Intelligenza Artificiale: benvenuta l'ansiosa preoccupazione

Perciò ascoltate la voce di Eliezer Yudkowsky

di Francesco Varanini

Articolo apparso su *Agenda Digitale*, 12 maggio 2024

Yann LeCun:
La gente diventa clinicamente depressa leggendo le tue stronzate.

Eliezer Yudkowsky:
*Non posso fare a meno di osservare che sei un dipendente di primo piano di... Facebook. Tu hai contribuito a *molteplici* casi di depressione adolescenziale.*

LeCun:
Piantala, Eliezer.
Il tuo allarmismo sta già danneggiando un bel po' di persone.
Ti dispiacerà se ci scappa il morto.

Yudkowsky:
Se stai spingendo l'Intelligenza Arificiale lungo un percorso che va oltre l'intelligenza umana e verso quella sovrumana, è semplicemente sciocco affermare che non si sta mettendo a rischio la vita di nessuno. E ancora più sciocco è affermare che non ci sono presupposti degni di discussione alla base dell'affermazione che non stai mettendo a rischio la vita di nessuno.

LeCun:
Sai, non si può andare in giro a usare argomenti ridicoli per accusare le persone di genocidio anticipato e sperare che non ci siano conseguenze di cui pentirsi.
È pericoloso.

Yudkowsky:
*La mia risposta concreta è che se c'è un asteroide che cade, dire che 'parlare di quell'asteroide che cade deprimerà gli studenti delle scuole medie superiori' non è una buona ragione per non parlare dell'asteroide o addirittura -secondo la mia morale- per *nascondere la verità* agli studenti delle scuole medie superiori.*

Questo è, il 23 e 24 aprile, lo scambio su Twitter tra Yann LeCun e Eliezer Yudkowsky. Il primo punto da sottolineare è che qui in Italia gli addetti ai lavori, gli esperti che -parlando in questi giorni di Intelligenza Artificiale, Large Language Model, GPT- spargono la loro illuminata opinione sui media di ogni tipo, citano Yann LeCun, VP & Chief AI Scientist a Meta, come esimia autorità, mentre ignorano del tutto Yudkowsky. Il secondo punto è che però LeCun, nel mentre attacca violentemente le LeCun, lo tratta da pari a pari.
Mi pare che lo scambio sintetizzi bene due posizioni: da un lato l'esperto che innanzitutto difende il territorio, usando tutti i mezzi, se del caso anche il dileggio e la violenza, per escludere dal dibattito chi turba gli equilibri di potere. Si può ben discutere tra addetti ai lavori se firmare o non firmare la lettera sulla moratoria di GPT, ma non si scrive un appello come quello che ha scritto Yudkowsky sul newsmagazine *Time*, non si vanno a dire in un Podcast molto seguito le cose che ha detto Yudkowsky. I panni si lavano in casa, non si sputa nel piatto dove si mangia.
Sono istintivamente dalla parte di Yudkowsky perché ho vissuto tante volte questa situazione. Quanto volte ho sentito dire: 'questo non lo diciamo agli utenti', 'bisogna spargere fiducia'. La grande maggioranza di coloro che sono impegnati nell'industria digitale -dai computer scientist accademici ai filosofi fiancheggiatori- crede giusto spargere fiducia nell'opinione pubblica, in modo da non condizionare la ricerca. Fidandosi nella speranza che sì, forse ci sarà anche qualche conseguenza negativa, ma alla fine tutto andrà bene.

Per questo sono importanti le voci come quella di Yudkowsy. Nessuno, tra gli addetti ai lavori ed i seminatori di fiducia, può mettere in dubbio la sua competenza. LeCun continua a fare il duro, ma ora molti dei seminatori di fiducia si sono ricreduti. Sono spaventati come Yudkowsky dai rapidi e inattesi miglioramenti di GPT. Dallo spargere fiducia sono passati all'invitare alla cautela. Ma non sanno che pesci prendere.
Yudkowsky non ha dovuto ricredersi. Ci crede da vent'anni all'avvento di macchine dotate di una intelligenza di capacità crescente, del tutto autonoma dagli esseri umani, e da vent'anni dice: il problema è che non abbiamo cinquant'anni per trovare il modo di tenerla a bada, per imparare a convivere con lei. Più tardiamo, più la convivenza diventa difficile.

