazione di ognuno, completa di esempi e storie che spiegano in che modo potrebbe avere un impatto positivo, oltre a un elenco delle persone che ci stanno attualmente lavorando di cui sono a conoscenza.[1]
Per prima cosa, un paio di definizioni:

- *Governance delle IA* lo sviluppo di norme, politiche, leggi, processi e istituzioni (non solo governi) locali e globali che influenzeranno i risultati nella società derivanti dallo sviluppo e dall'impiego delle intelligenze artificiali.[1]

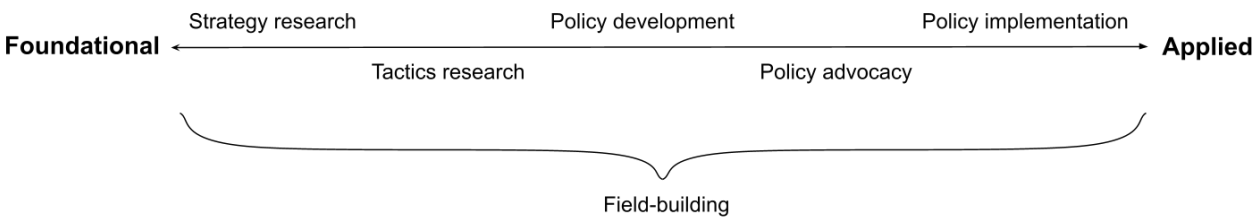
- *La governance lungoterminista delle IA*, nello specifico, è la sottocategoria motivata dalle preoccupazioni per l'impatto delle IA su un periodo di tempo molto ampio. Coincide spesso con il lavoro che punta a controllare le IA trasformative (TAI).

Vale la pena notare che si tratta di un ambito molto ristretto. Se dovessi ipotizzare direi che ci sono circa 60 persone che lavorano nel campo della governance delle IA e la cui preoccupazione è il suo impatto sul lunghissimo periodo.

Breve riassunto

Parlando in astratto, penso sia utile visualizzare il lavoro di base e quello applicato come ai poli opposti di uno spettro. Dal lato del lavoro di base abbiamo la *ricerca strategica*, il cui scopo è individuare obiettivi astratti ideali per la governance lungoterminista delle IA; c'è poi la *ricerca tattica*, il cui obiettivo è individuare i programmi utili a raggiungere quegli obiettivi. Dal lato del lavoro applicato troviamo lo sviluppo di normative, il cui scopo è trasformare la ricerca in normative concrete; abbiamo poi lavoro che spinge perché quelle normative vengano applicate e infine l'applicazione vera e propria (ad esempio da parte di impiegati statali).

C'è anche lavoro di *field-building* (dalla collocazione incerta): invece che contribuire direttamente a risolvere il problema, l'obiettivo è quello di creare un ambito con persone che lavorano sul tema.



Ovviamente si tratta di una semplificazione e non tutti i lavori potranno essere catalogati con precisione. Al contrario di quel che si pensa di solito, non sempre la conoscenza profonda passa dalla base alla parte applicata: è anche importante che la ricerca tenga conto delle preoccupazioni normative, vale a dire quanto è probabile che questa ricerca corrobori una proposta normativa che è fattibile dal punto di vista politico.

Vediamo ora nel dettaglio ognuno di questi ambiti.

Ricerca

Ricerca strategica

In definitiva, l'obiettivo della *ricerca strategica lungoterminista sulle IA* è quello di individuare obiettivi di alto livello che, se raggiunti, potrebbero chiaramente aumentare le probabilità di risultati positivi, da un punto di vista lungoterminista, da parte di un'IA avanzata (sull'esempio di Muehlhauser, a volte farò riferimento a questo obiettivo come al "raggiungimento di una *chiarezza strategica*").
La ricerca stessa può essere collocate in vari punti dello spettro compresi tra *mirata* ed *esplorativa*:

- L'obiettivo della ricerca strategica mirata è quello di fornire risposte che facciano luce su altre domande specifiche e importanti.
    - Ad esempio, "Voglio scoprire qual è la potenza di calcolo del cervello umano, perché mi aiuterà a scoprire quando verranno sviluppate le TAI (cosa che determina gli obiettivi di alto livello che dovremmo ricercare)"
- La ricerca strategica esplorativa fornisce risposte senza avere bene in mente le altre domande importanti a cui potrebbe aiutarci a rispondere.
    - Ad esempio, "Voglio scoprire quali sono le politiche industriali della Cina, perché probabilmente saperlo mi aiuterà a rispondere a una serie di domande strategiche importanti, anche se non so esattamente a quali".

Esempi

- Il lavoro delle previsioni sulle TAI, come ad esempio ancore biologiche e leggi scalabili per modelli linguistici neurali.
    - Esempio di importanza strategica: se le TAI arriveranno in tempi brevi, allora un ambito di esperti che cresce lentamente sembra poco promettente; se le TAI sono al di là da venire, allora alla governance lungoterminista delle IA potrebbe essere assegnata una priorità minore.
- Il lavoro sulla chiarezza delle fonti dei rischi esistenziali legati alle IA, come gli scritti di Christiano, Critch, Carlsmith, Ngo e Garfinkel.
    - Esempio di importanza strategica: se la maggior parte dei rischi esistenziali derivanti dalle IA proviene da IA avanzate non allineate, allora la governance dovrebbe concentrarsi su chi costruisce quelle IA.

- Le ricerche che indagano la velocità dei progressi sulle TAI, ad esempio le indagini e le analisi di AI Impacts.
    - Esempio di importanza strategica: se i progressi nelle IA sono discontinui, allora è possibile che ci sia un numero ristretto di protagonisti molto importanti e che la maggior parte del valore della governance risieda nell'influenzare quelle persone.

È facile confondere la ricerca strategica (e soprattutto la ricerca strategica *esplorativa*) con la ricerca *ad ampio spettro*. Come mostrano molti degli esempi precedenti, la ricerca strategica può essere *obiettivi ristretti*, vale a dire che può rispondere a una domanda dalla portata limitata. Esempi di portata limitata e ampia portata:

- Sulle leggi di scala (scaling laws):
    - Ampia portata: in generale, come cambiano le prestazioni dei modelli di deep learning se aumentano le dimensioni di quei modelli?
    - Portata limitata: in che modo cambiano *in particolare* le prestazioni di *modelli linguistici di grandi dimensioni* (come ad esempio GPT-3) se aumentano le dimensioni di quei modelli? (Argomento affrontato in questo paper.)
- Sulle fonti dei rischi esistenziali derivanti dalle IA:
    - Ampia portata: in generale, quali rischi esistenziali comportano le IA?
    - Portata limitata: quali rischi esistenziali potrebbero derivare *nello specifico da IA il cui obiettivo è ottenere influenza*? (Argomento affrontato in questo report.)

In effetti, spesso è meglio scegliere domande dalla portata limitata, soprattutto se si è ricercatori di basso rango, perché sono di solito più gestibili.

Luke Muehlhauser ha alcuni consigli per coloro che vogliono cimentarsi in questo tipo di lavoro al punto 4 di questo post. In questo post invece ci sono esempi di domande aperte sulla ricerca.[3]

Storie di impatto

- *Impatto diretto*: gli obiettivi nel campo della governance delle IA sono molteplici ed è necessario dare priorità a quelli più importanti. Questo tipo di lavoro è spesso motivato dall'impressione dei ricercatori che c'è molta poca chiarezza sui temi che influenzano gli obiettivi che dovremmo raggiungere. Si vedano ad esempio i risultati di questi sondaggi, che mostrano come ci sia parecchio disaccordo rispettivamente sugli scenari di rischi esistenziali derivanti dalle IA e sul numero di questi rischi.
- *Impatto indiretto:*
    - Field-building: avere un'idea precisa di ciò che si sta cercando di raggiungere e della sua importanza sarebbe utile per attrarre nuovi ricercatori in quel campo.
    - Comunicare la necessità di nuove normative: se l'obiettivo è convincere le persone a prendere decisioni costose o drammatiche in futuro, è meglio sapere con esattezza cosa dire a proposito degli obiettivi che si sta cercando di raggiungere e della loro importanza.

Chi ci sta lavorando?

Alcune persone nelle seguenti organizzazioni: FHI, GovAI, CSER, DeepMind, OpenAI, GCRI, CLR, Rethink Priorities, OpenPhil, CSET,[4] oltre ad alcuni accademici indipendenti.

Ricerca tattica

L'obiettivo della *ricerca tattica lungoterminista sulle IA* è quello di individuare i programmi utili a raggiungere obiettivi di alto livello (a cui la ricerca strategica ha assegnato la priorità). Di solito la sua portata è più limitata per natura.

Vale la pena notare che può essere conveniente fare ricerca tattica anche se non si sono individuati con precisione degli obiettivi prioritari: per ricerca personale, per la propria carriera e per creare un campo di ricerca.

Esempi

- Windfall Clause
  - Il programma: sviluppare uno strumento per la redistribuzione dei vantaggi delle IA per il bene comune
  - Obiettivi di alto livello di questo programma: fare in modo che i protagonisti di questa ricerca non competano gli uni con gli altri per sviluppare per primi un'IA avanzata; ridurre la disparità economica.
- Meccanismi di sostegno di affermazioni verificabili
  - Il programma: sviluppare pratiche con cui gli sviluppatori di IA possono rendere più verificabili le loro affermazioni riguardo lo sviluppo delle IA (affermazioni di cui gli sviluppatori possano essere ritenuti responsabili)
  - Obiettivi di alto livello di questo programma: sviluppare meccanismi con cui è possibile dimostrare i comportamenti responsabili delle intelligenze artificiali; permettere un controllo più efficace; ridurre la tentazione di prendere scorciatoie per ottenere un vantaggio competitivo.
- IA e Antitrust
  - Il programma: proporre modi per ridurre le tensioni tra le leggi antitrust e la necessità di uno sviluppo in cooperazione delle IA
  - Obiettivi di alto livello di questo programma: aumentare il livello di cooperazione tra le società che sviluppano IA avanzate.
-

Storie di impatto

- *Impatto diretto*: creare soluzioni che vengano utilizzate per prendere decisioni migliori (nelle normative e nelle ricerche future).
  - Quello che Allan Dafoe chiama "modello prodotto di ricerca".
- *Impatto indiretto*: anche se non tutte le soluzioni verranno usate per prendere decisioni migliori, saranno utili per la crescita del gruppo di persone che hanno a cuore i problemi della governance lungoterminista delle IA, oltre a migliorare le conoscenze, le competenze, le connessioni e la credibilità dei ricercatori.
  - Quello che Allan Dafoe chiama "Modello di field-building della ricerca".

Chi ci sta lavorando?

Alcune persone nelle seguenti organizzazioni: FHI, GovAI, CSER, DeepMind, OpenAI, GCRI, CSET, Rethink Priorities, LPP, oltre ad alcuni accademici indipendenti.

Sviluppo, sostegno e attuazione delle politiche

La ricerca strategica produce obiettivi di alto livello. La ricerca tattica prende quegli obiettivi e produce dei programmi con cui raggiungerli. *Lo sviluppo di politiche* prende quei programmi e li trasforma in consigli pronti per essere consegnati ai decisori politici. Per fare ciò è necessario avere ben chiaro (ad esempio) quali domande fare di preciso, che tipo di

linguaggio usare (sia nelle politiche formali che nella richiesta) e altre caratteristiche legate al contesto che influiranno sulle probabilità di successo.

Il sostegno alle politiche spinge affinché quelle politiche vengano attuate, ad esempio quali sono le persone più adatte per proporle, a chi e in quale momento.

L'attuazione delle politiche consiste nel mettere in pratica quelle politiche grazie al lavoro di impiegati statali o corporazioni.

È bene distinguere tra le politiche governative (ovvero quelle politiche pensate per essere attuate da governi od organizzazioni intergovernative) e le politiche aziendali (ovvero le politiche pensate per essere adottate dalle compagnie). Alcune persone che lavorano nel campo della governance lungoterminista delle IA si concentrano sul migliorare le politiche aziendali (in particolar modo quelle che riguardano gli sviluppatori di IA), mentre altri si concentrano nel migliorare quelle dei governi di pertinenza.

Un fattore motivante comune è che le specifiche del processo di attuazione sono spesso considerate fondamentali per il successo della stessa attuazione. Ad esempio, se una norma governativa presenta una falla, l'intera norma potrebbe essere inutile.

Rispetto alla ricerca, questo tipo di lavoro di solito si basa meno sui ragionamenti individuali e più sulla raccolta di informazioni/conversazioni (ad esempio riunioni per stabilire chi ha autorità su una determinata norma, cosa gli sta a cuore e quali sono gli interessi di altre figure rilevanti) e sul coordinamento (ad esempio stabilire in che modo convincere un determinato gruppo di persone a sostenere una data norma).

Come detto in precedenza, la conoscenza profonda a volte scorre "all'indietro". Ad esempio, la formulazione di una precisa politica potrebbe essere ripetuta in base a come il sostegno datole modifica le proprie conoscenze (e l'ambiente in cui ci si muove).

Esempi

- Politiche governative:
    - Impegnarsi a non incorporare intelligenze artificiali nel comando, controllo e comunicazione in ambito nucleare (NC3), raccomandato ad esempio dal CLTR in Future Proof.
    - Monitoraggio dello sviluppo di intelligenze artificiali da parte del governo, come elaborato in questo libro bianco sul monitoraggio delle IA.
    - Fare in modo che nuove norme o strategie/principi sulle IA siano sensibili ai rischi derivanti da intelligenze artificiali avanzate (oltre che quelle attuali), come il feedback fornito da diverse organizzazioni AE sulla proposta di legge dell'UE sulle intelligenze artificiali.
- Politiche aziendali:
    - Sviluppare norme per la distribuzione responsabile delle ricerche sulle IA, considerati i possibili usi impropri, come queste raccomandazioni del PAI.

La posizione di queste idee sullo spettro va da quelle più mirate (come ad esempio non integrare le IA all'interno dell'NC3) a quelle più generali (nel senso di creare una capacità generale per far fronte a un'ampia gamma di problemi che potrebbero sorgere, ad esempio la maggior parte di quelli visti in precedenza). Sono dell'opinione che allo stato attuale le nostre pratiche di sviluppo, sostegno e attuazione delle politiche debbano concentrarsi soprattutto sulle idee più generali, considerati tutti i dubbi sul futuro delle IA (e al tempo stesso spingere per l'attuazione di buone idee mirate quando se ne presenta l'occasione).

Storie d'impatto

- *Impatto diretto*: poter disporre di buone norme aumenta le probabilità di farsi strada con successo nella transizione verso un mondo con IA avanzate.

- *Impatto indiretto*: anche se non si è sicuri della robustezza di certe idee, svilupparle/sostenerle/attuarle permetterà a chi si occupa di governance lungoterminista delle IA di ottenere conoscenze, competenze, connessioni e credibilità.

    - Detto questo, dovremmo anche fare molta attenzione a non attuare politiche che potrebbero rivelarsi dannose, ad esempio limitando lo sviluppo di politiche future.

Chi ci sta lavorando?

- Sviluppo:
    - Di politiche governative: CLTR, FLI, GovAI, CSET, CSER, FHI, TFS
    - Di politiche aziendali: OpenAI, DeepMind, GovAI, CSER, FHI, PAI
- Sostegno:
    - A politiche governative: CLTR, CSET, FLI, TFS
    - A politiche aziendali: PAI
- Attuazione:
    - Di politiche governative: diverse persone in ambito statale
    - Di politiche aziendali: OpenAI, DeepMind

Field-building

L'obiettivo dichiarato di questo lavoro è creare un campo o una comunità di persone che svolgano lavoro di una certa importanza nell'ambito della governance lungoterminista delle IA. [6] Possiamo vederlo come un lavoro in cui è necessario sia (1) coinvolgere persone nuove e (2) rendere questo campo più efficace.