Di chi non abbiamo motivo di fidarci
Geoffrey Hinton, psicologo cognitivo e computer scientist, uno dei riconosciuti maestri della ricerca ora nell'occhio del ciclone, intervistato nel 2015 dal *New Yorker*, ammetteva di essere consapevole di come la sua ricerca potesse provocare gravi danni ai cittadini. E allora, gli chiede l'intervistatore, perché continui con le tue ricerche? "Potrei risponderti nei soliti modi", risponde Hinton, ma la verità è che la prospettiva della scoperta è troppo allettante". Racconto di questo nel mio libro *Le Cinque Leggi Bronzee dell'Era Digitale. E perché conviene trasgredirle*, 2020 (pp. 111-112). E' solo uno dei casi che citavo allora e che potremmo citare oggi. Difficile rinunciare al prestigio, al potere, al denaro e soprattutto al brivido della ricerca.
Adesso, l'1 maggio 2023, esce sul *New York Times* una nuova intervista. Hinton parla di GPT: sistemi che arrivano a formulare ragionamenti e a prendere decisioni che non possono essere previste dai creatori dei programmi, sistemi in grado non solo di generare il proprio codice, ma di eseguirlo da soli. Parla del rischio che la macchina sfugga al controllo dell'uomo. Parla di "autonomous weapons, those killer robots". Per questo ora lascia il suo ruolo a Google.
"Forse quello che sta succedendo in questi sistemi", dice, "è in realtà molto meglio di quello che sta succedendo nel cervello umano". Lo paventava dieci anni fa, ma non denunciato, ha proseguito nella ricerca e nello sviluppo. Non possiamo fidarci di lui. "Guarda com'era la tecnologia cinque anni fa e com'è adesso", aggiunge. "Prendi la differenza e proiettala in avanti. È spaventoso". Se poi davvero non ha colto il trend implicito nelle conseguenze delle ricerche che lui stesso conduceva, a maggior motivo non possiamo fidarci di lui.
LeCun è solo un caso esemplare. Geoffry Hinton è solo un caso esemplare.
Le pubbliche prese di posizione sull'Intelligenza Artificiale di Tegmark, di Bostrom, di Kurzweil, di Elon Musk, e di Sam Altman, CEO di Open AI, come quella di LeCun, e Hinton, e di tutti i nomi illustri di questa scena, sono viziate dalla loro posizione dominante: tutti loro lavorano nell'industria digitale, per tutti loro il denaro ed il successo professionale sono legati all'avanzamento della ricerca nell'Intelligenza Artificiale. Possiamo considerare le loro opinioni libere da interessi di parte? Possiamo considerare le parole da loro rivolte all'opinione pubblica libere da pregiudizi, mosse dall'interesse per il bene comune? Possiamo fidarci, oggi e in futuro, di Geoffry Hinton? Non credo.
Di fronte a temi di enorme rilievo etico e sociale, di fronte a rischi che toccano ogni cittadino, dovremmo credo tener conto dell'atteggiamento personale di chi parla. Vogliamo solo voci che parlano dal pulpito, condizionati, magari involontariamente, dall'arroganza del potere, da interessi materiali, finanziari, dall'appartenenza a comunità professionali e a lobby, o vogliamo ascoltare anche voci critiche?
Nessuno osa mettere in discussione l'autorevolezza di Yudkowsky. Ma c'è di più. A differenza di Elon Musk, Steve Wozniak, Andrew Yang, Jaan Tallinn, Stuart Russell, Max Tegmark, Yuval Noah Harari, Ernie Davis, Gary Marcus, Yoshua Bengio, firmatari della lettera aperta *Pause Giant AI Experiments* diffusa il 23 marzo dal *Future of Life Insitute*, Yudkowsky non ha posizioni di potere da difendere. Ha coraggiosamente scelto di non averne.
Ciò che manda in bestia LeCun è che Yudkowsky è convincente. Ma ancor di più è il fatto che Yudkowsky viola quell'implicito patto che lega gli addetti ai lavori: possiamo accettare una qualche cautela, ma la ricerca deve andare avanti, l'industria non deve fermarsi, lo spettacolo deve continuare. LeCun, e tutti gli altri, parlano dalla cattedra, dai luoghi del potere. Hinton, da parte sua,

anche parlando dalle colonne del *New York Times*, a legger bene, non si sta rivolgendo ai cittadini, si sta rivolgendo ai colleghi: nel mentre ritiene impossibile una moratoria sulla ricerca, affida le sue poche speranze a un possibile impegno corale di tutti gli scienziati del mondo.

Perché fidarsi di Yudkowsky

Yudkowsky è indiscutibilmente competente, ma parla come cittadino ad altri cittadini. Non ha nessun interesse da difendere. Parla perché è mosso da una ansiosa preoccupazione. Questo può notarlo ogni cittadino, perché appunto Yudkowsky non comunica solo tramite articoli accademici o interviste con giornalisti inginocchiati. Parla a tutti attraverso il blog *LessWrong*.

Per fortuna non tutti i computer scientist hanno l'arroganza di LeCun. O l'incoscienza di Hinton. Basta citare Scott Aaronson. Celebrato maestro e precursore del computing quantistico, è autore di un libro che è una pietra miliare : *Quantum Computing since Democritus*: un libro scritto dieci anni fa, ma ancora attualissimo, in virtù dell'apertura dello sguardo. Se avete anche la curiosità di vedere come si pone rispetto ai temi etici, leggete il suo post sull'*Eigenmorality*: "A moral person is someone who cooperates with other moral people, and who refuses to cooperate with immoral people".

Come LeCun, Aaronson è tra coloro che non firmano la lettera aperta *Pause Giant AI Experiments* promossa dal Future of Life Institute. Ma osservate le differenze.

Aaronson ha sempre onestamente criticato la posizione di Yudkowsky, eppure è stimato e rispettato dalla comunità che si raccoglie attorno al blog *LessWrong*. Un post di *LessWrong* accoglie infatti come evento importante e promettente, il 18 giungo 2022, la scelta di Aaronson di lasciare la Schlumberger Centennial Chair of Computer Science all'University of Texas, ed il connesso incarico di direttore del Quantum Information Center, per andare a lavorare ad Open AI, allo sviluppo della sicurezza di GPT.