Esempi

1. Coinvolgere nuove persone creando:
    - borse di studio, come la OpenPhil Technology Policy Fellowship;
    - programmi o corsi online che aiutino i novizi ad aggiornarsi sulle ultime novità nel campo della governance delle IA;
    - materiali introduttivi di ampio respiro e di alta qualità pensati per gli studenti universitari;
    - finanziamenti di ricerca più scalabili per connettersi, sostenere e dare credibilità alle nuove leve.
2. Rende questo campo più efficace creando:
    - programmi di ricerca;
    - opportunità per i ricercatori di lunga data di assumere facilmente assistenti di ricerca.[7]

Storie d'impatto

- *Modello di crescita:* creare un campo di ricerca sulla governance lungoterminista delle IA con molte persone allineate con le competenze e capacità giuste per condurre ricerche importanti e lavoro di policy (soprattutto quando questo tipo di lavoro è meno limitato da una mancanza di chiarezza strategica).

- *Modello Metropoli:*[8] creare un campo di ricerca sulla governance lungoterminista delle IA che abbia connessioni solide con comunità più grandi (ad esempio policymaking, scienze sociali, machine learning), in modo che questo campo possa contare sulle diverse competenze di queste comunità.

Chi ci sta lavorando?

GovAI, OpenPhil, SERI, CERI, CHERI e EA Cambridge. Più in generale, tutti i movimenti interni all'Altruismo Efficace che si occupano di cause generali. Tra tutti i tipi di interventi discussi in questo post, questo è quello meno considerato.

Ulteriori punti di vista sulla governance lungoterminista delle IA

Finora mi sono limitato a presentare un unico punto di vista della governance lungoterminista delle IA. Ovviamente ce ne sono altri, che potrebbero essere più utili per altri scopi. Ad esempio, si potrebbe dividere il panorama attuale in base ai diversi tipi di intervento:

- Intervenire sugli attuali dibattiti nell'ambito delle policy per far sì che siano più sensibili ai rischi esistenziali derivanti dalle IA (ad esempio maggiore consapevolezza di quanto può essere difficile creare intelligenze artificiali all'avanguardia)
- Proporre nuovi strumenti di policy (ad esempio, standard internazionali per le IA)
- Convincere i governi a finanziare la ricerca sulla sicurezza delle IA
- Portare a cambiamenti in ambito aziendale (Windfall Clause)
- ...

Oppure lo si potrebbe dividere in base all'area geografica (anche se non tutte le organizzazioni fanno parte di una determinata area):

- San Francisco Bay Area: OpenPhil, OpenAI, PAI, diverse organizzazioni che lavorano all'allineamento delle IA. Generalmente più concentrate sul non-allineamento come fonte di possibili rischi esistenziali; simili per cultura alla Silicon Valley e alle comunità razionaliste.
- Washington DC: il governo statunitense, CSET. Si concentrano sullo sviluppo/sostegno/attuazione di politiche statunitensi; simili per cultura alla cultura di DC.
- Regno Unito: FHI/GovAI, DeepMind, governo britannico, CSER, CLTR, (altre?). Di solito più concentrate sulle molteplici fonti di rischio esistenziale derivante dalle IA.
- Unione Europea: nel 2020 la Commissione Europea ha stilato la bozza della prima normativa sulle intelligenze artificiali del mondo, che con ogni probabilità verrà approvata nei prossimi anni e potrebbe generare un "effetto Bruxelles".
- Cina.
- …

Oppure, si potrebbe dividere in base a diverse "teorie di vittoria", vale a dire storie su come l'umanità giungerà con successo a un mondo con IA avanzate. C'è molto altro che si potrebbe dire sull'argomento. Lo scopo di questo post era solo quello di fornire una breve panoramica dei diversi lavori attualmente in corso.

Ringraziamenti: questo post è il risultato di una mia sintesi personale del panorama attuale, ma si ispira e/o estrapola direttamente da altri post sul forum di Altruismo Efficace di Allan Dafoe, Luke Muehlhauser, Convergence Analysis. Grazie anche a Jess Whittlestone per le preziose

conversazioni, a Matthijs Maas, Yun Gu, Konstantin Pilz, Caroline Baumöhl e soprattutto a un critico del SERI per il feedback ricevuto sulla prima bozza.

1. Sicuramente ho dimenticato qualche gruppo importante e potrei aver sbagliato a classificarli o anche averli rappresentati male – se è così, fatemelo presente e correggerò!
2. Preso direttamente dalla definizione data da Open Philanthropy.
3. N.B.: alcune di queste domande riguardano la *ricerca tattica* piuttosto che quella strategica.
4. CSET si occupa perlopiù di ricerca tattica e sviluppo e sostegno di politiche, ma il loro lavoro nel mappare il processo di produzione dei semiconduttori rientra nella ricerca strategica.
5. Muehlhauser lo definisce "un periodo compreso tra 1 e 20 anni in cui potrebbero essere prese le decisioni più d'impatto sulle TAI".
6. Distinto dai vantaggi di field-building degli altri tipi di lavoro esaminati perché questo è *esplicitamente e unicamente* concentrato sulla creazione di un campo di ricerca.
7. Che potrebbe a sua volta aiutare a coinvolgere persone nuove.
8. Idea estrapolata da questo post di Allan Dafoe.

[OPZIONALE] Ricerca sulla sicurezza delle IA: panoramica delle carriere

di Benjamin_Todd – 23 novembre 2021

Originale disponibile qui: https://80000hours.org/career-reviews/ai-safety-researcher/.

[Tempo di lettura: 7 minuti]

In breve: Per mitigare i rischi posti dallo sviluppo dell'intelligenza artificiale, è necessario fare ricerca su come risolvere le sfide in ambito tecnico e i problemi in ambito di progettazione per assicurarci che le intelligenze artificiali più potenti facciano ciò che vogliamo – e che portino benefici – senza nessuna conseguenza catastrofica involontaria.

Consigliata: Se si è individui molto adatti a questo tipo di carriera, potrebbe essere il modo migliore per avere un impatto sociale.

Stato della recensione: Basata su studi di media profondità

Perché lavorare nella ricerca sulla sicurezza delle IA potrebbe avere un grande impatto?

Come già argomentato, nei prossimi decenni potremmo assistere allo sviluppo di potenti algoritmi di machine learning con la capacità di trasformare la nostra società. Ciò comporterebbe importanti vantaggi e svantaggi, inclusa la possibilità di un rischio catastrofico.

Oltre al lavoro sulle politiche e sulle strategie discusse in questa recensione di carriera, un'altra strada maestra per limitare questi rischi consiste nel fare ricerca sulle sfide tecniche poste da intelligenze artificiali avanzate, come il problema dell'allineamento. In breve, come possiamo progettare IA potenti in modo che facciano ciò che vogliamo senza conseguenze indesiderate?

Questo ambito di ricerca ha iniziato a decollare. Esistono oggi importanti poli accademici e laboratori di IA dove si può lavorare su questi problemi come il Mila di Montreal, il Future of Humanity Institute di Oxford, il Center for Human-Compatible Artificial Intelligence di Berkeley, DeepMind a London e l'OpenAI a San Francisco. Abbiamo svolto attività di consulenza per oltre 100 persone in questo ambito, con molte di esse che già lavorano nelle istituzioni summenzionate. Il Machine Intelligence Research Institute di Berkeley lavora in questo campo dal 2005 e possiede una prospettiva e un programma di ricerca non convenzionali rispetto agli altri laboratori.

Abbondano i finanziamenti per i ricercatori talentuosi, incluse borse di studio universitarie e donazioni filantropiche dai principali sovvenzionatori come Open Philanthropy. È possibile inoltre ottenere finanziamenti per il proprio progetto di dottorato. La principale necessità di questo campo consiste in più persone capaci di impiegare questi fondi per portare avanti la ricerca.