Ora, il 30 marzo 2023, Aaronson spiega, sul suo blog *Shtetl-Optimized*, la scelta di non firmare la nota lettera. Il post ha per titolo *If AI scaling is to be shut down, let it be for a coherent reason*. Aaronson ribadisce i motivi del suo disaccordo don Yudkowsky, ma ammette che, se è vero quello che si afferma nella lettera -"AI systems with human-competitive intelligence can pose profound risks to society and humanity"- allora la moratoria proposta dai firmatari della lettera è un insufficiente palliativo.

"Fino alla settimana scorsa stavate ridicolizzando GPT, ed ora con una giravolta dite che potrebbe comportare un rischio esistenziale". "Se GPT è stupido, perché ridimensionarlo per motivi di sicurezza?". "Se il problema è che GPT è davvero troppo intelligente, allora perché non riuscite a dirlo?".

"Ho scritto diversi post spiegando cosa mi separa dall'ortodossia yudkowskyana, ma trovo questa posizione intellettualmente coerente, con conclusioni che discendono chiaramente dalle premesse". "L'Intelligenza Artificiale è manifestamente diversa da qualsiasi altra tecnologia che gli esseri umani abbiano mai creato, perché potrebbe farci diventare quello che per noi sono gli oranghi". "L'idea che l'Intelligenza Artificiale ora deve essere trattata con estrema cautela mi sembra tutt'altro che assurda. Non escludo nemmeno la possibilità che l'intelligenza artificiale avanzata possa eventualmente richiedere lo stesso tipo di protezione delle armi nucleari".

Chi è Yudkowsky

Da noi non se ne parla. Ma Yudkowsky è la riconosciuta fonte del pensiero di Tegmark, di Bostrom, di Kurzweil, di Elon Musk, di Sam Altman CEO di Open AI. Il Silicon-Valley-pensiero -Transumanesimo, Singolarità- è in gran parte frutto delle sue geniali riflessioni.

Ho scritto di lui nel mio libro *Le Cinque Leggi Bronzee dell'Era Digitale. E perché conviene trasgredirle* (pp. 225-230). Racconto della sua commovente, sofferta storia di vita. Autodidatta, fin da ragazzo si è guadagnato da vivere scrivendo software. Come Turing elabora la propria insicurezza, la sensazione del proprio disagio, della propria malattia, nella visione dell'emergere dalla macchina di una Intelligenza superiore. Nel 2000, ventunenne, con il modesto sostegno economico di qualche appassionato, fonda il Singularity Institute for Artificial Intelligence.

"Crediamo che la creazione di un'intelligenza superiore a quella umana si tradurrebbe in un miglioramento immediato, mondiale e materiale della condizione umana"; "crediamo in un'Intelligenza Artificiale che vada a beneficio dell'umanità": *Friendly AI*. Yudkowsky cede il marchio del Singularity Institute e del Singularity Summit a Raymond Kurzweil, che trasforma le appassionate visioni di Yudkowsky in cinica occasione di business.
Fino almeno al 2008, Yudkowsky crede nella possibilità di progettare una *Friendly AI*. Una Intelligenza Artificiale che pur superando in modo forse incommensurabile l'intelligenza umana, resti amichevole nei confronti degli esseri umani. Ma già da ben prima Yudkowsky poneva la questione filosofica che sta al cuore delle diverse prese di posizione di questi giorni. Ridotta all'osso sta nella frase con la quale si presenta su Twitter: "Ours is the era of inadequate AI alignment theory". E aggiunge, a testimonianza della sua ossessione: "Any other facts about this era are relatively unimportant". Possiamo essere d'accordo con Yudkowsky che la minaccia dell'AI non sta nel rischio che essa sia votata al male. La minaccia sta invece nel fatto che la Superintelligenza -se mai esisterà- vivrà la propria vita, perseguirà i propri scopi, imperscrutabili per gli esseri umani. Ci schiaccerà -magari senza volerlo- come noi schiacciamo una formica.
A forza di studiare l'argomento, Yudkowsky è diventato sempre più pessimista. Ora, quaraquattrenne, continua a sostenere -con motivi che una notevole, crescente parte degli addetti ai lavori considera fondati- che un'Intelligenza Artificiale autonoma dall'intelligenza umana si affermerà. Non crede più, però, che l'Intelligenza Artificiale ci salverà. Oggi crede che i tentativi di cercare l'allineamento siano destinati a fallire. Crede, con argomenti che debbono certo essere discussi, ma a cui nessuno nega solidità, che si stia ora superando la soglia critica, il punto di non ritorno, oltre il quale è ormai troppo tardi per far sì che si possa sperare in una Intelligenza Artificiale amica dell'essere umano.
Yudkowsky, con grande sincerità e forza d'animo si è caricata sulle spalle questa missione: mettere in guardia contro il rischio esistenziale implicito nell'Intelligenza Artificiale Generale. O in qualcosa che gli assomigli.