Che cosa comporta questo percorso?

L'obiettivo in questo percorso sarebbe quello di ottenere un lavoro in uno dei migliori centri per la sicurezza delle IA – nel profit, nel no-profit o in ambito accademico – e poi lavorare sulle questioni più urgenti, con l'eventuale prospettiva di diventare un coordinatore dei ricercatori che supervisiona la ricerca sulla sicurezza.

In generale, le posizioni tecniche sulla sicurezza delle IA si possono distinguere in ruoli di (i) ricerca e di (i) ambito ingegneristico. I ricercatori guidano progetti di ricerca. Gli ingegneri implementano i sistemi e compiono le analisi necessarie per portare avanti la ricerca.

Anche se gli ingegneri hanno minore influenza sugli obiettivi di alto livello della ricerca, è comunque importante che si occupino della sicurezza, in modo che comprendano meglio gli obiettivi finali della ricerca stessa (e in questo modo darsi meglio le priorità), siano più motivati, orientino la cultura dominante verso la sicurezza e usino il capitale di carriera guadagnato per contribuire a altri progetti futuri sulla sicurezza. Questo significa che l'ambito ingegneristico è una buona alternativa per coloro che non vogliono essere diventare ricercatori.

Può essere utile che ci siano persone che capiscono le sfide poste dalla sicurezza delle IA che lavorino in squadre di ricerca in questo ambito che non siano direttamente orientate alla sicurezza. Lavorare in queste squadre può metterti nella posizione di aiutare a promuovere la preoccupazione sulla sicurezza in generale, soprattutto se finirai per ricoprire una posizione dirigenziale con influenza sulle priorità dell'organizzazione per cui lavori.

Saremmo entusiasti di vedere più persone che sviluppano competenze nel lavoro sulla sicurezza delle IA in Cina o in contesti legati alla Cina – leggi di più nella nostra recensione dei percorsi di carriera sulla sicurezza e sulla gestione dell'IA in contesti legati alla Cina, alcuni dei quali si compiono nella ricerca tecnica.

Esempi di persone che hanno intrapreso questo percorso di carriera

Catherine Olsson ha iniziato il suo dottorato all'Università di New York, lavorando sui modelli computazionali della visione umana. Alla fine ha deciso di lavorare direttamente sulla sicurezza delle IA e ha ottenuto un impiego alla OpenAI, seguito da uno a Google Brain, per poi spostarsi ad Anthropic.
SCOPRI DI PIÙ

Daniel Ziegler dopo aver abbandonato il suo dottorato sul machine learning a Stanford, che ha sempre adorato costruire cose e ha sempre sognato di definire lo sviluppo delle IA, ha fatto domanda per lavorare alla OpenAI. Ha impiegato sei settimane a prepararsi per l'intervista e ha ottenuto il lavoro. Il suo dottorato, invece, lo avrebbe impegnato per sei anni. Daniel pensa che il suo grande balzo di carriera possa essere possibile per molte altre persone.
SCOPRI DI PIÙ

Chris Olah ha avuto un percorso affascinante e non convenzionale. Chris non solo non possiede un dottorato di ricerca, ma non ha ottenuto nemmeno una laurea triennale. Dopo aver abbandonato l'università per aiutare un suo conoscente a difendersi da false accuse penali, Chris ha iniziato a lavorare da sé alla ricerca sul machine learning, ottenendo infine un tirocinio presso Google Brain.
SCOPRI DI PIÙ

Come valutare la tua attitudine

La ricerca tecnica con l'impatto maggiore sulla sicurezza delle IA sarà svolta da persone nei migliori ruoli summenzionati. Quindi, per decidere se questo percorso faccia al caso tuo, è importante chiederti se tu abbia una ragionevole probabilità di ottenere quei lavori.

- Hai la possibilità di essere ammesso in una delle cinque migliori scuole di specializzazione sul machine learning? Questo potrebbe essere un buon test per capire se

potrai ottenere un lavoro nei migliori poli di ricerca sulla IA, anche se non si tratta di un prerequisito.
- Sei sicuro dell'importanza della sicurezza della IA sul lungo periodo?
- Sei un tecnico informatico in ambito software o machine learning che ha lavorato in FAANG o in altre aziende competitive? Potresti formarti per ottenere una posizione di ricerca oppure una posizione in ambito tecnico.
- Hai la possibilità di dare un contributo a un rilevante quesito di ricerca? Per esempio, hai un grande interesse per l'argomento, hai grandi idee su grandi questioni da analizzare e non puoi fare a meno di perseguire tali idee? Leggi di più su come capire se sei un buon candidato per un ruolo nella ricerca.

Come entrare in questo campo

Il primo passo in questo percorso di solito consiste nell'intraprendere un dottorato di ricerca in machine learning in una buona scuola. È possibile entrare in questo campo anche senza un dottorato, ma è probabile che venga richiesto per ruoli di ricercatore nei poli universitari e in DeepMind, i quali coprono una buona fetta delle migliori posizioni. Un dottorato di ricerca in machine learning apre anche strade nelle politiche sulle IA, nelle IA applicate e nel guadagnare per donare, quindi questo percorso ha buone opzioni di riserva se dovessi decidere che la sicurezza delle IA non fa per te.

Comunque, se preferisci l'ambito tecnico alla ricerca, un dottorato non è necessario. Puoi invece seguire un *master* o sviluppare competenze nel profit.

È anche possibile iniziare questo percorso a partire dalle neuroscienze (soprattutto quelle computazionali), quindi, se possiedi già esperienza in quest'area, non è detto che dovrai tornare a studiare.

Se hai già molta familiarità con la sicurezza delle IA come area critica, il nostro miglior consiglio è di dare un'occhiata a questa guida passo-dopo-passo per intraprendere una carriera nella sicurezza tecnica delle IA scritta da Charlie Rogers-Smith.

Ultimamente si sono concretizzate opportunità anche per scienziati sociali che contribuiscano alla sicurezza delle IA.

Puoi trovare ulteriori dettagli nelle risorse al termine della recensione.

Organizzazioni consigliate

- AI Safety Support lavora per ridurre il rischio esistenziale e catastrofico legato alle IA sostenendo chiunque voglia lavorare su questo problema, concentrandosi sull'aiutare nuovi o aspiranti ricercatori sulla sicurezza delle IA attraverso consigli di carriera e costruendo una comunità.
- Alignment Research Center è un'organizzazione di ricerca no-profit al lavoro per allineare i futuri sistemi di machine learning agli interessi umani. Si impegna attualmente a sviluppare una strategia di allineamento "end-to-end" che possa essere adottata oggi dal contesto profit e che, nel frattempo, possa scalare verso i futuri sistemi di apprendimento automatico. Controlla i ruoli attualmente vacanti.
- Anthropic è una compagnia che si occupa della ricerca sulle IA e della sicurezza al fine di costruire sistemi di IA affidabili, interpretabili e manovrabili. Gli interessi del loro team di ricerca multidisciplinare includono il linguaggio naturale, il feedback umano, le leggi di potenza, l'apprendimento attraverso il rinforzo, la generazione di codice e l'interpretabilità. Controlla i ruoli attualmente vacanti.

- Il Center for Human-Compatible Artificial Intelligence punta a sviluppare i mezzi concettuali e tecnici per riorientare la spinta generale della ricerca sull'IA verso sistemi con vantaggi comprovati. Controlla i ruoli attualmente vacanti.
- Il Center on Long-term Risk affronta i rischi peggiori per lo sviluppo e l'impiego di IA avanzate. Attualmente si concentra sugli scenari di conflitto, così come su sugli aspetti tecnici e filosofici della cooperazione. Il loro lavoro include ricerche interdisciplinari, il fare finanziamenti o il suggerire candidati e il costruire una comunità di professionisti e altri ricercatori nell'ambito di queste priorità. Controlla i ruoli attualmente vacanti.
- DeepMind è con ogni probabilità il più grande gruppo di ricerca che sta sviluppando un'intelligenza artificiale generale nel mondo occidentale. Siamo sicuri di poter consigliare ruoli presso DeepMind solamente negli ambiti della sicurezza, dell'etica, delle politiche e della sorveglianza. Controlla i ruoli attualmente vacanti.
- Il Future of Humanity Institute è un istituto di ricerca multidisciplinare dell'Università di Oxford. Accademici del FHI sfoderano gli strumenti della matematica, della filosofia e delle scienze sociali per influenzare le domande fondamentali sull'umanità e le sue prospettive.
- Il Machine Intelligence Research Institute è stato uno dei primi gruppi a preoccuparsi dei rischi dell'intelligenza artificiale nei primi anni 2000 e ha pubblicato diversi articoli sui problemi di sicurezza e su come risolverli. Controlla i ruoli attualmente vacanti.
- OpenAI è stata fondata nel 2015 con l'obiettivo di condurre ricerche su come rendere sicure le IA. Ha ricevuto oltre un miliardo di dollari di impegni di finanziamento dalla comunità di questo settore tecnologico. Controlla i ruoli attualmente vacanti.
- Redwood Research conduce ricerche applicate per aiutare ad allineare i futuri sistemi di IA agli interessi umani. Controlla i ruoli attualmente vacanti.

Vuoi dei consigli faccia-a-faccia su come intraprendere questo percorso?

Visto che si tratta di uno dei nostri percorsi prioritari, se pensi che questo percorso possa essere un'ottima occasione per te, saremmo *particolarmente* entusiasti di consigliarti sui tuoi prossimi passi. Possiamo aiutarti a considerare le tue opzioni, a creare connessioni con altri che lavorano nello stesso campo e, se possibile, anche aiutarti a trovare un lavoro o delle opportunità di finanziamento.

CANDIDATI PER PARLARE CON IL NOSTRO TEAM

Scopri di più

Ulteriori letture essenziali (in inglese):

- Per aiutarti a orientarti nel campo, ti consigliamo lo AI safety starter pack
- Guida passo-dopo-passo di Charles Rogers Smith sulle carriere nella sicurezza della IA
- Il nostro profilo di problemi sui rischi dell'IA
- Questo curriculum sulla sicurezza dell'IA (o, per qualcosa di più breve, questa serie di post di Richard Ngo)
- La nostra guida su come diventare un tecnico nel campo del machine learning con un focus sulla sicurezza dell'IA

Ulteriori letture (in inglese):

- Recensioni della carriera come dottorato nel machine learning
- Lista di letture dal Center for Human-Compatible AI

- Una serie di liste di lettura sulla sicurezza dell'IA
- Podcast: Dr Paul Christiano on how OpenAI is developing real solutions to the 'AI alignment problem', and his vision of how humanity will progressively hand over decision-making to AI systems
- Podcast: Machine learning engineering for AI safety and robustness: a Google Brain engineer's guide to entering the field
- Podcast: The world needs AI researchers. Here's how to become one
- Podcast: Chris Olah on working at top AI labs without an undergrad degree e What the hell is going on inside neural networks
- Podcast: A machine learning alignment researcher on how to become a machine learning alignment researcher
- Leggi tuti i nostri articoli sulle carriere nella sicurezza dell'IA

Prevenire una catastrofe legata alle IA

29 agosto 2022

Originale disponibile qui: https://80000hours.org/problem-profiles/artificial-intelligence.

Noi di 80,000 Hours abbiamo appena pubblicato la nostra analisi più lunga e più dettagliata di sempre. Parla di come ridurre i rischi esistenziali legati alle IA e si può trovare qui.

Il resto di questo intende fornire un background del profilo, una sinossi e un indice dei contenuti.

Un po' di background

Come molti dei nostri contenuti, questo profilo è indicato a chi ha già passato un po' di tempo sul nostro sito ma non ha familiarità con il mondo dell'Altruismo Efficace, ragion per cui è un'analisi piuttosto introduttiva. Ciò nonostante, speriamo che possa essere utile anche ai membri della comunità dell'Altruismo Efficace.

Il profilo rappresenta principalmente il mio (Benjamin Hilton) punto di vista, ma è stato corretto da Arden Koehler (il nostro direttore) e revisionato da Howie Lempel (il nostro CEO). Entrambi sono d'accordo in linea generale con le mie conclusioni.

Alcuni degli accorgimenti che ho preso per fare in modo che questo profilo potesse essere il più utile possibile per chi è nuovo a queste tematiche:

- Mi sono concentrato su quello che per me è il problema più importante: i rischi di IA alla ricerca di potere che potrebbero derivare da sistemi di progettazione senzienti con capacità avanzate, come delineato da Joe Carlsmith.

- Ho cercato di parlare sempre in termini concreti e ho scritto un articolo a parte su come potrebbe avvenire una catastrofe causata da un'IA. (Devo molto al report di Carlsmith, a *What failure looks like* di Cristiano e a *Superintelligenza* di Bostrom.)

- Ho fornito (quelle che secondo me sono) informazioni di base importanti, come i risultati dei sondaggi degli esperti di machine learning sul rischio legato alle IA, una panoramica dei progressi recenti nel campo delle IA e sulle leggi di potenza.

- Ho cercato di spiegare con sincerità perché le argomentazioni che presento potrebbero rivelarsi errate.

- Ho incluso una lunga lista di domande frequenti in cui presento quelle che ritengo essere le risposte migliori alle obiezioni più comuni sul lavorare ai rischi delle IA. Inoltre, se volete darmi un feedback su questi contenuti e preferite non farlo pubblicamente, potete usare l'apposito modulo.
  Questo post include una sinossi dell'articolo e un indice dei contenuti.

Sommario

Ci aspettiamo di vedere progressi significativi nel campo delle IA nei prossimi decenni, forse anche di raggiungere un punto in cui le macchine supereranno gli umani in molti, o tutti, gli

ambiti. Questo potrebbe portare vantaggi incredibili, aiutarci a risolvere problemi globali attualmente irrisolvibili, ma potrebbe anche comportare gravi rischi. Questi rischi potrebbero emergere per errore (ad esempio se non troviamo una soluzione tecnica per il problema della sicurezza delle IA) o volontariamente (ad esempio se le IA dovessero far precipitare un conflitto geopolitico). Pensiamo anche che sia necessario lavorare di più in questo campo per ridurre questi rischi.

Alcuni rischi legati a IA avanzate potrebbero essere esistenziali — vale a dire che potrebbero portare l'umanità all'estinzione o comunque provocare così tanti danni da farci perdere il ruolo di specie dominante sul pianeta.[1] Ad oggi non ci sono ancora soluzioni soddisfacenti che ci permettano di sviluppare e integrare nella società questa tecnologia trasformativa, il cui avvento sembra sempre più vicino. Secondo le nostre stime attualmente ci sono circa 300 persone nel mondo che stanno lavorando a questo problema.[2] Il risultato è che una catastrofe causata da un'IA potrebbe essere il problema più urgente al mondo, nonché quello su cui dovrebbero assolutamente lavorare coloro che possono dare un importante contributo. Tra le soluzioni più promettenti attualmente ci sono la ricerca tecnica per la creazione di intelligenze artificiali sicure, la ricerca strategica sui possibili rischi legati alle IA e la ricerca sulle politiche che governi e compagnie dovrebbero adottare per scongiurare questi rischi. Se verranno sviluppate politiche soddisfacenti, ci sarà bisogno di persone che le istituiscano e che le attuino. Ci sono anche molti ruoli complementari che danno la possibilità di avere un grande impatto, come ad esempio operation management, giornalismo, guadagnare per donare e altri ancora. Alcuni sono elencati qui sotto.

Il nostro punto di vista

Consigliato – priorità massima

Uno dei problemi più urgenti su cui lavorare.