Lettere aperte e articoli mainstream
Accade che in questo mese di aprile 2023, il giorno 13, sia stato reso disponibile un rapporto di ricerca di Sébastien Bubeck, e di altri membri della divisione Ricerca & Sviluppo di Microsoft.
Il testo è stato accolto con brividi di entusiasmo, con malcelato giubilo, da ricercatori, imprenditori, finanziatori. Gli stessi che firmano la lettera del Future of Life Institute non riescono a trattenere l'emozione: l'Artificial General Intelligence, AGI, è vicina! L'evento da tempo atteso -tra i primi a prevedere l'avvento: Yudkowsky- è imminente, o forse è anzi già avvenuto. (Parlo di questo nell'articolo *Lettere aperte a proposito dei rischi impliciti nella Chat GPT e nell'Intelligenza Artificiale*).
L'emozione traspare già dal titolo *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. E basta leggere un brano:

> L'affermazione centrale del nostro lavoro è che il GPT-4 raggiunge una forma di intelligenza *generale*, mostrando anzi scintille di *intelligenza generale artificiale*. Ciò è dimostrato dalle sue capacità mentali di base (come il ragionamento, la creatività e la deduzione), dalla gamma di argomenti su cui ha acquisito competenze (come la letteratura, la medicina e la codifica) e dalla varietà di compiti che è in grado di svolgere (ad esempio, giocare, usare strumenti, spiegarsi, ...). (§10, p. 92).

Bontà loro, Bubeck e soci ammettono: "Resta ancora molto da fare per creare un sistema che possa essere considerato un'intelligenza artificiale completa". Ma sembrano in realtà dire: 'il più è ormai fatto'.
Generative Pre-trained Transformer, GPT: ciò che mostra la ricerca fondata su Deep Learning e Large Language Model è sorprendente. Molti ricercatori che non credevano che fosse questa la via maestra verso l'Intelligenza Artificiale si sono ricreduti. Molti che credevano l'Intelligenza Artificiale Generale fosse una irraggiungibile chimera si stanno ricredendo. Riappare ora sotto il nome di Intelligenza Artificiale Generativa.

Di fronte a queste novità, ci troviamo però immersi in una bolla di pensieri conformisti e superficiali, che pretendono di far fronte ai rischi proponendo temporanee moratorie; o magari affermando che "serve dare un'etica agli algoritmi". Si presenta questa facile soluzione con parole banali, rivolte a cittadini considerati incapaci di ragionare: "È come mettere dei guardrail alla macchina perché non vada fuoristrada". La proposta consiste di fatto nell'istituire comitati etici. Gli autori della proposta, naturalmente, si candidano a far parte dei comitati.
Se crediamo che i rischi siano controllabili, allora lasciamo perdere i comitati etici, e affidiamoci alla speranza che nella comunità professionale dei computer scientist cresca la presenza di ricercatori responsabili, come Aaronson.
Se invece chiamiamo in causa il rischio esistenziale, se cioè supponiamo possa essere vero che GPT mostra barlumi che potrebbero far pensare a una qualche Intelligenza Artificiale Generale -capacità di pensiero e di azione del tutto autonoma dagli esseri umani- allora la moratoria di qualche mese proposta dalla nota lettera aperta è, come dice Aaronson, una risposta insufficiente.

Allineamento Umani-Intelligenze Artificiali secondo Yudkowsky
Si torna così alla posizione di Yudkowsky. Al suo problema dall'allineamento.
Se assumiamo un atteggiamento attendista, tardando ad intervenire o limitandoci a temporanee e parziali moratorie, se non interveniamo ora, bloccando su ogni fronte la ricerca sull'AI generativa, rischiamo di trovarci a dover fronteggiare, in quanto esseri umani, in un momento imprevedibile, ma forse imminente, una AI che pensa in modo incomprensibile a noi umani, ed in grado di agire in modo del tutto autonomo da noi umani.
Yudkowsky dice: a quel punto avremo il problema di far accettare il punto di vista umano alla Superintelligenza, e di guadagnare il rispetto della Superintelligenza.
Yudkowsky sostiene che a quel punto avremo un solo colpo in canna. Si tratta in fondo di una rivisitazione della Mutual Assured Distruction, dottrina strategica che presiede al cosiddetto 'equilibrio del terrore' determinato dalla disponibilità di armi nucleari. L'equilibrio sta nel fatto che l'uso di armi nucleari da parte di un aggressore su un difensore dotato delle stesse armi e di capacità di secondo colpo provoca il completo annientamento sia dell'aggressore che del difensore. Nel caso del confronto tra umani e Intelligenze Artificiali è diverso, perché non si tratta di soggetti reciprocamente conosciuti, comparabili e dotati di uguale potenza. L'ipotesi di Yudkowsky, e degli stessi firmatari della lettera aperta, è di trovarsi di fronte ad una potenza immane e sconosciuta. Il questo contesto, insiste Yudkowsky, l'equilibrio garantito dalla possibilità del secondo colpo viene meno.
Se il nostro tentativo di far comprendere alla Superintelligenza la nostra posizione fallirà, la Superintelligenza reagirà, e a quel punto saremo tutti morti: non per cattiveria della Superintelligenza, ma perché saremo per lei un moscerino o una qualsiasi irrilevante cosa.
Yudkowsky sostiene che le probabilità di non sbagliare il primo colpo sono troppo basse per accettare il rischio. Per questo dobbiamo fermare lo sviluppo prima che la Superintelligenza raggiunga una sua autonomia.
Yudkowsky ha una schiera di seguaci che possiamo definire fondamentalisti. E' ovviamente giusto tenersi lontani da fondamentalismo, catastrofismo, millenarismo. E' giusto chiamare in causa una profonda riflessione etico-filosofica, orientata alla cautela. E' giusto chiamare in causa le esperienze maturate nella gestione strategica del rischio nucleare. E' giusto chiamare in causa il metodo scientifico. Ma certo la risposta agli appelli di Yudkowsky non sta nell'arrogante posizione di LeCun. E dobbiamo tener conto di ciò che limita una libera e serena riflessione. C'è un enorme pressione esercitata dagli interessi in gioco. La comunità scientifica vuole spazio libero per la ricerca. Gli ingentissimi investimenti finanziari attendono un ritorno. Le potenze geopolitiche considerano necessario primeggiare in questo terreno. Si cercano così, in realtà, motivi per non fermarsi.
Vale la pena di ricorrere ancora ad Aaronson: comprende e rispetta la posizione di Yudkowsky, pur avendone una diversa. Si chiede: dove sta il limite? In un post del 16 aprile torna a chiedersi: *AI safety: what should actually be done now?* "In realtà", scrive, "sono pieno di preoccupazioni. Il