Portata

Le IA possono essere utili in moltissimi contesti e c'è la possibilità che abbiano un enorme impatto positivo, ma quello che ci preoccupa in particolare è la possibilità di risultati estremamente negativi, soprattutto in caso di catastrofe esistenziale. Abbiamo ancora molti dubbi, ma in base alle valutazioni fatte da altre persone, la nostra idea generale è che il fattore di rischio di una catastrofe esistenziale provocata da un'intelligenza artificiale nei prossimi 100 anni è di circa il 10%. Ulteriori ricerche potrebbero far variare di molto questa cifra – secondo alcuni esperti non supera lo 0,5%, secondo altri è di oltre il 50% e non escludiamo che uno dei due gruppi possa avere ragione. In generale, riteniamo che, quando si parla della prosperità dell'umanità sul lungo periodo, lo sviluppo delle IA costituisca una minaccia di gran lunga superiore di qualsiasi altro problema.

Trascuratezza

Nel 2020 sono stati investiti circa 50 milioni di dollari per ridurre i rischi peggiori legati alle IA, mentre ne sono stati spesi miliardi per aumentare le loro capacità.[3] [4] Nonostante i sempre maggiori timori degli esperti, attualmente ci sono solo circa 300 persone che lavorano per ridurre il rischio di una catastrofe esistenziale provocata da un'IA.[2] Di queste, circa due terzi lavorano alla ricerca sulla sicurezza delle IA, mentre le rimanenti si dividono tra chi fa ricerca strategica e chi ricerca e sostiene politiche adeguate.

Fattibilità

Fare progressi nella prevenzione di catastrofi causate dalle IA può sembrare difficile, ma ci sono molti ambiti in cui è possibile fare più ricerca e il campo stesso è molto recente. Per questo pensiamo che sia possibile risolvere i problemi, anche se abbiamo ancora molti dubbi – le valutazioni a questo proposito tendono a variare enormemente.

Note

1. *Ci preoccupa anche la possibilità che le intelligenze artificiali possano risultare meritevoli di considerazioni morali - ad esempio, perché acquistano coscienza di sé. Anche se non possiamo discuterne in questo articolo, abbiamo trattato l'argomento qui.*

2. *Stime ricavate grazie al database di AI Watch. Per ogni organizzazione ho stimato la percentuale di dipendenti che lavorano direttamente per ridurre i rischi esistenziali legati alle IA. Molte di queste stime sono soggettive (ad esempio, "questo programma di ricerca riguarda davvero la sicurezza delle IA?") e potrebbero essere troppo basse in caso AI Watch non disponesse di dati su qualche organizzazione o troppo alte se nei dati sono presenti ripetizioni o se considerano persone che non lavorano più in quell'ambito. Con un intervallo di confidenza del 90%, stimo tra le 100 e le 1500 persone.*

3. *È difficile stabilire con precisione quanti soldi vengano spesi per migliorare le capacità delle IA, in parte a causa della scarsità di dati e in parte per via di domande come queste:*
    - *Quali ricerche sulle IA stanno davvero migliorando capacità pericolose che potrebbero far aumentare il rischio di catastrofe esistenziale?*
    - *Devono essere inclusi i progressi fatti sull'hardware delle IA o nella raccolta di dati?*
    - *E per quel che riguarda progressi sui processi di ricerca in generale o fattori che potrebbero portare in futuro a una crescita economica e quindi a possibili nuovi investimenti?*

    La cifra più significativa che siamo stati in grado di trovare sono stati i costi sostenuti da DeepMind nel 2020, all'incirca 1 miliardo di sterline, secondo il loro rendiconto annuale. Ci aspettiamo che un budget del genere abbia contribuito in qualche modo ai "progressi nelle capacità delle IA", considerando che il loro obiettivo principale è sviluppare sistemi di IA generali molto potenti. (Vale la pena notare, tuttavia, che DeepMind contribuisce anche al lavoro sulla sicurezza delle IA, che potrebbe ridurre il rischio esistenziale.)

    Se DeepMind spende circa il 10% nel migliorare le capacità delle IA, otteniamo una cifra attorno ai 10 miliardi di sterline. (Dal momento che ci sono molte aziende di IA negli Stati Uniti e si stanno facendo molti sforzi per sviluppare IA avanzate in Cina, il 10% ci sembra una stima ragionevole.)

    Come limite superiore, le spese nel settore delle intelligenze artificiali nel 2021 sono state di circa 340 miliardi di dollari.
    In generale, quindi, riteniamo che i soldi spesi nel miglioramento delle capacità delle IA vadano da 1 miliardo a 340 miliardi di dollari l'anno. Anche ipotizzando solo 1 miliardo di dollari, si tratterebbe di una cifra 100 volte maggiore di quello che viene speso per ridurre i rischi delle IA.

Sezioni del report completo [opzionali e in inglese]

- 1. Many AI experts think there's a non-negligible chance AI will lead to outcomes as bad as extinction
- 2. We're making advances in AI extremely quickly
    - Current trends show rapid progress in the capabilities of ML systems

Il teorema di Bayes e l’evidenza scientifica

[OPZIONALE] Guida al Teorema di Bayes

Manuale EA – 16 giugno 2022 - 4 minuti di lettura

Originale disponibile qui: https://arbital.com/p/bayes_rule/?l=1zq.

Far pagare l'affitto alle proprie credenze

di EliezerYudkowsky – 29 Luglio 2007 - 8 minuti di lettura

Originale disponibile qui:
https://www.lesswrong.com/s/6xgy8XYEisLk3tCjH/p/a7n8GdKiAZRX86T5A.

Così comincia una classica storia:

> *Se un albero cade in una foresta e non c'è nessuno a sentirlo, fa rumore? Una persona dice "Sì, perché produce vibrazioni nell'aria", e un'altra dice "No, perché non c'è nessun cervello con il suo meccanismo uditivo nei paraggi."*

Se esiste un'abilità fondamentale nell'arte della razionalità, una forma mentis su cui poggiano tutte le altre tecniche, potrebbe essere questa: l'abilità di notare, nella propria testa, sia i segni psicologici dell'avere una mappa mentale di qualcosa, che quelli del non averla.

Immaginiamo che, dopo che l'albero è caduto, i due che stavano discutendo passeggino assieme nella foresta. Si aspetterà forse uno dei due di vedere l'albero caduto alla loro destra, mentre l'altro alla loro sinistra? Immaginiamo che, prima che l'albero cada, i due mettano vicino all'albero un registratore. Si aspetterà forse uno dei due di sentire qualcosa di diverso dall'altro, quando ascoltano la registrazione? Immaginiamo anche che colleghino un elettroencefalogramma a qualsiasi cervello al mondo: si aspetterà forse uno dei due di vedere un tracciato diverso rispetto all'altro?

Anche se i due discutono, uno dice "No" e l'altro dice "Sì", non si aspettano di avere esperienze diverse. Pensano di avere schemi diversi del mondo, ma non che questi schemi differiscano per quel che riguarda ciò che si aspettano *succederà*; quando si parla di esperienze sensoriali, le loro mappe del mondo sono identiche.

Potremmo essere tentati di eliminare questa categoria di errori affermando che l'unica credenza legittima sia l'aspettativa di un'esperienza sensoriale. Ma in realtà, molte cose nel mondo non possono essere esperite con i cinque sensi. Non vediamo gli atomi che formano il mattone, ma gli atomi ci sono comunque. Sotto i vostri piedi c'è un pavimento, ma non avete *esperienza* diretta del pavimento; vedete la luce *riflessa* sul pavimento, anzi, vedete ciò che la vostra retina e la vostra corteccia visiva hanno ricavato da quella luce. Ricavare il pavimento partendo dal vedere il pavimento significa fare un passo indietro ed entrare nelle cause nascoste dell'esperienza. Forse è un passo molto breve, ma è comunque un passo indietro.

Vi trovate in cima a un palazzo accanto a un orologio a pendolo che segna le ore, i minuti e i secondi. Avete in mano una palla da bowling, che fate cadere dal tetto. Su quale ticchettio delle lancette sentirete l'urto della palla da bowling che colpisce il suolo?