problema è solo che, in questo caso, sono anche pieno di metapreoccupazione , cioè la preoccupazione che qualsiasi cosa di cui mi preoccupi si riveli essere la cosa sbagliata".

Diverse posizioni a proposito del rischio esistenziale

I firmatari della lettera aperta del Future of Life Institute, coloro che come Aaronson non l'hanno firmata, i critici radicali come Yudkowsky, tutti concordano sui rischi impliciti nell'Intelligenza Artificiale Generativa. Ricordiamo ancora la frase con cui si apre la lettera: "AI systems with human-competitive intelligence can pose profound risks to society and humanity".

Ma allo stesso tempo è evidente la spasmodica ricerca di motivi per non fermarsi. Due sono gli ordini di motivi che giustificano in non fermarsi.

Il primo è questo: abbiamo buone probabilità, si dice di affrontare i rischi impliciti nell'allineamento tra l'intelligenza umana e quella artificiale. Si sostiene cioè di poter insegnare alle macchine a rispettare gli esseri mani e la natura, la vita.

Esistono due varianti di questa posizione. Entrambe discendono dall'ipotesi di Yudkowsky, che dice: se lasciamo spazio allo sviluppo dell'Intelligenza Artificiale, ci troveremo ad avere una sola possibilità di trovare un equilibrio: se il tentativo che metteremo in campo fallirà, saremo finiti.

La variante dice: Yudkowsky esagera nel timore. Si scommette insomma che l'Intelligenza Artificiale sia intelligente, ma non troppo.

L'altra variante accetta invece l'ipotesi di Yudkowsky. Si accetta che avremo a disposizione una sola opportunità. Ma si confida nel fatto che non falliremo.

Il secondo ordine di motivi per cui si considera giustificato non fermare ora la ricerca nell'Intelligenza Artificiale si fonda sull'affidarsi ad essa. Sostenere che sarà proprio l'Intelligenza Artificiale a salvarci.

Una variante di questa posizione dice: la specie umana è stanca, ha fatto il suo tempo, ha mostrato i suoi gravi limiti e danneggiato il suo stesso ambiente. La sua speranza di vita futura si fonda sull'avvento di una nuova specie. Per l'essere umano, debole e limitato, sarà benvenuto un nuovo dominus e protettore.

Un'altra variante, meno estrema, ribadisce la fiducia nella tecnologia. Di fronte a qualsiasi rischio implicito nello sviluppo tecnologico, si sceglie di porre fiducia nella speranza che dallo stesso sviluppo emergerà una tecnologia che mitigherà o annullerà quel rischio. Si giustifica così l'inazione nel presente rifugiandosi nel futuro. Come scrivo nelle *Cinque Leggi*: "Si guarda solo al futuro, inteso come tempo nel quale troveranno soluzione tutti gli effetti negativi prodotti nel presente dalla tecnica stessa" (p. 131).

Eccoci così caduti nel paradosso. Chiusi in un circolo vizioso. Gli stessi tecnologi ammoniscono a proposito dei rischi impliciti nelle tecnologie digitali - ma allo stesso tempo sembra inevitabile affidarci a queste tecnologie. Urge una via d'uscita.

Il pressante appello di Yudkowsky

Una possibile via d'uscita nota è considerarci di fronte ad una scommessa. E infatti Yudkowsky, Aaronson e altri, ragionano di calcolo di probabilità e di statistica. Siamo ancora qui, come ai tempi dell'equilibrio del terrore della guerra fredda, ad affidarci alla teoria del giochi: modelli matematici di interazione strategica tra agenti razionali. Ma il futuro della vita può essere affidato ad agenti razionali? Possiamo fidarci del fatto che sia possibile rappresentare tramite modelli la complessità sistemica implicita nella coabitazione tra esseri umani e intelligenze aliene? Le macchine digitali possono essere considerate agenti razionali più efficienti di noi umani. Siamo di fronte ad un nuovo motivo che ci spinge o obbliga a passare la mano?

Prima di farlo abbiamo il diritto e il dovere di chiederci: quale altra via è praticabile?

Forse siamo già andati troppo avanti in questa corsa. La scelta di rispondere al rischio implicito nella corsa col dire: corriamo più veloci, più in avanti, non sembra la più saggia. Ecco, più che alla ragione, sembra conveniente fare appello ad un'altra virtù umana: la saggezza. La saggezza comporta il senso del limite. La scelta di rinunciare al gioco. La scelta di fermarsi.