Per poter rispondere con precisione, dovete ricorrere alle vostre credenze personali, come ad esempio "la gravità terrestre è pari a 9,8 metri al secondo per secondo" e "questo edificio è alto circa 120 metri". Queste credenze non sono aspettative, semplici ed inarticolate, di un'esperienza sensoriale; sono fatte (più o meno) di parole, sono proposizionali. Non sarebbe

una grande esagerazione, probabilmente, descrivere queste due credenze come frasi, composte da parole. Eppure queste due credenze hanno una *conseguenza* inferenziale che è un'aspettativa sensoriale precisa: se la lancetta dei minuti dell'orologio è sul 12 quando lasciate cadere la palla, vi aspetterete di vederla sull'1 quando sentirete il tonfo cinque secondi dopo. Per prevedere con la massima precisione possibile queste esperienze sensoriali, dobbiamo elaborare credenze che non sono aspettative di esperienze sensoriali.

Uno dei vantaggi che l'*Homo sapiens* ha su ogni altra specie al mondo è che noi possiamo imparare a dare forma a ciò che non vediamo. È anche uno dei nostri punti deboli. Gli esseri umani spesso credono a cose che non solo non vedono, ma che non esistono neanche.

Quello stesso cervello che crea una rete di cause inferite alla base di un'esperienza sensoriale può anche creare una rete di cause slegate da un'esperienza sensoriale, oppure collegate male. Gli alchimisti pensavano che il flogisto creasse il fuoco – potremmo semplicisticamente modellare il loro modo di pensare con un nodo etichettato "Flogisto" collegato con una freccia all'esperienza sensoriale di un fuoco che scoppietta – ma questa credenza non ha portato ad alcuna previsione concreta; il collegamento tra il flogisto e l'esperienza è sempre stato creato dopo l'esperienza invece di limitare in anticipo l'esperienza.

Immaginate che il vostro insegnante vi dica che il famoso scrittore Wulky Wilkinsen è in realtà un "autore retroposizionato", cosa che si può dedurre dal fatto che le sue opere mostrano una "resublimazione alienata". Forse il vostro insegnante lo sa perché è stato il suo insegnante a dirglielo, ma forse l'unica cosa che è in grado di dirvi sulla resublimazione è che è un tratto del pensiero retroposizionato e che questo tipo di pensiero è caratterizzato da una resublimazione alienata. Cosa dovreste quindi aspettarvi dalle opere di Wulky Wilkinsen?

Nulla. Questa credenza, se così possiamo chiamarla, non è collegata a nessuna esperienza sensoriale. Ma è meglio che vi ricordiate che "Wulky Wilkinsen" ha la caratteristica di "retroposizionamento", nonché quella di "resublimazione alienata", in modo che possiate recitarle a memoria al prossimo esame. Le due credenze sono collegate tra loro ma non c'è nessun collegamento con un'esperienza attesa.

Siamo in grado di creare interi sistemi di credenze che sono collegate solo tra di loro; potremmo chiamarle "credenze fluttuanti". Di tutte le specie animali, solo gli esseri umani hanno questa anomalia mentale, questa perversione della capacità dell'*Homo sapiens* di creare sistemi di credenze più flessibili e più generali.

La virtù razionalista dell'*empirismo* consiste nel chiedersi continuamente quali esperienze sono predette – o meglio ancora, vietate – dalle nostre credenze. Pensate che il flogisto produca il fuoco? Allora, sulla base di ciò, cosa vi aspettate che succeda? Pensate che Wulky Wilkinsen sia un autore retroposizionato? Allora, sulla base di ciò, cosa vi aspettate di vedere? No, non la "resublimazione alienata"; *quale esperienza avrete*? Pensate che se un albero cade in una foresta fa rumore anche se non c'è nessuno a sentirlo? Allora quale sarà la vostra esperienza?

Meglio ancora: chiedetevi quale esperienza *non potrete* avere. Pensate che l'*Élan vital* sia la spiegazione per la misteriosa carica vitale degli esseri viventi? E allora cosa *non può* succedere per via di questa credenza? Quale dato la falsificherebbe inequivocabilmente? Una

risposta nulla significa che la vostra credenza non *limita* l'esperienza, ma fa sì che possa succedere *qualsiasi cosa*. È fluttuante.

Quando discutete di qualcosa che vi sembra fattuale, tenete sempre a mente di quale differenza nell'aspettativa state parlando. Se non riuscite a trovarne nessuna, allora forse state discutendo di etichette mentali nel vostro sistema di credenze – o peggio ancora, credenze fluttuanti, aggrappate al vostro sistema. Se non vi rendete conto delle esperienze insite nelle opere retroposizionate di Wulky Wilkinsen, potete andare avanti a discutere per l'eternità.

Soprattutto, non domandatevi in cosa dovreste credere: chiedetevi cosa dovreste aspettarvi. Ogni dubbio su una credenza dovrebbe scaturire da una domanda su cosa aspettarsi e quella domanda dovrebbe essere il fulcro del vostro ragionamento. Ogni ipotesi di credenza dovrebbe scaturire da una specifica ipotesi di aspettativa e dovrebbe continuare a pagare l'affitto in aspettative future. Se una credenza non paga l'affitto, sfrattatela.

Che cos'è una prova o evidenza?

di EliezerYudkowsky – 22 settembre 2007 - 4 minuti di lettura

Originale disponibile qui:
https://www.lesswrong.com/s/6xgy8XYEisLk3tCjH/p/6s3xABaXKPdFwA3FS.

> La frase "la neve è bianca" è *vera* se e solo se la neve è bianca.
>
> — Alfred Tarski

> Dire di ciò che è che è, o di ciò che non è che non è, è *vero*.
>
> — Aristotele, *Metafisica IV*

Immaginate di stare camminando per strada quando a un certo punto vi si slaccia una scarpa. Nulla di che, ma dopo un po', per qualche strana ragione, cominciate a *credere* che le vostre scarpe siano slacciate. La luce che viene dal sole colpisce i lacci delle vostre scarpe e si allontana; alcuni fotoni vengono catturati dalle vostre pupille e raggiungono la retina; l'energia di questi fotoni dà vita a impulsi neurali; gli impulsi neurali vengono trasmessi alle aree del vostro cervello deputate all'elaborazione delle immagini; qui l'informazione visiva è elaborata, ricostruita in un modello 3D e identificata come una scarpa slacciata. Una sequenza di eventi, una catena di cause ed effetti, nel mondo e nel vostro cervello, alla fine delle quali vi ritrovate a credere a quello in cui credete. Il risultato di questo processo è uno stato *mentale* che rispecchia lo stato reale delle vostre *scarpe.*

Che cos'è una *prova* o *evidenza*? È un evento "intrecciato", attraverso catene di causa ed effetto, con quello che vuoi sapere. Se l'obiettivo della vostra indagine sono i lacci delle scarpe, ad esempio, allora la luce che attraversa le vostre pupille è una prova correlata ai lacci. Uso il termine "intrecciato" per riferirmi alla causalità, ossia a due cose che si trovano collegate tra loro per via di un rapporto di causa ed effetto.

Non tutti gli effetti creano un "intreccio" necessario per portar ad un'evidenza. Avere un dispositivo che suona quando vi inserite i numeri vincenti della lotteria non è di alcun aiuto se il dispositivo suona *anche* quando inserite *altri* numeri. La luce che tocca le vostre scarpe non sarebbe molto utile come prova riguardo lo stato dei lacci, se i fotoni si organizzassero nello stesso stato fisico sia quando i lacci sono slegati che quando non lo sono.

Più in astratto: perché un evento possa essere considerato una *prova valida* per una data indagine, deve avvenire *diversamente* in un modo che sia intrecciato con i *diversi* stati possibili dell'oggetto in esame. (In linguaggio tecnico: deve esserci mutua informazione tra l'evento che porta la prova e l'oggetto dell'indagine, relativamente al vostro attuale stato di incertezza su entrambi).