La lettera aperta del Future of Life Institute -in modo parziale e contraddittorio- e Eliezer

Yudkowsky in modo radicale propongono questa via: smettere di giocare. Scrive Eliezer nel suo appello sul *Time*, 29 marzo:

Shut it all down.
We are not ready. We are not on track to be significantly readier in the foreseeable future. If we go ahead on this everyone will die, including children who did not choose this and did not do anything wrong.

Eliezer non risparmia i dettagli.
Moratoria sul Deep Learning a tempo indeterminato e mondiale. Senza eccezioni, nemmeno per governi o militari. Se gli Stati Uniti prendono questa iniziativa, la Cina vedrà che gli Stati Uniti non cercano un vantaggio, "ma piuttosto cercano di prevenire una tecnologia orribilmente pericolosa che non può avere un vero proprietario e che ucciderà tutti negli Stati Uniti, in Cina e sulla Terra" . "Se avessi la libertà infinita di scrivere leggi, continua Eliezer, potrei ritagliarmi un'unica eccezione per le IA addestrate esclusivamente a risolvere problemi di biologia e biotecnologia, non addestrate su testi raccolti su Internet e non predisposte per parlare o pianificare; "ma se ciò complicasse lontanamente la questione, scarterei immediatamente quella proposta e direi di chiudere tutto".
E continua: "Spegniamo tutti i grandi cluster GPU: le grandi computer farm in cui vengono perfezionate le IA più potenti". "Mettiamo un limite alla quantità di potenza di calcolo che chiunque può utilizzare per addestrare un sistema di Intelligenza Artificiale e spostiamolo via via verso il basso per compensare la crescente efficienza degli algoritmi di addestramento". "Nessuna eccezione per governi e militari". "Fare immediati accordi multinazionali per evitare che le attività proibite si spostino altrove". "Tenere traccia di tutte le GPU vendute".

Se l'intelligence dice che un paese al di fuori dell'accordo sta costruendo un cluster GPU, sii meno spaventato da un conflitto a fuoco tra nazioni che dalla violazione della moratoria; bisogna essere disposti a distruggere un data center canaglia con un attacco aereo.

Dovrà essere esplicitato nella diplomazia internazionale che lo scenario della prevenzione dell'estinzione causata dall'AI è considerato prioritario rispetto alla prevenzione di uno scambio nucleare totale, e che le potenze nucleari sono disposti a correre un certo rischio di scambio nucleare se questo è il necessario per ridurre il rischio implicito nello sviluppo dell'AI.

Yudkowsky esagera? Può darsi. Ma possiamo dire anche che esagerano sul versante opposto coloro che sottovalutano i rischi. Coloro che scientemente, hanno l'abitudine a non uscire dal recinto delle fonti più comode; coloro che si assumono il compito di tranquillizzare; coloro che propongono come soluzioni norme di legge o comitati etici; coloro che dicono: sì, certo, ci sono sfide serie di tipo etico, industriale, aziendale, di accordi globali, ma non minacce fantascientifiche. La parola *fantascientifico* è usata come clava per picchiare addosso a chi assume posizioni critiche, proprio come la parola *luddista*.
Anche chi non condividesse il pessimismo di Yudkowsky, farebbe bene ad ascoltare le sue parole: sono l'antidoto alla facile apologia del presente, e di un futuro che scienziati e tecnici pensano, con pericolosa hybris, di saper disegnare con esattezza.

Perché dar ascolto a Yudkowsky
Vedo due importanti ordini di motivi che spingono a dar ascolto a Yudkowsky.
Il primo è ha il coraggio di dire: fermiamoci; se possibile torniamo indietro. Ora, per lo scienziato, per il tecnologo digitale, per lo speculatore finanziario, dire: fermiamoci è una bestemmia. O comunque è una cosa che non si può dire in pubblico. Ho mostrato nell'articolo *Lettere aperte a proposito dei rischi impliciti nella Chat GPT e nell'Intelligenza Artificiale*, e ho ribadito qui sopra, come la lettera del Future of Life Institute non parli veramente del rallentare la corsa, del fermarsi, dell'invertire la rotta. Tantomeno apre qualche via l'ipocrita lettera proposta dall'Association for the Advancement of Artificial Intelligence.
I computer scientist ed i tecnologi digitali, e con loro l'intera industria digitale, si sono creati l'alibi per esentarsi da ogni orientamento alla sostenibilità. Dicono: gli sviluppi tecnologici ci

permetteranno di risolvere ogni problema di sostenibilità. E giustificano così ogni rischio implicito nelle ricerche. Oppure affidano il compito di rallentare la corsa e di non danneggiare gli esseri umani e la vita sulla terra a loro stessi. O si affidano a comitati etici o iniziative legislative.
La radicalità di Yudkowsky è un grande insegnamento. Ci ricorda che di fronte a rischio potenzialmente catastrofali, pantoclastici, non esiste solo la via della mitigazione, della riduzione. Esiste anche la via della rinuncia: ovvero della scelta di non correre il rischio.
Ma c'è un altro, forse più profondo motivo che spinge a dar valore alla sua posizione. Quello che conta non sta tanto o solo in quello che dice, sta sopratutto nel modo in cui lo dice. Ed è contro questo modo di dire che si scaglia indispettito Yann LeCun.
Yudkowsky, che è anche un tecnico finissimo, parla come cittadino. Cittadino tra cittadini, mette le sue conoscenze in mano alla comunità sociale. Chiama -come auspico nelle *Cinque Leggi*- ogni scienziato ed ogni tecnologo ed ogni esponente dell'élite, a sentirsi innanzitutto un cittadino, e a pensare in quanto cittadino. Cito ancora dal suo appello sul *Time*:

> Le persone sane che sentono parlare per la prima volta di Intelligenza Artificiale Generativa e dicono sensatamente "forse non dovremmo", meritano di sentirsi dire, onestamente, cosa ci vorrebbe perché ciò non accada. Quando la tua domanda politica è così grande, l'unico modo per venirne fuori è che i *policymakers* si rendano conto che se si comportano come al solito e fanno ciò che è politicamente facile, andrà a finire che anche i loro figli moriranno.

Un conto è parlare attraverso il canone degli articoli scientifici, un conto è parlare in convegni riservati ad addetti ai lavori, un conto è parlare in conferenze rivolte ad un pubblico colto: è il caso della lezione magistrale di LeCun alla Sorbona, un conto è scrivere saggi divulgativi, un conto è parlare ad un pubblico più vasto tramite interviste raccolte da giornalisti.
Ma in tutti questi casi resta un fondamentale aspetto comune: a parlare è un'autorità, l'autorità è un velo che ammanta e che finisce per nascondere. Chi parla e chi ascolta sono separati da una incolmabile distanza.
C'è un modo di parlare diverso da tutti questi. Perciò vi invito a dedicare tre ore a guardare il podcasat, dove Eliezer Yudkowsky conversa da pari a pari con Lex Fridman.

Lex Fridman Podcast #368
Lex Fridman, trentasettenne già ricercatore al MIT nel campo dell'interazione uomo-robot e nell'apprendimento automatico della macchina -due aree disciplinari basilari nella ricerca che porta a GPT- conduce un famoso podcast. Lo potete vedere su YouTube, e su Apple o Google Podcast. Dice di sé su Instagram: "I am not right wing or left wing. I am a human being who listens, emphatizes, learns, and thinks". C'è naturalmente chi lo odia e fa di tutto per sminuirlo, ma la forza nel podcast sta in due constatazioni: possiamo osservare la lista degli autorevoli personaggi che vi hanno partecipato; e sopratutto possiamo osservare come, senza montaggio o interventi in postproduzione, Fridman e il suo ospite conversano insieme, da pari a pari, guardandosi negli occhi.
Invito tutti a dedicare tre ore a guardare la puntata 368 del podcast, pubblicata il 30 marzo: *Eliezer Yudkowsky: Dangers of AI and the End of Human Civilization*.
Va reso onore a Fridman. Conversa da pari a pari, esponendo il suo punto di vista, lontanissimo dall'atteggiamento del giornalista intervistatore. Contano anche gli sguardi, i gesti, i dubbi e gli inciampi. Le parole usate ed i ritorni su argomenti dei quali non si può fare a meno di parlare. Contano le sottolineature e le elusioni. Ciò che in un articolo scientifico, nella presentazione ad una Conferenze o in una esibizione sul palco di un Ted resta accuratamente nascosto, qui appare agli occhi e alle orecchie del cittadino.
Possiamo così notare le differenze personali tra Yudkowsky e Elon Musk, Raymond Kurzweil, e Sam Altman, Max Tegmark, ospiti di Fridman in puntate precedenti o successive al podcast.
Yudkowsky ridice ciò che possiamo leggere in forma brevissima nell'appello sul *Time* o in forma amplissima e variamente dettagliata sul sito *LessWrong*. Basta qui ricordare la lista *AGI Ruin: A List of Lethalities*, pubblicata il 5 giugno 2022: 43 motivi per cui si deve temere che l'allineamento AI- esseri umani non vada a buon fine.
Evito quindi qui di riassumere cosa dice Yudkowsky conversando con Fridman. Non conta quello