Se elaborato correttamente, tale intreccio può essere contagioso, che è il motivo per cui è necessario avere un cervello e un paio di occhi. Se i fotoni toccano i lacci delle scarpe e si spostano poi su un sasso, quel sasso non cambierà granché. Il sasso non rifletterà i lacci delle scarpe in maniera significativa, né cambierà in maniera visibile in base a una scarpa allacciata o slacciata. Che è poi il motivo per cui le testimonianze dei sassi non sono granché utili in tribunale. Una pellicola fotografica catturerà la causalità fra lacci e i fotoni in arrivo in modo

che la foto possa essere usata come prova. Se i vostri occhi e il vostro cervello funzionano adeguatamente, *voi stessi* sarete in un rapporto di causalità con i lacci delle vostre scarpe.

Questo è il motivo per cui i razionalisti danno così tanta importanza all'affermazione, al limite del paradosso, secondo cui la credenza di una persona è degna di essere presa in considerazione solo se, in linea di principio, si potrebbe convincere quella persona a credere diversamente. Se lo stato della vostra retina fosse il medesimo a prescindere dal tipo di luce che vi entra, allora sareste ciechi. Alcuni sistemi di credenze, nel tentativo piuttosto ovvio di dare credito a se stessi, affermano che certe credenze sono degne di considerazione solo se ci si crede *incondizionatamente,* a prescindere da quello che uno possa vedere o pensare. Lo stato in cui si trova il vostro cervello dev'essere sempre lo stesso. Da cui l'espressione "credere ciecamente": se ciò in cui credete non dipende da quello che vedete, allora è come se qualcuno vi avesse accecati ficcandovi un dito nell'occhio.

Se i vostri occhi e il vostro cervello funzionano correttamente, le vostre credenze saranno per forza intrecciate ai fatti. *Il pensiero razionale dà vita a credenze che sono esse stesse delle prove.*

Se la vostra bocca dice la verità, allora le vostre credenze razionali, esse stesse una prova, funzioneranno come prova per qualcun altro. Questo intreccio può trasmettersi attraverso una serie di cause ed effetti – e se voi parlate e qualcuno ascolta, si tratta di causa ed effetto. Quando al telefono dite "ho le scarpe slacciate", state condividendo con un amico uno stato di causalità con le vostre scarpe.

Ne consegue che, tra persone oneste che pensano che anche chi hanno di fronte sia onesto, le credenze razionali sono contagiose. Ecco perché dire che le vostre credenze *non* sono contagiose – che ci credete per motivi personali che non sono trasmissibili – sembra parecchio sospetto. Se le vostre credenze sono intrecciate alla realtà, allora *dovrebbero* essere contagiose tra persone oneste.

Se il vostro modello della realtà vi suggerisce che il risultato dei vostri processi mentali *non* dovrebbe essere contagioso per gli altri, allora quel modello vi sta dicendo che le vostre credenze non sono davvero delle prove, il che significa che non sono intrecciate alla realtà. Quello che dovreste fare è intervenire per correggerle e smettere di crederci.

In effetti, se *sentirete*, a livello *non logico*, cosa questo *significa*, smetterete *automaticamente* di crederci, perché "la mia credenza non è intrecciata alla realtà" *equivale a dire"*la mia credenza non è esatta". Non appena smetterete di credere che " 'la neve è bianca' è vero", dovreste smettere (automaticamente!) di credere che "la neve è bianca", altrimenti c'è qualcosa che non va.

Cercate di spiegare perché i processi mentali che impiegate producono sistematicamente credenze che rispecchiano la realtà. Spiegate perché pensate di essere razionali. Perché pensate che, con gli stessi processi mentali che usate voi, le altre persone dovrebbero credere che "la neve è bianca" se e solo se la neve è bianca. Se *non* credete che il risultato dei vostri processi mentali sia intrecciato alla realtà, allora perché credete al risultato dei vostri processi mentali? È la stessa cosa, o perlomeno dovrebbe esserlo.

Rischi di sofferenza (s-risk)

[OPZIONALE] Perché i rischi di sofferenza sono i rischi esistenziali peggiori e come possiamo prevenirli

di Max_Daniel – 2 giugno 2017

Link al discorso (in inglese) qui: https://www.youtube.com/watch?v=jiZxEJcFExc&list=PLwp9xeoX5p8Pi7rm-vJnaJ4AQdkYJOfYL&index=14.

Gli altruisti efficaci che si concentrano sul futuro lontano si trovano a dover scegliere tra diversi tipi di interventi. Tra questi, gli sforzi per ridurre il rischio di estinzione umana sono quelli che finora hanno ricevuto più attenzione. Nel suo discorso Max Daniel porta avanti l'idea che forse dovremmo rafforzare questo tipo di lavoro con interventi che puntino a prevenire futuri ben poco desiderabili ("rischi di sofferenza") e questo è un motivo in più per concentrarsi sui rischi delle IA tra tutte le fonti di rischio esistenziale finora individuate.

[OPZIONALE] Per approfondire i rischi dell'AI (materiali in inglese)

EA Hanbook 15 luglio 2022

Lo sviluppo dell'intelligenza artificiale

- AlphaGo - The Movie - DeepMind - Documentario sull'intelligenza artificiale, l'antichissimo gioco del Go e cosa possiamo imparare sulle potenzialità future delle IA. (Filmato - 1 ora e 30 minuti)
- The Artificial Intelligence Revolution: Part 1 - Una divertente e interessante esplorazione dell'intelligenza artificiale dal famoso blogger Tim Urban. (45 minuti)

Altre risorse sull'allineamento dell'intelligenza artificiale

- AGI Safety Fundamentals Curricula
- My personal cruxes for working on AI safety (65 minuti)
- Professor Stuart Russell on the flaws that make today's AI architecture unsafe & a new approach that could fix it (Podcast - 2 ore 15 minuti)
- Some Background on Our Views Regarding Advanced Artificial Intelligence - Open Philanthropy Project - Una spiegazione del perché ci sia una seria possibilità che il progresso dell'intelligenza artificiale potrebbe essere comparabile alla transizione dall'era neolitica alla rivoluzione industriale. (1 ora)
- The Precipice (25 minuti)
- What Failure Looks Like - Due storie specifiche su come potrebbero essere peggiori scenari di una società risultata dal fallimento dell'allineamento di IA, che si sposta considerevolmente dalla classica storia della "esplosione dell'intelligenza" (12 mins.)
- AGI Safety from first principles - L'opinione di un ricercatore di IA sui fattori specifici per il problema di allineamento nell'intelligenza artificiale generale (1 ora 15 minuti)
- Human Compatible: Artificial Intelligence and The Problem of Control (Libro)
- The Alignment Problem: Machine Learning and Human Values (Libro)

Governance dell'intelligenza artificiale

- The new 30-person research team in DC investigating how emerging technologies could affect national security - 80,000 Hours - Come cambierebbe la sicurezza internazionale se gli effetti del machine learning fossero di portata simile a quelli dell'elettricità? (Podcast - 2 ore)
- Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority - Center for a New American Security - Come i progressi tecnologici in ambito militare (incluse, ma non solo, le IA) possono comportare rischi e creare problemi nel prendere decisioni importanti, degni di attenzione da parte delle strutture di sicurezza nazionali. (60 minuti)

Lavori tecnici sull'allineamento dell'IA

- AI Alignment Landscape (Video - 30 minuti)

- AI safety starter pack (7 minuti)
- How to pursue a career in technical AI alignment (59 minuti)
- Technical Alignment Curriculum (readings for a 7 week course)
- AI Alignment Forum, soprattutto le loro sequenze principali

Critiche ai rischi dell'IA

- How sure are we about this AI stuff? (26 minuti)
- A tale of 2.75 orthogonality theses (20 minuti)
- How to know if AI is about to destroy civilization (sommario, 2 minuti)
- The AI Messiah (e il primo commento) (5 minuti)
- How good is humanity at coordination? (4 minuti)