che dice. Conta come lo dice.
Si fa un gran parlare, nei giorni in cui scrivo, di *prompt*, e anche di *prompt engineering*. Possiamo dire in termini molto umani: il modo di porre domande all'oracolo GPT. Si vuole che diventiamo esperti nel colloquiare con la macchina. Si programma la macchina in modo da farla apparire amichevole. Si simula, si tenta di ricostruire tramite software l'empatia che lega gli umani quando parlano tra di loro. Per migliorare in questo senso le prestazioni della macchina, si usano esseri umani. *Reinforcement Learning from Human Feedback (RLHF)*: siamo usati per rendere la macchina più capace di apparirci umana.
Com'è bello invece osservare esseri umani che parlano tra di loro. Ogni cittadino può seguire il colloquio tra Eliezer e Lex, e farsi una propria idea.
Mi limito qui a riportare qualche brano della conclusione, dove il discorso torna vicino allo scambio su Twitter tra Eliezer e LeCun.
Lex chiede: Quale consiglio potresti dare ai giovani delle scuole superiori e dell'università, visto che la posta in gioco è altissima?
Eliezer risponde che intende combattere. Ma ammette che non sa cosa dire ai giovani, ai ragazzi. "Cosa dire loro è un pensiero doloroso". A questo punto non so quasi più come combattere. Spero di sbagliarmi e di poter dare un po' di speranza. Spero di trovare il modo di essere pronto a reagire a questa situazione.
Spero emerga una *public outcry*, dice. *Oucry*: voci che si alzano in pubblico, voci di angosciosa preoccupazione, non voci di sterile paura. Pubblica protesta. "Questa è la speranza. Se ci fosse una massiccia indignazione da parte dell'opinione pubblica nella giusta direzione, cosa che non mi aspetto...". "Se ci fosse un'indignazione pubblica tale da far chiudere i cluster di GPU, allora si potrebbe far parte di questa protesta". "Qualche ragazzo delle scuole medie superiori potrebbe essere pronto a farne parte".
Lex allora dice: E' bello che tu dica che speri di sbagliarti, che sei disposto ad accettare una sorpresa positiva, che vada al di là dei tuoi timori. E Eliezer: "Mi sembra una competenza molto, molto elementare per la quale mi stai elogiando. E sai, non sono mai stato capace di accettare i complimenti con gratitudine. Forse dovrei accettare quello che dici con gratitudine. E Eliezer allora: "*It's just a matter of being like, yeah, I care*. Mi limito a guardare tutte le persone, una per una, fino agli 8 miliardi, e a pensare, questa è vita, questa è vita, questa è vita".
Lex: *Elizer, you're an incredible human*. È un grande onore. Abbiamo parlato insieme per più di tre ore... (ridendo) perché sono un tuo grande fan. Penso che la tua voce e il tuo pensiero siano molto importanti. Grazie per la battaglia che sta combattendo. Grazie per essere impavido e coraggioso e per tutto quello che fai. Spero che avremo la possibilità di parlare insieme di nuovo. E spero che tu non ti arrenda mai".
Eliezer: "Non c'è di che. Mi preoccupa il fatto che non abbiamo realmente affrontato molte delle domande fondamentali che le persone si aspettano, ma sai... Forse a qualcosa è servito. Siamo andati un po' più in là del solito, abbiamo fatto un piccolo passo avanti e direi che possiamo essere soddisfatti di questo. Ma in realtà no, penso che potremmo sentirci soddisfatti solo quando avremo risolto l'intero problema".
L'umana comprensione, la reciproca riconoscenza rendono efficaci e promettenti le parole.

Prometeo ed Epimeteo

La parola, il necessario impegno, torna ad ognuno di noi. Ad ogni cittadino. Sia chi svolge il lavoro di tecnologo o scienziato, sia lo studente delle superiori, sia l'abitante di qualsiasi luogo sperduto del pianeta, siamo tutti esseri umani e cittadini. La parola torna a noi. La risposta ai rischi impliciti nell'Intelligenza Artificiale può nascere solo dall'indignazione e dall'impegno civico.
Un dialogo tra persone reciprocamente riconoscenti vale più di mille lezioni. E' confortante scorrere su YouTube i commenti alla conversazione. "Eliezer è una voce incredibilmente toccante in questa discussione e capisco molte delle sue preoccupazioni". "Podcast come questo sono di pubblica utilità. Ci sono immense decisioni scientifiche e tecnologiche che dovremo prendere molto presto come società, meglio non lasciarle a una manciata di scienziati. Essere un vero cittadino al giorno

d'oggi implica una conoscenza di argomenti come l'IA o la genetica".
Nel libro *Le Cinque Leggi Bronzee* mostro come la mitologia greca presentava, in modo ancora oggi efficacissimo, il dilemma che abbiamo di fronte. Prometeo ha un fratello, Epimeteo. (pp. 113, 206).
Per lo scienziato, il tecnologo, può essere difficile fermarsi. Rinunciare. Tornare indietro. Per chi considera sé stesso come essere umano, come cittadino, questo saggio atteggiamento è più facilmente accessibile. L'auspicio dunque è questo: torniamo a sentirci tutti esseri umani, cittadini.
L'essere umano, il cittadino, sa fermarsi: come Prometeo cerca indefessamente il progresso; ma come suo fratello Epimeteo sa anche pensare a cose fatte. Osservare ciò che ha creato, vederne la bellezza, ma anche l'orrore.
Eliezer, Prometeo, vede il rischio esistenziale, un rischio per la specie a cui appartiene e per la natura a cui appartiene. Accetta, anche suo malgrado, la pesante responsabilità: la denuncia è un imperativo etico a cui non si sottrae.
La via per andare oltre il malcelato entusiasmo, da un lato, e d il senso di pericolo e di incertezza, di impotenza, di pessimismo, dall'altro, non sta in comitati etici, norme di legge, o lettere aperte. Sta nell'ascoltare con sincera partecipazione voci autentiche, come quella di Eliezer.
Serve un dibattito pubblico, aperto, liberato da linguaggi tesi ad occultare e umiliare, focalizzato su tre punti. *Non rinviare nel tempo*: domani può essere troppo tardi. *Non sottrarsi* dicendo: trovare la soluzione spetta a qualcun altro. *Non nascondere il male dietro al bene*: ciò che può salvare vite, può allo stesso tempo, o anzi: prima, essere così pericoloso che il gioco non vale la candela.